ULPGC
Universidad de
Las Palmas de
Gran Canaria

IUMA
Instituto Universitario de
Microelectrónica Aplicada

IUMA

Exploring Medical Artificial Intelligence Interpretability through the Lens of Information Theory

Doctorado en Tecnologías de Telecomunicación e Ingeniería Computacional

Tesis Doctoral
Abián Hernández Guedes
Las Palmas de Gran Canaria, Diciembre 2024

ULPGC
Universidad de
Las Palmas de
Gran Canaria

IUMA
Instituto Universitario de
Microelectrónica Aplicada

IUIBS
Instituto Universitario de
Investigaciones Biomédicas y Sanitarias

Universidad de Las Palmas de Gran Canaria
Escuela de Doctorado

Programa de Doctorado en Tecnologías de Telecomunicación e Ingeniería Computacional

Título de la Tesis

Exploring Medical Artificial Intelligence Interpretability through the Lens of Information Theory

Tesis Doctoral presentada por **D. Abián Hernández Guedes**

Dirigida por el **Dr. Gustavo Marrero Callicó**

Codirigida por el **Dr. Juan Ruiz Alzola**

Dr. Gustavo Marrero Callicó
(firma)

Firmado digitalmente por MARRERO CALLICO GUSTAVO - 42875067P
Fecha: 2024.12.04 22:34:06 Z

Dr. Juan Ruiz Alzola
(firma)

Firmado digitalmente por RUIZ ALZOLA JUAN BAUTISTA - 42098781H
Fecha: 2024.12.04 22:42:54 Z

D. Abián Hernández Guedes
(firma)

Firmado por HERNANDEZ GUEDES ABIAN - ***9199** el día 04/12/2024 con un certificado emitido por AC FNMT Usuarios

Fecha: December 4, 2024

ULPGC
Universidad de
Las Palmas de
Gran Canaria

IUMA
Instituto Universitario de
Microelectrónica Aplicada

IUIBS
Instituto Universitario de
Investigaciones Biomédicas y Sanitarias

Exploring Medical Artificial Intelligence Interpretability through the Lens of Information Theory

by

Abián Hernández Guedes

A thesis submitted in conformity with the requirements
for the degree of Doctor of Philosophy in Telecommunication Technologies and
Computational Engineering (T2IC)

Research Institute for Applied Microelectronics
University of Las Palmas de Gran Canaria

Gobierno de Canarias
Consejería de Economía,
Conocimiento y Empleo
Agencia Canaria de Investigación,
Innovación y Sociedad
de la Información

Canarias
avanza
con Europa

Unión Europea

FONDO SOCIAL EUROPEO

Exploring Medical Artificial Intelligence Interpretability through the Lens of Information Theory

Abián Hernández Guedes
Doctor of Philosophy in Telecommunication Technologies and Computational Engineering (T2IC)
Research Institute for Applied Microelectronics
University of Las Palmas de Gran Canaria
2024

Abstract

Artificial intelligence, particularly deep learning models, has rapidly advanced recently, demonstrating exceptional performance in areas such as image processing, natural language understanding, and, more critically, healthcare. These models hold immense potential in automating tasks traditionally performed by experts, such as disease diagnosis or medical imaging interpretation. However, despite their success, a significant barrier to the full integration of deep learning models into critical fields like medicine is their lack of transparency and interpretability. Often referred to as "black boxes", these models make decisions where understanding the processes within the hidden layers is nearly impossible, raising concerns about their reliability in critical scenarios.

This thesis aims to address the challenge of interpretability in deep learning models, particularly in medical contexts, by adopting an approach grounded in information theory. This perspective treats deep learning models as information processors that operate at multiple levels of abstraction, enabling them to extract useful features from complex, high-dimensional data. Throughout the dissertation, various deep learning methods are proposed to tackle key challenges in medical data analysis, including medical screening with limited datasets, feature selection for clinical data, or anomaly detection in imbalanced clinical data. Furthermore, hyperspectral imaging, an emerging modality for clinical purposes, has been used in experiments, particularly in the signal decomposition task known as hyperspectral unmixing. The thesis not only addresses these problems but also focuses on understanding the underlying mechanisms behind the models' operations.

The work provides theoretical and experimental insights into how deep learning models balance data compression and prediction, offering an interpretable framework based on information theory to gain more in-depth insights into their decision-making processes. The application of information theory allows for a more general and quantifiable interpretation of deep learning models, making it possible to evaluate how well these models transmit and process the information required for specific tasks. By analyzing the models from this abstract perspective, the thesis demonstrates that it is possible to design more interpretable architectures without compromising their performance.

The results presented in this dissertation demonstrate the potential of deep learning models in critical applications like healthcare, while also underscoring the importance of interpretability to ensure their safe and effective use. This work contributes to ongoing efforts to make deep learning-based artificial intelligence more transparent and reliable in high-impact fields such as medicine.

Resumen

La inteligencia artificial, en particular los modelos basados en redes neuronales, ha avanzado en los últimos años, mostrando un rendimiento excepcional en áreas como el procesamiento de imágenes, la comprensión del lenguaje natural y, de manera especialmente relevante, en ámbitos críticos como la salud. Estos modelos tienen un inmenso potencial para automatizar tareas que tradicionalmente realizaban expertos, como el diagnóstico de enfermedades o la interpretación de imágenes médicas. Sin embargo, a pesar de su éxito, una barrera significativa para la plena integración de los modelos de aprendizaje profundo en campos críticos como la medicina es su falta de transparencia e interpretabilidad. A menudo denominados "cajas negras" o "cajas opacas", estos modelos toman decisiones cuyo proceso en las capas ocultas es imposible de entender, lo cual genera inquietudes sobre su fiabilidad en escenarios críticos.

Esta tesis tiene como objetivo abordar el desafío de la interpretabilidad en modelos de aprendizaje profundo, particularmente en contextos médicos, adoptando un enfoque basado en la teoría de la información. Esta perspectiva trata a los modelos de aprendizaje profundo como procesadores de información que operan en múltiples niveles de abstracción, permitiéndoles extraer características útiles de datos complejos y de alta dimensionalidad. A lo largo de esta tesis, se proponen varios métodos de aprendizaje profundo para enfrentar desafíos clave en el análisis de datos médicos, como la clasificación con conjuntos de datos con muestras limitados, la selección de características en datos clínicos o la detección de anomalías en datos clínicos desbalanceados. Además, se explora el uso de imágenes hiperespectrales, una modalidad emergente en el ámbito clínico, aplicándola especialmente en la descomposición espectral de señales. La tesis no solo aborda estos problemas, sino que también se enfoca en comprender los mecanismos subyacentes en las operaciones de los modelos propuestos.

El trabajo proporciona perspectivas teóricas y experimentales sobre cómo los modelos de aprendizaje profundo equilibran la compresión de datos y la predicción, ofreciendo un marco interpretativo basado en la teoría de la información para obtener una comprensión más profunda de sus procesos de toma de decisiones. La aplicación de la teoría de la información permite una interpretación más general y cuantificable de los modelos de aprendizaje profundo, haciendo posible evaluar qué tan bien estos modelos transmiten y procesan la información necesaria para tareas específicas. Al analizar los modelos desde esta perspectiva abstracta, la tesis demuestra que es posible diseñar arquitecturas más interpretables sin comprometer su rendimiento.

Los resultados presentados en esta tesis demuestran el potencial de los modelos de aprendizaje profundo en aplicaciones críticas como la salud, al tiempo que subrayan la importancia de la interpretabilidad para garantizar su uso seguro y efectivo. Este trabajo contribuye a los esfuerzos en curso por hacer que la inteligencia artificial basada en redes neuronales sea más 'transparente' y confiable en campos de alto impacto como la medicina.

En memoria de Lukas.

Acknowledgements

After so many years, I have finally accomplished the greatest challenge of my life so far: the development of this thesis. In it, I have tried to reflect my perspective on the problems I address and the ways I believe they should be approached. However, the outcome of this work is the result of my personal and professional growth, shaped by interactions with people who have enriched my life, my thinking, and my personality. With this in mind, I want to express my gratitude to those I consider fundamental in this journey, as Alan Turing said:

The isolated man does not develop any intellectual power.

First, to my family. To my parents and sister, who are the pillars of my personal, ethical, and moral development. To my partner, Karelis, thank you for being my refuge during the most difficult moments and for patiently enduring my emotional vulnerabilities. And to Luna, for your unconditional love. You are the reason I never give up.

To my thesis advisors, Gustavo and Juan, thank you for giving me the opportunity to work and, above all, for granting me the freedom to think. Your trust in my decisions, even when I chose unconventional paths, has been crucial to my development as a researcher.

To Himar Fabelo, the main reason I started this journey and a constant role model throughout the doctorate. Your natural leadership is a skill I deeply admire.

To my colleagues from the MACbioIDi project, who made my first steps in research an unforgettable experience. To Asmaa and Marilola, for believing in the 'princeso' from the beginning, a gesture that greatly motivated me. To Nayra, whose ability to share knowledge reminds me of the best teachers I have had. Finally, to my friends: Iban, who validates my theory of friendship with his constant references to *The Simpsons*, and Guillermo, the most significant influence on my research profile, shaping the foundation of who I am in this field today.

I would also like to thank Idafen, Mónica, and Borja for the valuable conversations that helped me formalize many of the ideas presented in this thesis. When I recall the most creative moments of this work, one of you was always present.

To my colleagues at the Institute of Applied Microelectronics (IUMA): Raquel, Bea, Antonio, Laura, María, and Carlos. You are an example of how young people can be a source of inspiration. A special mention to Samuel Ortega, whose professionalism was one of the greatest motivations to improve during my early years.

To my collaborators at the Instituto Astrofísico de Canarias, especially Natalia and Enrique. Thank you for your commitment and for turning the article writing process into a creative and fun endeavor, with revision sessions full of memes that should never be lost.

To the people from the Department of Mechanical Engineering at Tokyo University of Science. I am deeply grateful to Professors Takemura and Takamatsu for allowing me to join their laboratory and creating an environment where I felt comfortable and motivated. Thanks also to Ryodai, Manami, and Alberto for their friendship and support. Alberto, especially, I cannot imagine my time in Japan without your company and advice.

Outside the academic context, I want to thank my lifelong friends: Cristian, Carlos, and Juan, for showing that time does not affect friendship. To Alejandro and Moisés, for being a source of admiration. To Aitor, who sets the tempo of my music. To Marlon and Alejote, the eternal 'super colegas del infierno'. And to the 'comité de ocio nocturno': Samuel, Noelia, and María Elena.

A special thanks to Guzmán, for being my confidant during the hardest moments, and to Constantine, who listened to me during the most critical days of the final stage of this thesis.

Finally, I want to mention my friends from Japan: Umar, my ‘besto friendo’, Nicharee, and Memi, for the happy moments we shared. Your laughter and company made my stay in Japan unforgettable.

To all the teachers and professors who guided and motivated me throughout my academic life: thank you.

Contents

List of Tables

List of Figures

List of Algorithms

Acronyms

AAE Adversarial AutoEncoder. 176, 177, 180

ACO Ant Colony Optimization. 102, 125

AD Anomaly Detection. 6, 7, 14, 171–177, 179, 181–183, 185, 186, 188, 191–193, 196–198, 201, 206

ADeLEn Anomaly Detection in Latent Spaces using Entropy-based score. 173, 177, 179, 181–189, 191–194, 196–198, 201, 202, 206

AE AutoEncoder. xxiv, xxvi, 25, 26, 28, 35, 65–67, 76, 77, 79, 80, 83, 85, 88–90, 92, 95–98, 100, 136, 140, 141, 143, 146, 147, 149, 155, 157, 175–177, 179, 182, 183, 185, 196, 198, 200, 201, 209–211, 214

AI Artifical Inteligence. 1–6, 11, 12, 26, 28–30, 36, 63, 64, 72, 95, 98, 99, 115, 132, 136, 172, 199

ANC Abundance Non-negativity Constraint. 137, 139, 143

ANN Artificial Neural Network. 19–21, 55

ASC Abundance Sum-to-one Constraint. 137, 139, 143, 150

AUC Area Under the Curve. 186, 188, 192, 194

BRATS BRAin Tumor Segmentation Dataset. 185, 194, 195

CDF Cumulative Distribution Function. 106, 108

CE Cross-Entropy. 17, 49, 75, 79–81, 127, 145, 183

CIFE Conditional Infomax Feature Extraction. 108, 109

CLHU Contrastive Learning for blind Hyperspectral Unmixing. xix, xxiv, xxv, 136, 137, 142–155, 157–165, 167–169, 201, 206, 214

CNN Convolutional Neural Network. 20, 22, 25, 26, 72

CPU Central Processing Unit. 20

CT Computed Tomography. 33, 63, 172, 177

CV Computer Vision. 4, 20, 22, 24, 26, 28, 32, 41, 63, 97, 100

Deep SAD Deep Semi-supervised Anomaly Detection. 176, 197, 198, 202

DeepLIFT Deep Learning Important FeaTures. 32, 67, 84, 96

DFU Diabetic Foot Ulcer. xxvi, 71–73, 76, 77, 82, 84, 85, 88, 90, 92, 95, 96, 99, 131, 200, 209, 217, 225, 226

DIP Deep Image Prior. 35, 140

DL Deep Learning. xxiii, 2, 4–7, 11, 19–24, 26–29, 31–36, 43–45, 63–67, 69, 72, 84, 89, 95, 97–101, 103–106, 108, 109, 111–113, 123, 124, 127, 128, 130–132, 135–137, 140–143, 146, 148, 149, 151, 155, 160, 168, 171–180, 182, 189, 191, 196, 198–201, 204–206, 214, 220, 223, 225

DNN Deep Neural Network. xxiii, 5, 7, 20–26, 28–31, 45, 56, 64–70, 75, 78, 85, 89, 90, 92, 95, 97, 98, 101, 103–107, 109, 111–115, 118, 123, 126, 128, 129, 131, 132, 136, 141, 142, 146, 149, 151, 155, 169, 171, 172, 178, 185–189, 191, 192, 197, 198, 200, 201, 204, 205, 210, 214

DPI Data Processing Inequality. 52, 65–69, 81, 85, 89, 90, 92, 96, 97, 111, 112, 128, 149, 150, 165, 166, 169, 183, 200, 201, 209

DPM Diffusion Probabilistic Model. 20, 114, 176, 177, 183, 197

ELBO Evidence Lower Bound. 19, 179, 180, 214, 215

ELMM Extended Linear Mixture Model. 137, 138, 141, 152–155, 157–159

EM Expectation-Maximization. 14, 18

ET Estimated Temperature. 217, 219, 225

ETD Estimated Temperature Difference. 217, 219, 226

FCLS Fully Constrained Least Squares. xix, xxv, 139, 152–155, 158, 159, 161

FLAIR Fluid-Attenuated Inversion Recovery. 195

FS Feature Selection. xix, xxiii, xxiv, 6–8, 37, 99–116, 118–126, 128–132, 135, 136, 141, 142, 171, 200, 205, 206, 211, 217, 219, 220, 223, 225

FSSU Fractional Sparse Spectral Unmixing. xxv, 155, 157–159, 167

GA Genetic Algorithm. 102, 125

GAN Generative Adversarial Network. 20, 183

GBM Generalized Bilinear Model. 138

GLMM Generalized Linear Mixture Model. 138, 141

GMM Gaussian Mixture Model. 14, 18, 174

GP Gaussian Process. 16, 31, 191

GPU Graphic Processing Units. 20

Grad-CAM Gradient-weighted Class Activation Mapping. 32, 43, 67

HSE Hot Spot Estimator. 217, 219, 225, 226

HSI HyperSpectral Imaging. xxiii, xxiv, 6, 7, 11, 38–41, 99, 101, 113, 115, 116, 119, 120, 123–126, 131, 135–137, 141, 142, 152, 162, 167–169, 200, 201, 205, 206, 227

HSIC Hilbert Schmidt Independence Criterion. 103

HSIC Lasso Hilbert Schmidt Independence Criterion Lasso. 103

HSU HyperSpectral Unmixing. xxiv, xxv, 7, 135–142, 144–146, 148, 150–152, 154, 155, 157, 160–164, 167, 168, 171, 201

IB Information Bottleneck. 54–56, 65, 66, 69, 78, 90, 95, 97, 98, 111, 113, 128, 148, 149, 200, 201, 206

ICA Independent Components Analysis. 55, 100, 136, 150, 168

IP Information Plane. xxiv–xxvi, 68–70, 79–82, 85–87, 90, 92, 96, 101, 128, 129, 131, 164–166, 200, 201, 209, 227

IQR Interquartile Range. xxiii, 117, 126, 174, 206

IR Infrared Radiation. 70, 71

IT Information Theory. 3, 7, 31, 44–49, 53, 54, 56, 64, 65, 68–70, 72, 78, 95, 99, 101, 108, 109, 111, 112, 128–132, 135–137, 141, 146–149, 164–167, 169, 172, 173, 178, 182, 183, 197, 198, 200, 201, 204–206, 211, 217, 227

KL Divergence Kullback-Leibler Divergence. 19, 31, 37, 49, 179–181, 183, 187, 197

LASSO Least Absolute Shrinkage and Selection Operator. xxiv, 102–104, 123, 124, 130, 219, 220, 223, 225

LCA Lateral Calcaneal Artery. 217, 220, 222, 226

LIME Local interpretable Model-agnostic Explanations. 33, 67

LMM Linear Mixture Model. 136–139, 142–145, 149, 150, 152, 165–168

LPA Lateral Plantar Artery. 217, 220, 222, 226

LSTM Long Short-Term Memory. 20, 175

MAP Maximum A Posteriori estimation. 16, 17

MCA Medial Calcaneal Artery. 217, 220, 222

MESMA Multiple-Endmember Spectral Mixture Analaysis. xxv, 155, 157–159

MI Mutual Information. xxvi, 49–52, 54, 55, 57, 66–70, 80–82, 85–87, 89, 92–94, 96, 98, 101–103, 108–111, 146–150, 165, 166, 182, 200, 210–212, 214, 227

ML Machine Learning. 2, 11–20, 26–31, 33–38, 43, 45, 49, 56, 65, 72, 99, 100, 103, 172, 174, 175, 205

MLE Maximum Likelihood Estimation. 17, 18, 178

MLP Multi Layer Perceptron. 140

MPA Medial Plantar Artery. 217, 220, 222

MRI Magnetic Resonance Imaging. 33, 63, 172, 177, 194, 195

mRMR minimum-Redundancy Maximum Relevance. 102, 103, 108, 109

MSE Mean Squared Error. 25, 76, 79, 175, 181

MVC Minimum Volume Constraint. 140

NIR Near-infrared. xxiii, 115, 116, 127, 162, 163

NLP Natural Language Processing. 4, 20, 22, 24, 28, 173

NMF Non-negative Matrix Factorization. 139, 140

NMM Non-linear Mixture Model. 136, 138, 167, 168

NTR Normalized Temperature Ranges. 217, 219, 223, 225, 226

OC-SVM One-Class SVM. 175, 176, 186–188, 197, 201

OCSP Orthogonal Complements Subspace Projection. xxvi, xxvii, 117, 118, 212, 213

PCA Principal Component Analysis. 100, 101, 136, 142, 146, 168, 176

PDF Probability Density Function. 47, 49, 56, 174, 176

PHATE Potential of Heat-diffusion for Affinity-based Trajectory Embedding. xxvii, 117, 212, 213

PPI Pixel Purity Index. 139

RBF Radial Basis Function. 57, 58, 79, 128, 164, 213, 222

ReLU Rectified Linear Unit. 21, 90, 95, 151, 164, 185

RF Random Forest. 13, 29, 219, 220, 223, 225

RFE Recursive Feature Elimination. 102, 223

RKHS Reproducing Kernel Hilbert Space. 56, 57, 103

RMSE Root Mean Squared Error. 151, 153, 154, 157, 158

SAD Spectral Angle Distance. xix, xxv, 136, 151, 154, 157–161, 167

SAM Spectral Angle Mapper. 136

SGD Stochastic Gradient Descent. 23, 24, 34, 36, 182

SNR Signal-to-Noise ratio. xix, xxv, 139, 152, 154, 161, 165

SUnSAL Sparse Unmixing by Variable Splitting and Augmented Lagrangian. 157, 167

SVDD Support Vector Data Description. 175, 176

SVM Support Vector Machine. xxvi, 56, 102, 117, 125, 175, 212, 213, 217, 222, 223

SWIR Short-Wavelength Infrared. 39

TCI Thermal Change Index. 217, 219, 225, 226

VAE Variational AutoEncoder. 20, 28, 112, 114, 176–180, 182, 183, 197, 201, 214

VCA Vertex Component Analysis. xix, xxiv, xxv, 139, 144, 152–155, 157–159, 161, 163–165

VNIR Visual and Near-Infrared. 39

Chapter 1

Introduction

> Cuando toca decantar lentamente lo que siento, yo me impaciento. Luego lo vuelvo a intentar, tinta y tiempo
>
> *Jorge Drexler - Tinta y tiempo*

Artifical Inteligence (AI) has recently made substantial advances in perception (the interpretation of sensory information), allowing machines to better represent and interpret complex data. However, it is crucial to understand that machines are not genuinely intelligent; they are merely proficient at executing instructions.

When I began my university studies in computer engineering, a professor highlighted this distinction by pointing out that while computers excel at performing operations swiftly, they lack true intelligence. To illustrate this, he conducted an exercise where he asked a colleague to follow precise instructions to open the classroom door: “Take two steps forward, turn right, take three steps forward, turn the handle, and pull the door”. Consequently, the door struck the colleague’s feet and did not open. This example demonstrates how computers operate: they follow instructions meticulously without considering contextual factors. If a machine fails to perform a task correctly, it is because it has not been provided with accurate instructions. This raises the fundamental question: what constitutes machine intelligence?

Although there are many formal and informal definitions of ‘intelligence’, in the context of AI research, there are two main perspectives [1]: one that emphasizes task-specific skill (achieving goals) and another that focuses on generality and adaptation. The second perspective, inspired by Turing’s work [2], is especially relevant for AI as it focuses on how machines can learn new skills. Following this approach, Friedberg provides an intuitive understanding of AI [3]:

> *If we are ever to make a machine that will speak, understand or translate human languages, solve mathematical problems with imagination, practice a profession or direct an organization, either we must reduce these activities to a science so exact that we can tell a machine precisely how to go about doing them, or we must develop a machine that can do things without being told precisely how.*

AI consists of algorithms that exhibit ‘intelligence’ because someone has designed instructions to describe the behavior of the machine (which will strictly follow these instructions), such as the commonly used algorithms in video games, and algorithms aimed at learning from data. This latter group of AI algorithms is known as Machine Learning (ML). While more details will be covered in Chapter 2, it is important to clarify that these models are designed to perform tasks without being explicitly told how to do so; they must learn to accomplish these tasks autonomously.

There is a pragmatic reason why there has always been an interest in AI: facilitating our lives. For instance, the application of logic and statistical pattern recognition to problems in medicine has been proposed since the early 1960s [4], focused on creating systems that could replicate the decision-making processes of human experts. However, these 'simpler' methods faced significant limitations in handling the complexity and variability of real-world medical data.

Over the past few decades, advances in ML have revitalized the field of medical AI, proclaiming that AI can diagnose a range of diseases. The implication is often that they can reduce the workload of the specialist and accomplish objectives with minimal manual intervention [5]. These AI models can outperform traditional statistical methods and even obtain a performance on par with healthcare professionals [6]. However, there is a trend towards using Deep Learning (DL) models, whose details are explained in Section 2.2, characterized by their complexity, being complicated how these models determine their decisions.

The adoption of AI in critical fields like medicine has raised several ethical and practical concerns. A critical issue is AI paternalism, where decisions made by AI systems could potentially override human judgment, leading to a loss of autonomy for users. For instance, in healthcare scenarios, this paternalistic potential leads to several ethical issues [7] such as implicit biases in the (quantified) notion of health and potential privacy concerns. Additionally, the use of more complex models, sometimes driven by the myth that there is necessarily a trade-off between accuracy and interpretability [8], poses another critical issue. Using extremely complex models that we cannot understand can contribute to problems associated with AI paternalism.

As an engineer, in this thesis I aim to provide insights for enhancing the interpretability of AI models, specifically DL models, to mitigate some critical issues associated with their use in essential fields. To achieve this, the key point is *information*.

1.1 Information is All You Need

Nowadays, *information* plays a central role in society, and almost every scientific discipline uses the concept of information within its context. The analysis of *information* is fundamental for data processing. However, while the meaning of *information* might seem intuitive, its absolute definition is extremely complex. This complexity is well-reflected in the skeptical view expressed by Bodgan [9]:

> *My skepticism about a definitive analysis of information acknowledges the infamous versatility of information. The notion of information has been taken to characterize a measure of physical organization, a pattern of communication between source and receiver, a form of control and feedback, the probability of a message being transmitted over a communication channel, the content of a cognitive state, the meaning of a linguistic form, or the reduction of an uncertainty.*

This skepticism arises from focusing on the meaning of the data, but this 'meaning' is less evident

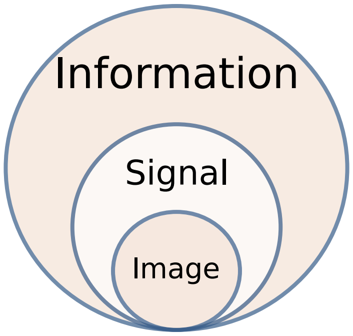


Figure 1.1: Different level of abstraction that defines the way that data is processed. The information corresponds to the highest level of abstraction, showing that the information processing is more general.

when viewed more abstractly. For this reason, as Warren Weaver claimed [9], *"information must not be confused with meaning"*.

Shannon's Information Theory (IT) [10], which will be introduced in Chapter 3, provides a perspective where the meaning of a transmitted message is irrelevant. Instead, the focus is on quantifying the information contained in that message. This theory simplifies the information and quantifies it using 'bits'. Although Shannon did not consider the implications for physics and physicists did not initially need 'bits', this approach offered a new perspective for understanding the world. As John Archibald Wheeler, the last collaborator of Einstein and Bohr, expressed [11], "*It from Bit*", indicating that information is fundamental and every 'it' originates from information, the 'bits'.

The main benefit of processing information is abstraction, a key concept for a general viewpoint. It is assumed that data contains information, and processing data aims to extract that information. However, in AI, some algorithms are designed exclusively for working with specific data types. This specificity is particularly relevant if you are interested in the 'meaning' of the information contained in the data, but it comes with constraints, such as the versatility of the algorithms and the performance achievable for complex tasks.

In image processing, for instance, it is assumed that information is contained within the images to be processed. Consequently, image processing involves acquiring images or extracting features of interest from them. These features could be designed to describe and characterize data with a specific structure, making them unsuitable for data with different structures; in other words, they are applicable only to images. Moreover, these features might be tailored for specific domains within the images, limiting their applicability to images from different contexts or image modalities.

As a result, a natural abstraction when working with images is to consider them as signals. In this way, a signal noise reduction algorithm can be easily adapted to different structures, such as images. However, the analysis of information is even more abstract than signal processing. Assuming that data contains information, the information analysis can be applied to any data type. Figure 1.1 illustrates a Venn diagram depicting these different levels of abstraction in data processing.

The task of creating an intelligent machine is complex. Describing the instructions for detecting, for instance, a dog in an image is challenging. Developing an AI system for general-purpose object detection is even more complicated. Currently, these tasks are carried out through AI that it can learn from data. However, this learning process has been described at a high level of abstraction, such as describing the learning process based on information processing. As a result, AI extracts features that contain information, but the 'meaning' of this information might not be understandable from our perspective.

The substantial advances in AI are a consequence of the increasing importance of information. The amount of information is growing, sometimes excessively, driven by advances in telecommunications technologies and computer science. "Too much information" is a phrase that likely resonates with many, reflecting the exhaustion that can be experienced today. It is intuitive to understand that your data contains information that is not relevant for a specific task, it is specially relevant in high-dimensional data. Processing this vast amount of information is exhausting, and designing a well-performing ad-hoc algorithm based on 'understandable' features is unapproachable. For this reason, it is necessary that AI to work at a higher level of abstraction.

As will be discussed in future chapters, in general, DL models outperform other AI models because they can operate at different levels of abstraction. This capability arises because these models process information rather than just images or signals. They can extract 'useful features' that contain the necessary information from the data, even though these features can be difficult to understand. The term 'useful features' refers to the fact that your data might contain a lot of information that is not relevant to your task. These irrelevant features are discarded by the model, which only extracts features that are pertinent to your task.

As a result, it can be stated that DL models, these extremely complex models, can process information more effectively than other AI methods. Therefore, for a more in-depth understanding of the problem-solving mechanisms of these methods, it is necessary to analyze how the information is processed in these models.

1.2 Motivation

Recently, the impact of DL has been particularly notable in various aspects of our lives. In 2022, generative models based on DL, such as DALL-E [12] and later advancements like Midjourney [13], transformed image generation from text, showcasing significant improvements. This transformative trend continued into 2023, marked as the year of AI for text generation with innovations like ChatGPT [14].

The dominance of DL models extends across different domains, notably in fields such as Computer Vision (CV) or Natural Language Processing (NLP), where classical methods have taken less relevance, and many of these traditional techniques can be considered 'outdated'. The superiority of DL models lies in their easily adaptable framework, surpassing ad-hoc methods. Leveraging available data, these models offer end-to-end solutions that are considered user-friendly, even for individuals who may not be considered experts.

Considering that the impact of DL applications in society is unquestionable, in my opinion, there are still many gaps to solve to get a total integration of these models in our lives. The current technological advances by DL model are demonstrating that many processes and tedious tasks, current carried out by humans, will be automatized. However, the emergence of ethical problems associated with AI-based automation will depend on how these models are integrated into society.

Currently, these problems have arisen in the community. The utilization of images or text generated by AI is causing certain issues that have not been adequately considered yet. For instance, misinformation creates a situation where users cannot identify those images or text that have been generated by AI, resulting in people accidentally sharing fake news. Additionally, it raises other concerns associated with copyright, as these models can be trained on data that original authors have not consented to. These problems might be caused because these tasks that have been addressed by AI are commonly associated with creative tasks that traditionally can only be carried out by humans.

The current applications where the AI is being specially relevant are not essential in our lives; most of them are associated with entertainment. Excluding the non-technical reasons, there is a specific reason why AI-based on DL is still not considered for integration into crucial fields: those models are non-reliable. A Deep Neural Network (DNN), the fundamental DL model, is still considered a 'black-box', as they are uninterpretable models where it is not possible to understand what is happening in the hidden layers.

As example, autonomous driving and medical screening are critical application fields where being inaccurate implies disastrous consequences. The output of DL models has higher uncertainty than the output obtained by classical, and simpler, methods. It is easier to understand classical methods, and in complex models like DNNs, it is much more complicated to understand why a small change in the input can drastically affect the model's performance. In this way, classical methods might be considered more robust.

Considering the medical fields, for instance, the regulation of future therapies using DL-powered AI is linked to reducing the uncertainty produced by these models. For this reason, opening the 'black-box' is a fundamental task for the integration of this technology in our lives.

In addition, several advantages will come with this approach aimed at reducing the uncertainty of these models. This task will lead to a more intuitive understanding about the problem-solving mechanism of these models. As a result, architecture design based on these understanding can obtain more efficient architectures.

1.3 Hypothesis

It is evident that the use of AI algorithms based on DL is specially relevant in contemporary scientific research, with many authors incorporating these algorithms to solve previously unapproachable problems. However, the inherent lack of 'transparency' in these models introduces a high level of uncertainty due to their complex problem-solving mechanisms. This uncertainty raises concerns about their application in critical scenarios.

Considering these factors, the balance between capacity and interpretability becomes pivotal in choosing suitable methods for different applications. Sacrificing model accuracy is often a conscious choice to enhance the interpretability of the obtained results, thus mitigating uncertainty.

In the light of these observations, the hypotheses of this dissertation are encapsulated in the following questions:

1. Is it possible to improve the interpretability of DL models?

2. Can the intuition regarding the mechanisms of these models be formalized within a common framework?

These questions will be addressed throughout this dissertation, with the results providing insights into various topics. While the versatility of DL models allows for the exploration of these questions in diverse domains, this dissertation primarily focuses on medical topics. The medical field is particularly critical, and there is ongoing debate about integrating DL models into it.

1.4 Objectives

This Ph.D. thesis aims to enhance the interpretability of DL models, providing a more profound understanding of how these models work. This Ph.D. thesis addresses various problems commonly encountered in the medical context where AI can be beneficial. Different methods are proposed in various chapters to achieve consistent results for these problems while offering a perspective on the problem-solving mechanisms employed by DL models.

The specific problems addressed in this dissertation are:

- **Medical screening with scarce dataset:** This is a common challenge in the medical context where data are often limited, yet DL models typically require large labeled datasets. The goal is to describe the expected features of the DL model, based on theoretical approaches, and determine if it is possible to assess whether the model is generalizing effectively, despite the limited number of samples.
- **Feature Selection in medical data:** This task is particularly relevant for enhancing data interpretation and reducing variability in clinical decision-making. This thesis aims to provide an interpretable DL solution to this task, including a theoretical description of the approach.
- **Signal decomposition:** Although this task will be focused on a specific image modality, HyperSpectral Imaging, which will be introduced in Section 2.6, the importance of signal decomposition for medical imaging will be discussed. An interpretable DL solution for addressing this problem will be provided.
- **Anomaly Detection:** This task has a fundamental advantage in that it does not require extremely comprehensive datasets. An alternative state-of-the-art approach using DL will be provided, along with a theoretical description to enhance the interpretability of the model.

This thesis aims to reduce the gap found in the use of DL-powered AI in critical fields such as medicine. Additionally, it seeks to emphasize the use of more interpretable models and to design more efficient architectures that consider the specific task being solved.

1.5 Collaborations and Acknowledgments

This Ph.D. thesis presents the outcomes achieved during the close collaboration between the Institute for Applied Microelectronics (IUMA) and the Research Institute of Biomedical and Health Sciences (IUIBS) of the University of Las Palmas de Gran Canaria (ULPGC) and the following research institutions:

- IACTEC Medical Technology Group, Institute of Astrophysics of the Canary Islands (Spain)
- Exploratory Oncology Research and Clinical Trial Center, National Cancer Center (Japan)

- Department of Electrical and Computer Engineering, Tokyo University of Science (Japan)

In addition, this research was conducted as part of the ITHaCA (Hyperspectral Identification of Brain Tumors) project, funded by the Canary Islands Government under Grant Agreement ProID2017010164.

Finally, this Ph.D. thesis has been co-financed by the Canary Islands Agency for Research, Innovation and Information Society of the Ministry of Universities, Science, Innovation and Culture and by the European Social Fund Plus (FSE+) Integrated Operational Program of the Canary Islands 2021-2027, Axis 3 Priority Topic 74 (85%).

1.6 Thesis organization

This dissertation is structured into three main parts: Part I provides a comprehensive literature review, Part II presents the original research contributions of this Ph.D. thesis, and Part III includes the appendix. A brief explanation of each chapter is presented below.

- ***Part I: Literature Review and Modern Practices***
  - **Chapter 2: Preliminar — Literature Review.** This chapter lays the groundwork by presenting essential concepts and the current state-of-the-art in the field, offering a foundation for understanding subsequent research contributions.
  - **Chapter 3: Information-Theoretical Framework.** The fundamentals of the IT, a theory of communication that operates within a probabilistic framework, are introduced in this chapter. This theory will be fundamental for achieving a general-purposed interpretability of the DL models.
- ***Part II: Research Contribution***
  - **Chapter 4: Information-theoretical Evaluation in Deep Learning.** This chapter introduces an IT perspective for improving the interpretability of the problem-solving mechanism used in DNN. A methodology for analyzing DNNs with limited datasets is introduced and validated within this framework.
  - **Chapter 5: Feature Selection with Deep Learning: Hyperspectral Band Selection.** This chapter proposes a Feature Selection (FS) method based on DL, aimed at enhancing data interpretability. The mechanism of this model is analyzed and validated using the IT approach outlined in the previous chapter.
  - **Chapter 6: Signal Decomposition: HyperSpectral Unmixing.** Focusing on HyperSpectral Imaging, this chapter introduces an innovative method for HyperSpectral Unmixing based on DL models. Additionally, this method is described and validated through the lens of IT.
  - **Chapter 7: Anomaly Detection by Deep Learning.** This chapter addresses Anomaly Detection in medical imaging using a parametric statistical approach with a generative DL model. As in previous chapters, its problem-solving mechanism is examined using the IT perspective.

  - **Chapter 8: Conclusions and Future Lines.** This chapter concludes the work presented in this dissertation by summarizing the main contributions of this thesis and expressing the importance of striving for more interpretable models. Potential areas for future research are also identified.

- ***Part III: Appendix***
  - **Appendix A: Supplementary Material**. This appendix contains the supplementary material related to the research presented in Part II.
  - **Appendix B: Feature Ranking for Diabetic Foot Ulcers Thermograms**. This section presents the results of the proposed FS method applied to different data modalities. The method provides a ranking of features extracted from thermographic images using state-of-the-art techniques.
  - **Appendix C: Code Availability**. This appendix provides the location of the repositories containing the code necessary for the reproducibility of this Ph.D. thesis.
  - **Appendix D: Resumen en español**. A brief summary of the dissertation is provided in Spanish in this appendix.

Part I

Literature Review and Modern Practices

Chapter 2

Preliminar — Literature Review

> The grand aim of all science is to cover the greatest number of empirical facts by logical deduction from the smallest number of hypotheses or axioms.
>
> *Albert Einstein (1950)*

This chapter is designed to provide the foundational insights by presenting essential concepts and the current state of research in the field. It covers both foundational and advanced topics in ML, which are crucial for a more in-depth understanding of these models, with particular focus on the learning process. The discussion will naturally transition into the specialized domain of DL, explaining key concepts and prevalent concerns associated with these technologies.

A significant focus of this thesis is the interpretability of DL models. This chapter examines how this critical aspect is addressed in contemporary literature, providing context for the methodologies applied in this work. Additionally, for clarity on the challenges discussed throughout this dissertation, there is an exposition on the nature of the inverse problem and the concept of sparse representation.

Lastly, although this thesis adopts a broad and general perspective, a brief introduction to HyperSpectral Imaging (HSI) is included. Since HSI is a specialized imaging modality referenced in various chapters, this section ensures that all readers, regardless of their previous exposure to the technology, gain a fundamental understanding of it.

2.1 Artificial Intelligence and Machine Learning

Artifical Inteligence encompasses algorithms aimed at replicating or 'mimicking' human-like intelligence, with a recent surge in interest specifically in ML [15]. AI algorithms are designed to solve specific problems intelligently, often through heuristic rules designed for those problems. An example of such an algorithm is the Minimax algorithm, which makes decisions by minimizing potential losses in worst-case scenarios. This algorithm is widely used in strategic games like chess or tic-tac-toe [16] and it is necessary to define a score-rule based on the state.

Machine Learning, which is a subset of AI originally termed by Arthur Lee Samuel in 1959 [17], has a specific feature that can be deduced from the name, 'learning'. This term implies the acquisition of knowledge. For humans, learning occurs through study or experience. Similarly, in ML, machines acquire knowledge or information through this learning process, enabling them to perform tasks or make decisions based on learned patterns within the data, rather than relying solely on predefined rules or instructions. Essentially, in ML, data describe the task and the expected solution, but the machine learns how to accomplish it by observing and analyzing the data

The widespread adoption of ML techniques in contemporary times can be attributed to the abundance of available data for different topics, a direct consequence of the interconnectedness and digital recording of information [15]. This is a direct result of the 'Information Age', also known as the Third Industrial Revolution, which began at the middle of the XX century, linked to the development of the transistor in 1947 [18, 11]. Nowadays, it can be considered that we are in the Fourth Industrial Revolution [19, 15], motivated primarily by the current AI revolution.

The primary challenge with data in ML is the lack of explicit information about the specific task to be solved. For instance, consider a ML model tasked with identifying a dog from an input image. The model must extract relevant features from the data (the image) that enable it to identify the presence of a dog. However, the data may contain irrelevant information, such as background scenery like a park, which is not directly related to the task of dog identification. Due to this reason, the majority of the problems addressed in ML aim to extract features from the data that are utilized to address the corresponding issue.

Although ML models are designed to learn from data, it is worth noting that the task to carry out is fundamental for the design of the ML algorithm. Mitchel [20, 21] provides the following definition of the learning process:

> *A computer program is said to learn from experience E with respect to some class of tasks T and performance measure P, if its performance at tasks in T, as measured by P, improves with experience E.*

Considering this definition, there are three elements for the description of the type of ML model: the task T, the performance measure P, and the experience E [21]. Based on these three elements, it can be defined the type of ML algorithm.

2.1.1 Types of Machine Learning Techniques

In accordance with Mitchell's definition of ML [20], a wide variety of ML algorithms can be categorized based on three fundamental elements. While providing an exhaustive description of ML methods is beyond the scope of this dissertation, this section aims to offer an intuitive overview for understanding the taxonomy of ML methods. Figure 2.1 illustrates a simplification of the taxonomy of ML.

The experience E is pivotal for categorizing ML algorithms. Experience describes how AI systems learn, and it can be classified into four main categories:

- **Supervised learning**: Supervised learning involves extracting features to describe the function that maps an input to an output, based on a labeled dataset [22]. In this approach, ML models are trained using a dataset that contains samples associated with labels or targets [21].

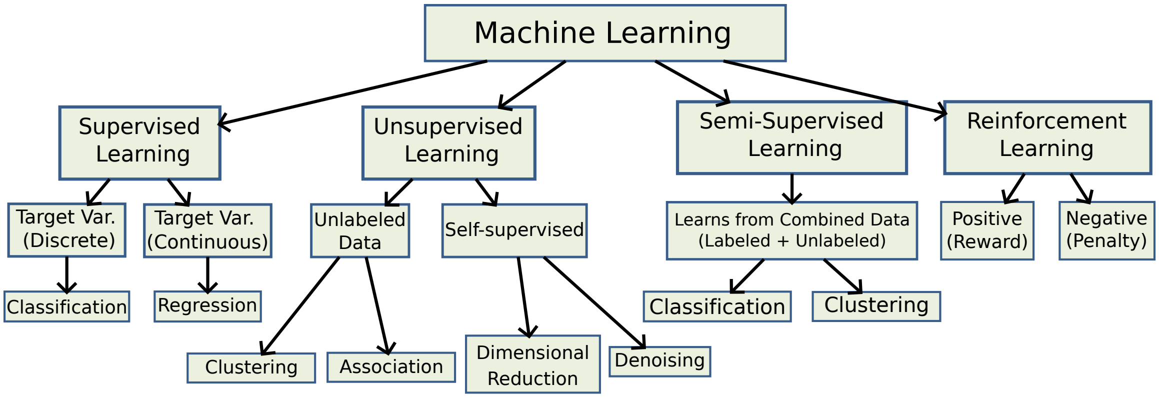


Figure 2.1: Taxonomy of Machine Learning techniques by learning methods. Some tasks are indicated as reference.

- **Unsupervised learning**: unsupervised learning consists of analyzing unlabeled datasets without human intervention to extract features that describe the underlying structure of the data [21, 15, 22].
- **Semi-supervised learning**: this type of learning combines aspects of both supervised and unsupervised learning, operating on both labeled and unlabeled data [15, 22].
- **Reinforcement learning**: This learning process is a fusion of the trial-and-error 'law-of-effect' tradition in psychology, optimal control theory in engineering, the secondary reinforcement tradition in learning, and the use of decaying stimulus [23]. In ML, this approach involves a software 'agent' interacting with an 'environment' to automatically evaluate optimal behaviors, with rewards for desired behaviors and penalties for undesirable ones [24, 25]. In reinforcement learning, the dataset is not fixed but provided by the environment.

The task T to be performed by the ML algorithm is inherently linked to the learning process, it is determined by the available data and the nature of the learning process. Figure 2.1 illustrates various tasks associated with the learning process. Similarly, the choice of performance measure P is heavily dependent on the specific task being tackled and the model selected.

For instance, in classification tasks, the objective is for the machine to determine which of k categories some input belongs to. To accomplish this task, the learning algorithm typically aims to produce a function $f : \mathbb{R}^n \to \{1, \ldots, k\}$. Given this map, from $\mathbb{R}^n$ to $\{1, \ldots, k\}$ the data has to be indicated by a labeled dataset, indicating that it is a supervised method. However, the performance measure is carried out by the *loss function*, which minimization corresponds to a better performance.

It is worth noting that some ML algorithms, like Random Forest (RF), do not rely on a *loss function* to improve model performance [26]. Instead, RF employs a process of random selection to obtain subsets of the data, knowing as 'bagging', and generates different decision trees to tackle classification tasks. These decision trees are uncorrelated, and the final solution is determined by majority voting among the trees. In this case, there's no explicit loss function involved. Instead, the 'learning' process involves generating decision trees by partitioning nodes based on impurity metrics like entropy or the Gini index [26].

In unsupervised learning, where there are no predefined targets for describing the expected solution, the choice of the *loss function* depends on the specific ML algorithm being used. For

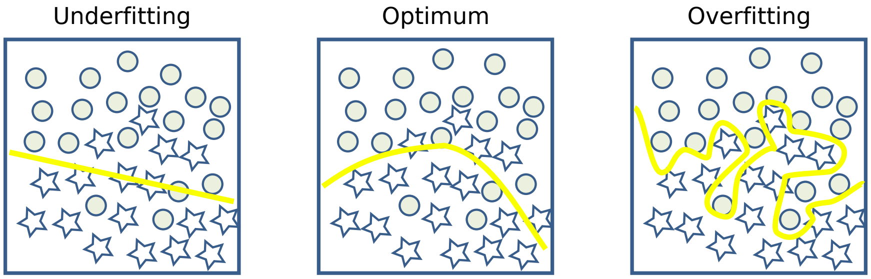


Figure 2.2: Example of from an underfitting model to an overfitting model, from left to right.

instance, in clustering, two commonly used ML approaches are the K-Means algorithm [27] and Expectation-Maximization (EM) algorithm using a Gaussian Mixture Model (GMM) [28], and the optimization processes for these approaches differ.

In the K-Means algorithm, there is defined a loss function that involves minimizing the squared distance of each data point to the closest hypersphere centroid [27]. As a result, the K-Means algorithm provides a model based on centroids. On the other hand, in a GMM, there is not a loss function in the traditional sense. Instead, the optimization of this model is performed using the EM algorithm, which entails maximizing the log-likelihood of the data given the model parameters, the mixture of Gaussian distributions [28]. However, this log-likelihood plays a similar role in guiding the optimization process and can be interpreted as a loss function aimed at minimizing negative log-likelihood. These loss functions guide the learning process in each algorithm, enabling them to identify meaningful patterns or structures in the data without the need for labeled targets.

In semi-supervised learning, which is commonly used for Anomaly Detection (AD) [29, 30], the approach is considered a classification task, as depicted in Fig. 2.1. Here, the algorithm utilizes a subset of labeled data, which is significantly smaller than the unlabeled data. In this type of task, the computer program sifts through a set of events or objects and flags some of them as being unusual or atypical [29].

On the other hand, reinforcement learning is frequently employed for training ML algorithms in controlled environments such as video games or robotics [25]. In reinforcement learning, an agent interacts with an environment and learns to take actions to maximize some notion of cumulative reward. This approach is particularly useful when the model needs to learn through trial and error, refining its actions based on feedback received from the environment.

2.1.2 Overfitting and Underfitting

The primary challenge in ML, especially in supervised learning, is achieving good performance on unseen inputs — that is, inputs that were not used during the training process. The goal is to obtain a model that can generalize well to new, unseen data and effectively solve the task at hand.

Two common problems encountered in ML are underfitting and overfitting. Underfitting occurs when the learning model is too simple and cannot capture the underlying relationships present in the data [31, 32]. This typically happens when the model has insufficient parameters to adequately represent the complexity of the data and capture the nuances of the input-target relationships. On the other hand, overfitting occurs when the model is overly complex, with an excessive number of parameters, and essentially memorizes the training data without truly learning the underlying patterns [31, 32]. As a result, an overfit model may perform well on the training data but fail to generalize to new, unseen inputs. Both underfitting and overfitting lead to poor performance on unseen inputs. Figure 2.2 illustrates both problems, underfitting and overfitting.

In practice, evaluating ML models often involves dividing the dataset into a training set, used exclusively for training, and a test set [21, 31, 32]. This approach allows us to estimate the model's performance on unseen data. However, certain assumptions must be considered when using this method, such as the i.i.d. (independent and identically distributed) assumption, which implies that the training and test sets are drawn from the same distribution. Additionally, independence among instances is assumed.

By using separate datasets, the expected performance of the model can be assessed. Underfitting occurs when the model fails to achieve a sufficiently low error (as indicated by the loss function) on the training set. On the other hand, overfitting happens when the model performs well on the training set but poorly on the test set, indicating that it has learned to memorize the training data rather than generalize to new data.

Validation in Machine Learning

It is commonly to find a third subset of the dataset, the known validation set. This set is often employed during the training for obtaining the optimal parameters of the ML model based on the validation error. This validation set, which contains samples with known provenance, is not applied for training the parameters of the model [33], it is exclusively for evaluating the performance of the model using as it as unseen inputs. As with the test and training set, the i.i.d. assumption is considered.

In addition, given that the number of samples in the dataset is limited, the K-Fold cross validation technique [34, 35] is commonly used to validate the algorithm design, not a particular training. This is specially relevant for a more in-depth understanding of the possible *hyperparameters*, those parameters that are not trainable of the model, and how it affects the model. This method involves randomly dividing the dataset into k groups, or folds, of approximately equal size. One fold is used as the validation set while the model is trained on the remaining $k-1$ folds. This process is repeated k times, with each fold serving as the validation set exactly once [35].

2.1.3 The Relevance of Probability in Machine Learning

Probability plays a fundamental role in ML algorithms due to the nature of ML tasks, which often involve dealing with uncertain or stochastic quantities. For instance, in the previous section, the challenges of overfitting were discussed, and probability theory provides a fundamental approach to address this issue. Even when ML models aim to produce consistent outputs for the same input, there can still be inherent uncertainty caused by factors such as noise in the data or randomness in the learning process. Therefore, probability theory provides a framework for reasoning about uncertainty and making probabilistic predictions or decisions in ML tasks. In essence, as famously expressed by Laplace:

> *Probability theory is nothing but common sense reduced to calculation.*

Uncertainty about decisions is a fundamental concept in ML. In this context, uncertainty in ML can arise from three main sources [21]:

1. Stochastic model: The system being modeled exhibits inherent stochastically, meaning that its behavior cannot be predicted with certainty.

2. Incomplete observability: Even in deterministic systems, uncertainty can arise when it is not possible to observe all variables, leading to apparent stochastic behavior.

3. Incomplete modeling: ML models may not fully utilize all the information provided by the data, resulting in uncertainty in the model's predictions. This occurs when relevant information is discarded or overlooked during the modeling process, see Section 2.1.2.

It is important to note that not all ML models are inherently probabilistic. Some ML models are deterministic, assuming that the data is known with certainty and that there is no randomness involved in the learning process. However, the probabilistic approach to describing ML algorithms offers a more versatile framework. By incorporating probability theory, even deterministic models can be generalized to account for uncertainty, being a special case of probabilistic models. This allows for a more comprehensive understanding and handling of uncertainty in ML tasks.

Problem Statement — Bayesian Approach

Considering the supervised learning case, see Section 2.1.1, for the problem statement, there is a dataset with samples $X \in \{x_1, \ldots, x_N\}$ and targets $Y \in \{y_1, \ldots, y_N\}$, where N is the number of samples or elements. As a consequence, the ML algorithm has to define a function $f : X \to Y$, i.e., a function that maps X into Y. The parameters of the ML model are denoted by ω, so $\hat{y} = f(x; \omega)$ where $x \in X$. Following the Bayesian approach, the supervised problem is described as follows:

$$p(\omega|X, Y) = \frac{p(Y|X, \omega)p(\omega)}{p(Y|X)} \propto p(Y|X, \omega)p(\omega), \tag{2.1}$$

where $p(Y|X, \omega)$ is the *likelihood*, $p(\omega)$ is the prior distribution, and $p(\omega|X, Y)$ is known as posteriori distribution. The nature of the target data is the condition that allows to differ this supervised learning problem between, for instance, a classification (Y as discrete variable) or regression (Y as continuous variable).

As it has been previously indicated, the loss function is the measure considered for training the model. This loss function arises naturally based on the maximum a posteriori probability in Bayesian statistics. Based on this loss function, the problem to solve is indicated by:

$$\underset{\omega}{\operatorname{argmin}}\ \mathcal{L}(\omega; X, Y) = \underset{\omega}{\operatorname{argmin}}\ -\log p(Y|X, \omega)p(\omega) \Leftrightarrow \underset{\omega}{\operatorname{argmax}}\ p(\omega|X, Y), \tag{2.2}$$

where $\mathcal{L}(.)$ is the loss function that corresponds to the negative log-likelihood. The use of maximization corresponds to the Maximum A Posteriori estimation (MAP). Considering, for instance, in regression problems in ML, using for instance a linear model, is intuitively considering the minimization of the distance $d(\hat{y}, y)$ as a loss function, so the solution is obtained by finding the parameters that minimize this distance. If you consider the use of Gaussian Process (GP) [36], which follows a Bayesian approach, for a simple case where it assumed that the data follows a Gaussian noise, likelihood is described as follows:

$$-\log(p(Y|X,\omega) = -\log\left(\frac{1}{(2\pi\sigma_1^2)^{\frac{N}{2}}}\, e^{\left(\frac{-||Y-f(X;\omega)||_2^2}{2\sigma_1^2}\right)}\right) \propto ||Y-f(X;\omega)||_2^2, \tag{2.3}$$

where σ corresponds to the standard deviation. As it can be observed, the squared Euclidean distance ($||.||_2^2$) appears in the equation for solving the regression problem. Additionally, the prior distribution with the aforementioned assumption is described as:

$$-\log(p(\omega)) = -\log\left(\frac{1}{(2\pi\sigma_2^2)^{\frac{K}{2}}}\, e^{\frac{-\|\omega\|_2^2}{2\sigma_2^2}}\right) \propto \|\omega\|_2^2 \tag{2.4}$$

where ω is composed by K parameters. In this specific case, it is assumed that $\omega \sim \mathcal{N}(0;\sigma_2)$, so the parameters will tend to 0.

The specific case indicated in Eq. (2.3) is describing a regression problem. In the case of a classification task, with discrete variables, the $-\log p(Y|X,\omega)$ is associated to the Cross-Entropy (CE), which will be defined in Section 3.1.2. In practice, the categorical CE is commonly used and is expressed as follows:

$$CE(\hat{y};y) = -\sum_{c} \mathbb{1}(y=c)\log(p(\hat{y})), \tag{2.5}$$

where $\hat{y} = f(x;\omega)$, $\mathbb{1}(\cdot)$ is the indicator function and $p(\hat{y})$ denotes the softmax probability of sample x (*Softmax:* $\mathbb{R}^d \rightarrow [0,1]$). The ground truth label corresponds to y and c denotes the class category.

The unsupervised problem follows the same description but removing the target variable Y. In this way, the Bayesian approach for an unsupervised learning problem is described as:

$$p(\omega|X) = \frac{p(X|\omega)p(\omega)}{p(X)}. \tag{2.6}$$

As happens with the supervised problem, the optimization corresponds to MAP estimation depicted in Eq. (2.2).

Maximum Likelihood Estimation

In most ML models, the problem is optimized by using Maximum Likelihood Estimation (MLE). Unlike MAP estimation, here, the parameters of the model, ω, are not considered random variables. In this approach, the loss function using MLE exclusively depends on the likelihood maximization:

$$\underset{\omega}{\text{argmin}}\ \mathcal{L}(\omega;X,Y) = \underset{\omega}{\text{argmin}} -\log p(Y|X;\omega) \Leftrightarrow \underset{\omega}{\text{argmax}}\ p(Y|X;\omega). \tag{2.7}$$

An alternative to the notation $p(Y|X;\omega)$ would be $p_\omega(Y|X)$, indicating that the parameters are not considered random variables. The expression depicted in Eq. (2.7) can be adapted for unsupervised approaches, as happens in the Bayesian approach.

This approach is a particular case of the Bayesian estimator described in Eq. (2.1). The estimator obtained by MLE coincides with the most probable model given a uniform distribution of the parameters ω. In other words, it is equivalent to setting $p(\omega)$ to a uniform distribution in Eq. (2.1). Using this uniform prior distribution corresponds to not applying any constraints in the solution space, i.e., all solutions are equiprobable. An interpretation of this approach is that the model does not have any prior assumption about which solution is more probable, there is no information. This lack of prior assumptions means that the MLE approach relies entirely on the data to determine the best-fitting parameters, without any additional bias or preference for particular parameter values.

MLE involves using probabilities for optimization. However, the use of probability is not inherent in all ML models, as some model's optimization processes are not described using probabilities. A good example of this is the unsupervised method previously mentioned: the centroid-based model obtained by the K-Means algorithm and GMM used by the EM algorithm.

The model obtained by K-Means algorithm is optimized by minimizing its loss function, defined as the sum of the squared distances between each data point and the centroid of its assigned cluster. In contrast, GMM used by the EM algorithm is obtained by maximizing the likelihood. In the case of GMM, the use of MLE is evident, but the centroid-based model obtained by K-Means algorithm can also be described within this framework. Specifically, the model obtained by K-Means can be considered a special case of GMM where the data follow isotropic Gaussian noise (resulting in 'spherical' solutions), solving the problem described by the MLE in Eq. (2.7). In other words, while the likelihood in GMMs is defined by $\mathcal{N}(\mu, \Sigma)$, where Σ is the covariance matrix, the likelihood of the deterministic model obtained by K-Means could be expressed as $\mathcal{N}(\mu, \sigma I)$, where I is the identity matrix.

This demonstrates that while not all ML models inherently use probabilities for optimization, they can be reinterpreted within a probabilistic framework.

2.1.4 Bayesian Inference

Bayesian inference corresponds to the branch of ML where Eq. (2.1) or Eq. (2.6) are applied directly. Consequently, it is possible to quantify the uncertainty about the unknown parameters of the model by the posterior distribution, which describes the parameters ω after observing the training set.

For an easier annotation, it will be considered the unsupervised task depicted by Eq. (2.6). Once it is trained the ML model with parameters ω and the training set X, to obtain the confidence of $p(x')$ where $x' \in X'$ and X' is a subset of test, it necessitates marginalization, which is achieved through:

$$\hat{p}(x') = p(x'|X) = \int_{\omega} p(x'|\omega)p(\omega|X)d\omega. \tag{2.8}$$

However, this marginalization is often computationally intractable, especially in high-dimensional problems. Consequently, $p(x'|X)$, denoted as the marginal distribution $\hat{p}(x')$, provides the uncertainty of the unobserved event x' using the parameters ω trained by the training set X.

In order to achieve a feasible solution in practice, it is imperative to consider some approximate inference methods to compute the integral described in Eq. (2.8). Variational inference plays a fundamental role in achieving this goal.

Variational Inference

Variational inference is a technique employed to approximate complex probability distributions when exact inference is computationally expensive or infeasible [37]. The idea is to replace the posterior distribution with an easy-to-sample approximate parametric distribution, $q_\phi(\omega|X) \approx p(\omega|X)$.

As it has been previously described, the objective of the model is to minimize $\log p(X|\omega)$ but in this case, it is necessary to solve directly the marginal distribution $\hat{p}(x)$. Considering the $x \in X$, $p(x) = \int_\omega p(x|\omega)p(\omega)$, the idea is to include this new distribution $q_\phi(\omega|x)$:

$$\hat{p}(x) = \int_\omega q_\phi(\omega|x)\frac{p(x|\omega)p(\omega)}{q_\phi(\omega|x)} = E_{\omega\sim q_\phi(\omega|x)}\left[\frac{p(x|\omega)p(\omega)}{q_\phi(\omega|x)}\right]. \tag{2.9}$$

Solving $\log \hat{p}(x)$ and following the Jensen Inequality [38], the problem follows:

$$\log \hat{p}(x) = \log\left(E_{q_\phi(\omega|x)}\left[\frac{p(x|\omega)p(\omega)}{q_\phi(\omega|x)}\right]\right) \geq E_{q_\phi(\omega|x)}\left[\log\left(\frac{p(x|\omega)p(\omega)}{q_\phi(\omega|x)}\right)\right]. \tag{2.10}$$

This expression is the Evidence Lower Bound (ELBO) [37]. Based on this bound, it can be obtained the next expression:

$$\begin{aligned}\log \hat{p}(x) \geq \int_w q_\phi(\omega|x)\left(\log p(x|\omega) + \log\frac{p(\omega)}{q_\phi(\omega|x)}\right) = \\ = E_{q_\phi(\omega|x)}\left[\log p(x|\omega)\right] - E_{q_\phi(\omega|x)}\left[\log\frac{q_\phi(\omega|x)}{p(\omega)}\right].\end{aligned} \tag{2.11}$$

Observing Eq. (2.11), it can be observed that the first expression, $E_{q_\phi(\omega|x)}\left[\log p(x|\omega)\right]$, corresponds to the log-likelihood and the second is the well-know Kullback-Leibler Divergence (KL Divergence) [39], denoted as $D_{KL}(.||.)$. The equality, and optimal solution, is only obtained when $D_{KL}(q_\phi(\omega|x)||p(\omega)) = 0$, i.e., when $p(\omega)$ fits perfectly to the approximate parametric distribution $q_\phi(\omega|x)$.

As a result, the loss function used for training is defined by:

$$\underset{\omega}{\text{argmin}} \; -E_{q_\phi(\omega|x)}\left[\log p(x|\omega)\right] + D_{KL}(q_\phi(\omega|x)||p(\omega)). \tag{2.12}$$

2.2 Deep Learning

Deep Learning is a subset of the ML paradigm, representing a specialized field within ML (see Fig. 2.3) where model architectures consist of multiple processing layers. These layers enable the model to learn data representations and extract features from various levels of abstraction [40, 41, 21]. DL is rooted in Artificial Neural Network (ANN), which are ML models inspired by biological learning processes and the structure of the human brain. For example, insights from the standard model of the visual cortex proposed by Hubel and Wiesel [42] suggest that the brain initially processes edges, then progresses to patches, surfaces, and finally identifies objects. This hierarchical representation enables the extraction of features at different levels of abstraction [40]. Although ANNs have been around since the 1940s, they were relatively unpopular for many years until experiencing a surge in popularity in recent times [21].

The term DL has a rich history, initially used by Rina Dechter in 1986 [43] in the context of efficiency in searching algorithms. However, contemporary use of DL to describe ML methods with multiple representation learning gained popularity around 2006 [21, 40]. The first precursor of DL was *cybernetics*, a field rooted in the idea of biological learning, which emerged in the 1940s with contributions from Norbert Wiener, who established the principles of control and communication in machines and living organisms [44], and the McCulloch-Pitts Neuron model, introduced in 1943 [45].

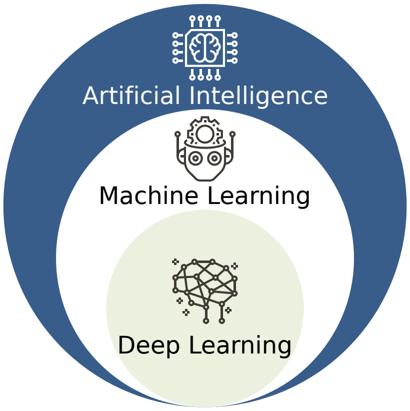
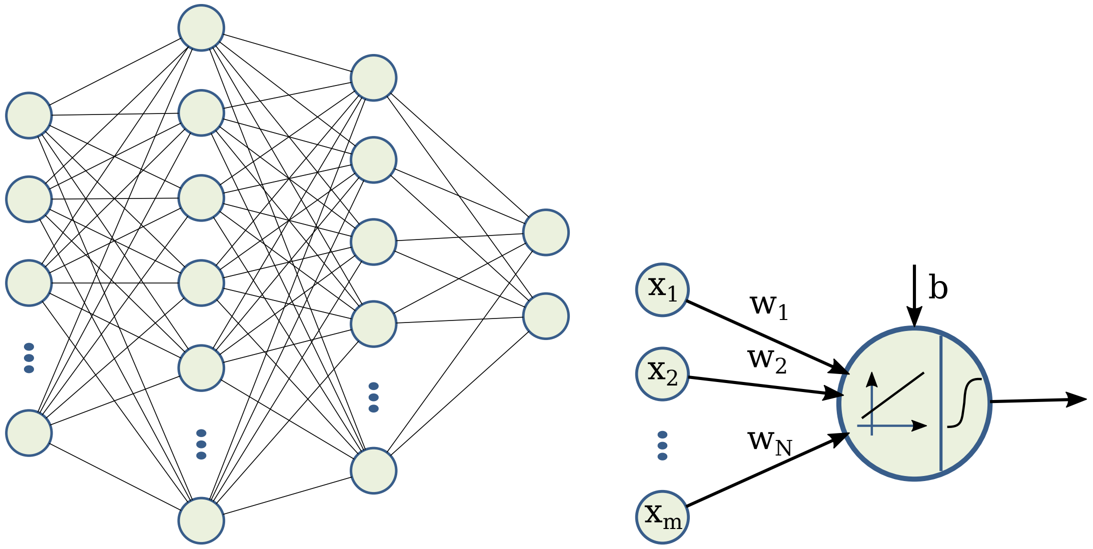


Figure 2.3: Venn diagram illustrating the logical relationships among artificial intelligence, machine learning and deep learning.

Cybernetics expanded until the 1960s, with Ivakhnenko and Lapa being notable contributors during this period [46]. Subsequently, during the 1980s to 1990s, *connectionism*, inspired by cognitive science, gained popularity [47, 21]. During this period, the concepts of ANN and *hidden layers* were introduced, enabling parallel signal processing across various branches of networks. During this second precursor of DL, significant figures such as Yann LeCun [48] and Schmidhuber [49] have made significant contributions. LeCun is widely regarded as a founding father of Convolutional Neural Network (CNN), heavily influenced by the visual cortex mechanism, and Schmidhuber's work on Long Short-Term Memory (LSTM).

The contemporary surge in popularity of DL has been significantly driven by the utilization of Graphic Processing Units (GPU), which dramatically reduces processing times for extremely DNN, models that have millions of parameters. LeCun's groundbreaking work demonstrated the advantages of using CNNs in CV tasks, achieving record-breaking performance in handwritten digit identification with the widely used MNIST database [48]. Notably, this implementation was conducted on Central Processing Unit (CPU), as GPUs lacked programmable data processing units, commonly known as shaders, until 2001 [50]. In 2006, Chellapilla et al. presented an early GPU-based implementation of CNNs, which was reported to be four times faster than its CPU counterpart [51]. However, predating this, Oh and Jung developed the first reported GPU-based DL model, which was a Fully-Connected DNN,which was reported to be 20 times faster than CPU-based implementations [52].

Recently, since around 2020, there has been a surge in interest in generative models within the field of DL. Generative models constitute a subfield of ML where models possess the ability to 'imagine' and create new entities, akin to a notable feature of human cognition [53]. Presently, some of the most powerful generative models rely on DNNs and can be employed for tasks such as image generation, as seen in Midjourney [13], or for NLP, exemplified by models like ChatGPT [14]. It is important to highlight that generative models are often formalized using Bayesian approaches, as discussed in Section 2.1.3. Leveraging current DNN architectures such as Transformers [54], along with generative approaches in DL such as Variational AutoEncoder (VAE) [55], Generative Adversarial Network (GAN) [56], and Diffusion Probabilistic Model (DPM) [57], enables the attainment of performance that surpasses that of any previous ML models.



(a)

(b)

Figure 2.4: Figure illustrating in (a) a Fully-Connected Neural Network with 2 hidden layers. In (b) a representation of a single neuron.

2.2.1 Neural Networks

As it has been previously mentioned, DL models refers to ANNs or DNNs. An example of a fully-connected DNN is illustrated in Fig. 2.4a. In this section, it is going to include a brief description of these models.

An ANN commonly consists of three specific components known as neurons, hidden layers, and activation functions. The neuron has the same function that neurons in our brains and are characterized by having multiple inputs and a unique output. A combination at the same level generates correspond to the layer. In the case of a fully-connected DNN, as the illustrated in Fig. 2.4a, all neurons have the same input, but the output is different. Finally, the activation function has the purpose of including the non-linearity in the model, describing the criteria used by each neuron to decide whether to activate the output. The most common activation function used is the Rectified Linear Unit (ReLU), but other commons activation functions are the hyperbolic tangent ($tanh : \mathbb{R} \to [-1, 1]$) or the sigmoid function ($sigmoid : \mathbb{R} \to [0, 1]$).

In Fig. 2.4b, it is illustrated a single neuron with m inputs. The neuron conducts a linear operation with a bias b, followed by the application of a non-linear activation function. Generalizing to n outputs (i.e., n neurons in the layer), the output of a layer given an input $x \in \mathbb{R}^m$ is determined by the parameters θ, which consist of $W \in \mathbb{R}^{n\times m}$ and $b \in \mathbb{R}^n$, along with an activation function represented by σ. The operation of the i^{th}-layer, f_i, can be expressed as:

$$f_i = \sigma\left(\begin{bmatrix} w_{1,1} & w_{1,2} & \dots & w_{1,m} \\ w_{2,1} & w_{2,2} & \dots & w_{2,m} \\ \vdots & \vdots & \ddots & \vdots \\ w_{n,1} & w_{n,2} & \dots & w_{1,m} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_m \end{bmatrix} + \begin{bmatrix} b_1 \\ b_2 \\ \vdots \\ b_m \end{bmatrix}\right) \tag{2.13}$$

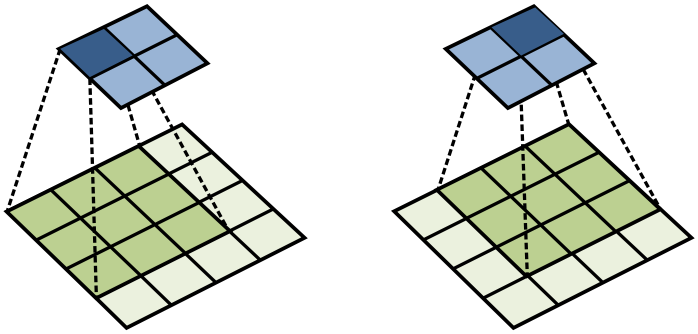


Figure 2.5: Convolution operation using a kernel size of 3 × 3 over 4 × 4 input.

This matrix operation, considering $\theta \in \{W, b\}$ corresponds to $f_i = \sigma(W^T X + b)$.

In this way, considering the example of Fig. 2.4a, the output corresponds to the sequence of the different functions f_i, where i is the index of the hidden layer. Therefore, the output of the DNN in Fig. 2.4a is expressed as $out = f_3(f_2(f_1(x)))$.

The use of DNNs have proven to be effective to solve problems in complex domains, such as supervised-learning for CV or NLP [40]. It is a consequence of the high number of parameters commonly associated with the DNN.

Convolutional Neural Network

The widespread use of CNNs in CV can be traced back to LeCun's proposal of this type of DNN for character recognition tasks [48]. The popularity of CNNs in contemporary CV is primarily driven by their excellent performance across various tasks, particularly highlighted by their success in beating the state-of-the-art in the ImageNet [58] image classification challenge [59].

Even CNNs are mainly attributed to LeCun, the first properly CNN was known as *Neocognitron* [60] proposed by Fukushima and Miyake in 1982. This *Neocognitron* was designed to detect pattern in images, inspired by the hierarchy in the standard model of the visual cortex proposed by Hubel and Wiesel [42].

The CNNs proposed to replace the linear operation $f_i = W^T X + b$, observed in the fully-connected DNN, by a convolution operation:

$$f_i = \sigma\left(K * X + b\right), \tag{2.14}$$

where $K \in \mathbb{R}^{k \times k}$ corresponds to the kernel, $X \in \mathbb{R}^{m \times m}$ represents the 2-dimensional input (the image of resolution $m \times m$) and $B \in \mathbb{R}^{(n+k-1)\times(n+k-1)}$ is the matrix of biases. The $\sigma(.)$ represents the non-linear operation.

In convolutional operations, there are various parameters to consider beyond just the kernel size (k) and the input image resolution. Other essential parameters include the 'stride', which determines how the convolutional operation moves across the image, and the 'padding'. For those interested in gaining an intuitive understanding of the interplay among different parameters in convolutional operations, I recommend consulting "A guide to convolution arithmetic for deep learning" by Dumoulin and Visin [61]. This comprehensive guide provides detailed explanations and visualizations that can elucidate how these parameters interact and affect the output of convolutional layers in DL models.

Convolutional layers typically have sparse interactions. Specific interaction between each output unit with every input unit is not found, and each convolutional kernel (or filter) is smaller than the input. In this way, the feature maps, which represent the locations in the input where features (represented by the filter) are present, contain spatial information.

Figure 2.5 provides a example of a convolution operation over a 2-dimensional input. The intuitive idea about this operation is related to the 'locality' by the sparse interactions. Specific interaction between each output unit with every input unit is not found, and each convolutional kernel (or filter) is smaller than the input. For this reason, what a convolutional operation proposes is to extract 'local' features, in this case spatial features which are representative of the subset of the input size. Depending on the 'stride', the displacement of the kernel through the image will make a difference. For instance, in Fig. 2.5, the kernel size is 3 × 3, requiring movement to four different positions to cover the entire input when the 'stride' is set to 1. As depicted, the output undergoes size reduction, a phenomenon often mitigated through the padding.

It is noteworthy that using a kernel size equivalent to the spatial resolution will discard the sparse interactions. As a result, this specific case fits to a fully-connected DNN, described in the previous section.

2.2.2 Optimization

The optimization of DL models is obtained by updating the gradient descent optimization method. As it was described in Section 2.1, the model is optimized by using a performance measure commonly known as loss function. The idea is to obtain the parameters that minimize this loss function, as it was described in Eq. (2.7). Based on the gradient descent, considering the parameters of the DL model denoted by θ, the first order optimally conditions are:

$$\frac{\partial \mathcal{L}(\theta; X, Y)}{\partial \theta} = 0. \tag{2.15}$$

DNNs are trained iteratively, updating step by step the parameters of the model based on the gradient. For instance, the parameters of the output layer of a DNN is depicted as:

$$\theta^{t+1} = \theta^{t} - \alpha \cdot \frac{\partial \mathcal{L}(\theta; X, Y)}{\partial \theta}. \tag{2.16}$$

The updating the weight in previous layers, it is carried out by using the *chain rule*. This process is known *back-propagation*, since the input is introduced in the model for obtaining the output and following the *chain rule*, the parameters of the layers are 'back-propagated'. The α in Eq. (2.16) is commonly known as *learning rate*.

Optimizers

The *optimizer* is a crucial concept in optimization problems, especially in DL, where the main optimizer is Stochastic Gradient Descent (SGD). Considering a supervised problem with the loss function defined as $\mathcal{L}(\theta; X, Y)$, SGD computes the gradient using a randomly sampled batch $x^b, y^b \in$

$\{X, Y\}$ of size B as follows:

$$\frac{\partial \mathcal{L}(\theta; X, Y)}{\partial \theta} = \frac{\partial}{\partial \theta} \frac{1}{B} \sum_{i}^{B} \mathcal{L}(\theta; x_i^b, y_i^b) \tag{2.17}$$

Optimization plays a critical role in DL, and alternatives to SGD have been proposed to achieve optimal model solutions. For instance, the first-order momentum, proposed by Sutskever et al. [62], is particularly important for model convergence. Algorithm 1 illustrates the SGD with Momentum approach.

Algorithm 1 Stochastic Gradient Descent optimizer with Momentum

```
Requirements: $\alpha$, $\mu$                ▷ $\alpha$ as learning rate and $\mu$ as the momentum
Initialize: $v = 0$
while not converged do
    $g \leftarrow \frac{\partial \mathcal{L}(\theta; x, y)}{\partial \theta}$
    $v \leftarrow \mu \cdot v - \alpha \cdot g$
    $\theta \leftarrow \theta + v$                ▷ Update parameters $\theta$
end while
```

Lately, alternatives to traditional SGD have emerged. Duchi *et al.* introduced AdaGrad [63], where the learning rate α is inversely proportional to the past gradient energy. Another popular optimizer is RMSProp [64], proposed by Alex Graves, which improves upon AdaGrad by introducing a moving average for the accumulated gradient.

The Adam *optimizer*, proposed by Diederik Kingma and Jimmy Ba [65], can be considered the next step beyond RMSProp, incorporating two momentum terms. While RMSProp considers only the second-order moment of the gradient with exponential decay, Adam adds a first-order term over the gradient. Algorithm 2 outlines the Adam optimizer, where $\odot$ represents the Hadamard element-wise dot product.

Algorithm 2 Adam optimizer

```
Requirements: $\alpha$, $\beta_1$, $\beta_2$                ▷ $\beta_1, \beta_2 \in [0, 1]$
Set: $\xi = 10^{-8}$
Initialize: $r = 0$, $s = 0$
while not converged do
    $g \leftarrow \frac{\partial \mathcal{L}(\theta; x, y)}{\partial \theta}$
    $t \leftarrow t + 1$
    $s \leftarrow \beta_1 \cdot s(1 - \beta_1) \cdot g$
    $r \leftarrow \beta_2 \cdot r(1 - \beta_1) \cdot g \odot g$
    $\hat{s} \leftarrow \frac{s}{1 - \beta_1^t}$
    $\hat{r} \leftarrow \frac{s}{1 - \beta_2^t}$
    $\theta \leftarrow \theta - \alpha \frac{\hat{s}}{\sqrt{\hat{r}} + \xi} \odot g$                ▷ Update parameters $\theta$
end while
```

2.2.3 AutoEncoder Architecture

Given the diversity of tasks to be addressed, numerous DNN architectures are available, including the Transformer [54], initially designed for NLP but also applicable to tasks like CV [66]. Attempting to

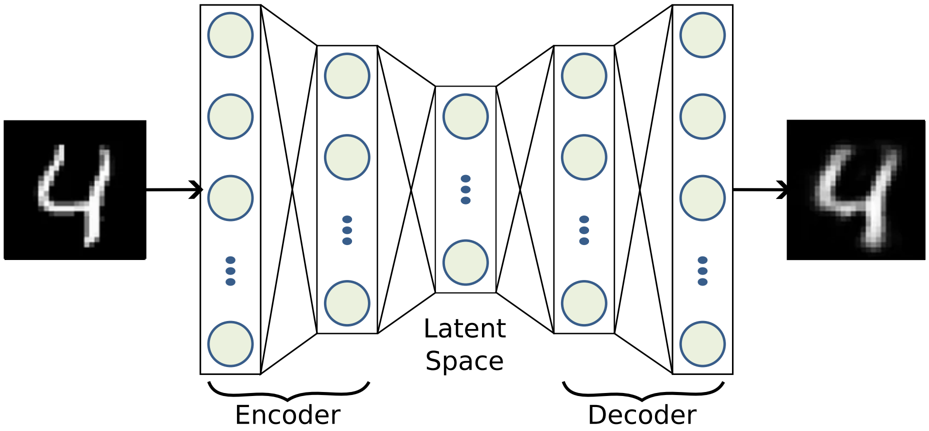


Figure 2.6: AutoEncoder architecture.

enumerate all DNN architectures proposed in the literature is unfeasible. However, the AutoEncoder (AE) [67] holds significant importance in self-supervised learning, making it a key component in this dissertation.

The AE architecture was proposed as a self-supervised method (see Fig. 2.1) for obtaining a low-dimensional data representation, a compression of a high-dimensional data. The idea is simple, the AE has the purpose of compressing the data in a latent space, Z, and obtaining a reconstruction from this. A example of this architecture is illustrated in Fig. 2.6. Even the AE illustrated in Fig. 2.6 use a fully-connected DNN, it can be implemented by using CNN.

The AE architecture [67] is characterized by three main components: the *encode path* or encoder, the *bottleneck*, which represents the compressed latent space, and the *decode path*. The encoder and decoder consist of a series of layers, $\{T_{E_1}, ..., T_{E_L}\}$ and $\{T_{D_L}, ..., T_{D_1}\}$, respectively, with L denoting the number of layers. An *encoder* comprises the encode path followed by the bottleneck, i.e., $T_E \in \{T_{E_1}, ..., T_{E_L}, T_Z\}$, while the *decoder* encompasses the decode path. In the example depicted in Fig. 2.6, both the encoder and decoder consist of two layers, making the AE a sequence of layers denoted as $f \in \{T_{E_1}, T_{E_2}, Z, T_{D_2}, T_{D_1}\}$.

For the loss function, a reconstruction error is calculated by comparing the input and the output obtained. Typically, the Mean Squared Error (MSE) is employed:

$$\mathcal{L}(\theta; X) = MSE(f(X;\theta), X) = \frac{1}{N}\sum_{x\in X}(x - \hat{f}(x;\theta))^2, \tag{2.18}$$

where N represents the number of samples in the dataset.

This architecture finds extensive use in image denoising [68] and has inspired applications such as image segmentation. For instance, U-Net [69], originally designed for medical image segmentation, employs an architecture heavily influenced by the AE, utilizing skip-connections that represent a direct connection between the encoder and decoder by concatenate outputs along the decode path. As a result, the input to a decode layer T_{D_i} is composed of the output of the corresponding encoding layer T_{E_i} and the output of the previous decoding layer $T_{D_{i-1}}$.

2.2.4 Preventing Overfitting in Neural Networks: Dropout Technique

As discussed in Section 2.1.2, overfitting occurs when an ML model memorizes the training data rather than learning the underlying patterns, typically associated with complex models. DNNs, for instance, often consist of tens of thousands or even millions of parameters to be learned. For example, MobileNetV3 [70], an efficient CNN designed for mobile applications, contains around 5 million parameters. Notably, architectures like Transformers [54] are notorious for their propensity to overfit [71]. While this large parameter count affords significant learning capacity, it also increases the risk of overfitting, making it a crucial concern to address.

Dropout [72, 73], or Binary Dropout, is a regularization technique commonly used in DL models to reduce overfitting. This mechanism consists in applying a multiplicative Bernoulli noise for each hidden unit in the neural network during the training, i.e, it is a feed-forward operation described as:

$$\begin{aligned} &z \sim Bern(\rho_d), \\ &\hat{\theta} = \theta \odot z, \\ &y = \sigma(\hat{\theta}x + b), \end{aligned} \tag{2.19}$$

where x denotes the vector of inputs into the layer, θ and b are the weights and biases in the layer, σ is a non-linear activation function such as, for instance, the sigmoid function and $\odot$ is a Hadamard dot product. The ρ_d value from the Bernoulli distribution, $Bern(.)$, also known as the dropout rate value, is a hyperparameter and represents the probability of an element being zeroed.

Dropout, according to the authors [73], breaks up complex co-adaptations between neurons that, rather than learning a good criterion, may learn to compensate for errors made by other neurons. A brief intuition about why dropout technique works is to understand that the connectivity of the layer to the prior layer is constantly changing, so you can interpret that the layer is learning using ‘different’ prior layers.

Interestingly, dropout has also been revisited as a Bayesian approximation for representing and estimating the model uncertainty [74].

2.2.5 Invariant Representation Unsupervised Learning

As previously discussed, one of the key features of DL is its ability to extract data features from different levels of abstraction, resulting in diverse data representations. However, learning an effective representation space without human interaction, such as data annotation, is generally challenging. One common approach to address this challenge is through the use of self-supervised AEs, see Section 2.2.3. This task becomes even more challenging when the goal is to obtain an invariant representation space.

In AI, there exist algorithms focused on obtaining features that are invariant to various transformations. For example, in CV, SURF (Speeded Up Robust Features) [75] and SIFT (Scale-Invariant Feature Transformation) [76] are commonly used to extract features from images that are invariant to affine transformations.

In DL, it is often desirable to obtain invariant features. This entails ensuring that any variation in the input, such as shifts or rotations, results in the same 'point' in the feature space derived from the hidden layer. This ‘point’ represents an invariant representation of the input. For instance, in

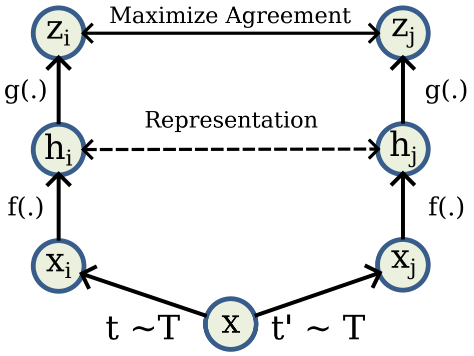


Figure 2.7: Framework proposed by SimCLR for contrastive learning.

the case of an image input, the invariant feature space should map both the original image and its rotated version to the same 'point'.

Invariant features are crucial for classification tasks, as they enable the model to identify the same object regardless of its position or orientation in the input space. The most direct method to achieve this in ML is through supervised learning, where labeled data guides the learning process and explicitly defines relevant features to extract. However, it is essential for these features to exhibit invariance to ensure robust and generalizable learning.

A significant challenge in DL lies in obtaining these invariant features through unsupervised learning. Researchers have explored using DL to derive an invariant representation space for years [77], with contrastive learning playing a pivotal role in this pursuit.

Contrastive Learning

Contrastive learning, a technique employed in DL, aims to learn a compact, invariant representation of the input X. In essence, given an image x and a transformed version $\hat{x}$ (e.g., rotated), the compressed representations of both images in Z should be close together. Conversely, dissimilar images should be located distantly in Z. This 'contrastive' approach aims to create a space where similar samples are clustered closely together and dissimilar ones are positioned farther apart. The primary challenge of this method lies in learning these effective representations without human supervision, making it an instance of unsupervised learning (see Section 2.1.1).

Although contrastive learning has been used for several years [77, 78], only recently has achieved comparable outcomes to those obtained with supervised methods [79]. In image classification, SimCLR [80] has produced results comparable to the state-of-the-art supervised method on ImageNet dataset [58]. SimCLR generates two augmented views of the same image and maximizes the similarity between them in a latent space. Figure 2.7 illustrates the framework proposed in SimCLR, where the input x is transformed using a subset of transformation defined in T, generating the pair samples x_i and x_j, and $f(.)$ represents the encoder that generates the representation space and $g(.)$ is known as projection head which generates the latent space Z in which the contrastive approach is carried out.

In the following years, methods based on SimCLR have been proposed. The enhanced SimCLR version, SimCLRv2 [81], achieves a better performance by utilizing a larger architecture. In addition, other methods such as MoCo (Momentum Contrast) [82], BYOL (Bootstrap Your Own Latent) [83] or SwAV (Swapping Assignments between multiple Views) [84] obtain state-of-the-art results on multiple benchmarks in image classification and object detection.

As it can be observed in Fig. 2.7, in contrastive learning-based methods, it is necessary to measure the similarity between the vectors in the latent space, Z. A common approach for high-dimensional data is to use the cosine angle to quantify this similarity. A high similarity is obtained when the angle is close to 0. Based on this quantification, it is necessary to define a loss function to encourage the model to learn to distinguish between similar and dissimilar samples. An example of such a loss function is the NT-Xent (Normalized Temperature-scaled Cross Entropy) [80, 85]. In this way, in the latent space the similar samples, those which are likely to belong to the same class, are projected close to each other, thus defining a specific cluster per class.

Generative Models for Invariant Representation

It is important to note that while contrastive learning is often associated with discriminative models, it is not the only method for acquiring a invariant representation. The use of a generative model for this purpose is intuitive, as the objective function aims to estimate the marginal distribution $p(X)$ for generating samples. The process of generating coherent samples necessitates the extraction of some invariant features. For example, in images of faces, the positions of the eyes and mouth are invariant features.

In the context of generative models, Bayesian inference can be considered. However, as previously discussed in Section 2.1.4, estimating $p(X)$ is often computationally intractable, making variational inference a suitable alternative.

Through variational inference, if the approximate parametric distribution $q_\phi(X)$, where ϕ represents the parameters, is tightly constrained and there is significant variability in samples X, the model will be compelled to achieve high levels of feature invariance among samples, which is particularly relevant. This task can be accomplished using VAEs, a AE that applies variational inference in the latent space Z. For instance, Dohi *et al.* [86] proposed the use of a VAE with constraints to control domain shifts in the latent space Z for anomalous sound detection. Another approach to achieving shift-invariant visual representation, similar to the one proposed in SimCLR, using VAEs was introduced by Chadebec and Allassonniere [87], utilizing data augmentation.

2.3 Interpretability in Machine Learning

The previous sections have delved into ML and DL methodologies, with examples on general tasks oriented to CV and NLP. However, the application of AI spans a wide array of domains beyond CV and NLP, including biology, healthcare, autonomous driving, and many others. While providing a comprehensive review of AI applications, particularly in DL, is a monumental task that lies beyond the scope of this thesis, it is important to note that the upcoming chapters in Part II will focus on healthcare applications. Healthcare is a domain where the utilization of DL models is gaining significant traction, albeit their deployment has several challenges.

The aforementioned challenges of DL in healthcare, which could also extend to other critical fields such as autonomous driving, primarily concern the reliability of the DNNs. Addressing these challenges will require tackling several ethical considerations. While state-of-the-art algorithms based on DL have showcased performance comparable to that of healthcare professionals [6], the lack of transparency of these models, often denoted as 'black-box', could lead to severe consequences in their predictions.

The decision carried out by a ML model is complicated to justify since its decision-making process cannot be discussed, particularly when the inner workings of the model are not well understood. The best way to justify a ML model is by addressing the problem by Bayesian inference (Section 2.1.4) since it is possible to quantify the uncertainty about the unknown parameters of the model by the posterior distribution, which describes the parameters after observing the training set. In this way, it can be deduced how the parameters affect the accuracy of the model. However, in general, humans do not feel comfortable with these models due to a lack of understanding of the ML problem-solving mechanisms. This uncertainty becomes more pronounced with the adoption of complex ML models, such as the DNNs. For instance, Decision Trees are known for their interpretability, making them among the most easily understandable ML methods, second only to simple linear regression models. However, in general, models like RF, which utilize multiple Decision Trees, sacrifice interpretability for improving the accuracy.

As it has been previously mentioned, the effectiveness of DNNs in solving complex problems in different topics is well-established [40, 6]. The Universal Approximation Theorem [88, 89] underpins this capability, asserting that a neural network with a single hidden layer can approximate any function to the desired level of accuracy, given certain conditions. These conditions usually involve employing a non-polynomial activation function and ensuring a sufficient number of hidden nodes. Essentially, this theorem suggests that DNNs can be considered as non-parametric estimators, particularly when the number of hidden nodes and layers can scale proportionally with the number of observations. This adaptability helps DNNs to capture and model intricate relationships within data without being confined by specific parametric forms. However, a major criticism of DNNs is their interpretability; these models have a decision-making mechanism that is challenging for humans to understand.

DL models have gained immense popularity in recent times due to their ability to provide highly adaptable outputs with high accuracy. However, it is important to remember the aphorism expressed by George Box [90]:

> *All models are wrong, but some are useful.*

Nowadays, DL architectures are becoming increasingly deep [91], driven by the pursuit of better results. This trend is often fueled by the myth that there is necessarily a trade-off between capacity and interpretability [8], i.e., better results are obtained by using a more complex model. However, sometimes a simple preprocessing of data allows obtaining better results using a simpler model. Consequently, rather than making an effort to create models that are inherently interpretable, there has been a recent explosion of work on 'explainable ML', where a second model is created to explain the first 'black box' model [8].

While it is true that current DL models have surpassed traditional techniques in AI in many areas, relying solely on specific features of these models, such as the number of hidden layers or the number of parameters, cannot be considered the exclusive way to improve current results. A deeper understanding of these models would provide a better intuition about their problem-solving mechanisms, and, for this reason, the interpretability of ML models should be considered a key feature in assessing their utility, making them more useful.

2.3.1 Interpretability versus Explainability

It is necessary to have a description of the concepts of ‘interpretability’ and ‘explainability’, but it is worth noting that, to the best of my knowledge, there is still no consensus on these definitions [92]. Although there is no definitive explanation for this, it might be a consequence of the current popularity of DNNs, albeit interpretability has been an active area of AI research for many years [92]. However, this thesis will use both expressions with the following descriptions:

Interpretability is associated with explaining the meaning of something [1] in understandable terms. This definition based on “understandable terms” is reflected in the literature [92]. Although it is not a formal definition, it effectively conveys the main idea of interpretability as approached in this dissertation.

Interpretability is fundamental for understanding and reasoning about the inner workings of a ML model. Interpretability involves deriving logical insights from the observation and establishing a hypothesis about the problem-solving mechanism employed by the model. This interpretation should identify specific features in the model, which can be validated through experimentation, ensuring they are satisfied.

By defining a human-comprehensible framework for understanding how the underlying AI operates, interpretability allows us to validate or correct the model through experimentation. However, interpretability should not be considered absolute. The way a model is interpreted may vary depending on the context, the observer’s perspective, and the objectives of the analysis. Interpretability is, therefore, a domain-specific notion [8].

In the literature, there are authors who consider that a model it is interpretable if, observing the parameters, it is possible to understand how the prediction is made [93] or the degree to which a human can understand the cause of a decision [94, 95]. In this thesis, these definitions fit better to an ‘explainable ML’ [8].

Explainability in the context of ML goes beyond interpretability by providing an explanation in a human-readable form of how and why a model came up with a prediction [96]. Therefore, explainability is associated with the ability of the model to provide an understandable justification for its decisions. Explainability relies on interpretability as a building block and aims to answer questions like “why did the AI make this particular decision” [96].

In practice, explainable ML for complex models, such as DNNs, is based on *post-hoc* techniques [91, 92] where a second and simpler model is used that provides a score for defining what part of the input data is responsible for the final decision. However, the use of specific techniques, such as attention mechanisms, can enhance model explainability by providing better intuition about the prediction process, as attention mechanisms assign weights to different parts of the input.

Explainability depends on interpretation, which tends to be subjective and based on the observer’s perspective. For instance, in a clinical decision, the explanation of why the model made a specific decision could be based on the score of the input features, and it is expected to be validated based on experts’ considerations. However, in high-dimensional data, there could be significant variability among experts regarding this validation, as each expert might consider different features relevant for the decision.

The aforementioned descriptions regarding interpretability and explainability fit with the Euro-

[1] Cambridge Dictionary, accessed 2024-04-18

pean regulation[2]. However, although explainability is crucial for trust in ML models, it may not be necessary if systems are sufficiently interpretable [8].

2.3.2 Interpretability in Deep Learning

As previously discussed, interpretability is a key concept of this thesis. Although there is still no consensus regarding the definition of interpretability within the DL community, this thesis will understand interpretability as the ability to provide explanations in understandable terms. Specifically, in this thesis, 'understandable terms' are oriented towards an engineering perspective, meaning that this interpretation is based on well-known systems. Consequently, the goal is to interpret the 'black box' by relating it to well-studied models or systems and ensuring that the DL model satisfies the features present in these systems.

One illustrative example of this approach is the Bayesian interpretation of the dropout technique in DL by Gal and Ghahramani [74]. While there are various interpretations of why the dropout technique works [73], Gal and Ghahramani viewed dropout in DNN as an approximation of a deep GP, a probabilistic Bayesian model. This technique minimizes the KL Divergence between an approximate distribution and the posterior of a deep GP, following the variational approach (see Section 2.1.4). As a result, given the validation of this interpretation as a deep GP, they proposed a computationally efficient method for quantifying model uncertainty using Monte Carlo methods.

Another inspiring work for this thesis is the research conducted by Tishby and Zaslavsky [97]. In this study, the authors use Information Theory (IT) to gain a theoretical understanding of DNNs, describing them as communication systems and examining the information flow within these channels. This approach will be further explored in the upcoming chapters.

This interpretation of DL models provides a more in-depth understanding of their problem-solving mechanisms. As demonstrated in both works [74, 97], this approach helps establish specific features inherent in DL based on that interpretation. Since evaluating interpretability is complex, this approach allows for assessments based on the requirements of the aforementioned systems, providing a more in-depth understanding of the model.

It is worth noting that sparsity is a useful measure of interpretability in models that use structured data [8]. Sparse models enable a more comprehensible understanding of how variables interact with each other. This perspective will also be emphasized throughout this thesis.

2.3.3 Explainability in Deep Learning

Most work oriented to understanding DL models is commonly associated with this 'explainable ML' approach [91, 92]. These methods seek to elucidate the 'black boxes' by explaining how the model makes decisions, focusing on the relevance of input features or model parameters. However, in extremely complex models, such as the DNNs, the explanation based on model parameters is unfeasible, so it is more commonly to approach this explanation based on the input features. As it has been previously mentioned, these 'post-hoc' methods are those that are typically applied after a model has been previously trained, and they aim to explain the model's predictions. They are often model-agnostic, meaning they can be applied to any type of model. The idea of these methods is to establish a clear connection between the input and the model's prediction [91]. However, it is worth

[2] TechDispatch #2/2023 - Explainable Artificial Intelligence, accessed 2024-08-02

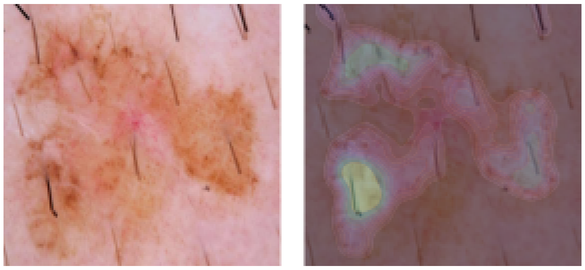


Figure 2.8: Grad-CAM to estimate the feature contribution for skin-cancer detection [102].

noting that employing these methods is resource-intensive, and achieving generalizability is complex because each sample must be analyzed independently.

Backpropagation-based Approaches

There are several methods that are designed explicitly for DL, which means that those methods cannot be considered model-agnostic. A general approach of post-hoc methods for DL are the methods known as backpropagation-based approaches. One of the first works related to these methods was proposed by Simonyan *et al.* [98], in the context of classification task of CV, in which there is computed a 'saliency map' using the gradient of the output and backpropagating them to the input convolutional layer.

The next relevant work of the aforementioned method was the Integrated-Gradient proposed by Sundararajan *et al.* [99]. These methods consist of computing the integration of the gradient-based feature attribution for an input x with respect to a baseline x'. In images, the baseline could be a black image and the gradient is integrated from the baseline image to the input x, which can be computed by obtaining n interpolation between x and x'.

An alternative to the Integrated-Gradient is Deep Learning Important FeaTures (DeepLIFT) [100]. In this case, it is computed the gradients, but it is compared the gradients of the inputs x with a baseline x'. In this case, it is important that the baseline represent an input without information. This baseline could be a black image or an image with random noise. However, the selection of a suitable baseline is crucial [100, 101], as it influences the attribution of feature importance.

However, regarding backpropagation-based methods, Gradient-weighted Class Activation Mapping (Grad-CAM) [103] stands out as one of the most popular. Grad-CAM computes a coarse grained feature-importance map by associating the feature maps in the final convolutional layer with particular classes based on the gradients of each class w.r.t. each feature map, and then using the weighted activations of the feature maps as an indication of which inputs are most important. For instance, in [102], we applied this method for validating the use of an attention mechanism with the purpose of reducing the number of parameters of a MobileNet model [70]. An illustration of Grad-CAM result can be observed in Fig. 2.8.

Perturbation-based Approaches

Perturbation-based techniques involve perturbing data around an individual prediction and analyzing the impact on the model's performance [104]. The explanation is derived from the conclusions obtained by, for instance, occluding part of the input image with a mask or replacing a word in a sentence with its synonym and observing the changes in the model's output. Unlike backpropagation-based methods, these techniques are model-agnostic because only perturbations in the input are required.

In the state-of-the-art is possible to find works in which the performance of DL models is affected by small perturbation. For instance, a well classified image can be drastically wrong by a small changes, indistinguishable to the naked eye [105, 106]. Those analyses are known as adversarial example-based sensitivity analysis [104]. In this particular type of analysis, the adversarial examples pertain to datapoints whose characteristics have been impacted by a subtle yet substantial amount, sufficient to cause a ML model to make incorrect predictions about them.

One of the most popular methods using this approach is Local interpretable Model-agnostic Explanations (LIME) [107]. This method involves generating randomly-sampled data around the neighborhood of a given instance and its corresponding prediction. Subsequently, predictions are made for the generated instances using the model in question and weighted by their proximity to the input instance. Finally, by using a simple model such as a Linear Regressor or a Decision Tree, is trained using this perturbed dataset [107, 104]. LIME, as the authors claimed, provides a local explanation for individual predictions by approximating the intricate decision boundaries of a model using the aforementioned simpler model. As a result, an understanding about which features contributed the most is obtained. A theoretical validation of LIME was carried out by Garreau and Luxburg in 2020 [108].

However, it is possible to find perturbation-based method designed exclusively for DL. For instance, Zeirver and Fergus [109] proposed to apply occlusion in different segments of an input image for evaluating the impact in the later layers.

2.4 Inverse Problems

Inverse problems are distinguished by their aim to recover information from noisy data [110]. One illustrative example is the medical imaging, including techniques such as Computed Tomography (CT) or Magnetic Resonance Imaging (MRI). In this context, the measurements are taken from the external surface of a body, and the challenge is to infer properties of the inaccessible 'interior' or 'hidden' information [111]. In general, the restoration problems, those in which the input signal is 'noisy' and the goal is to obtain the clean version of this, are an inverse problem.

In inverse problems, it is commonly assumed that the noise corresponds to an additive white noise. In this case, if it is considered an inverse problem, the degradation model is defined as follows [110]:

$$X = f(X') + \epsilon$$

where the noise signal X is obtained by a $f(.)$ transformation over the clean signal X' and ϵ is the noise injected into the X'. The key to inverse problems is that there is an unknown transformation of the signal, which is corrupted by measurement noise, and the goal is to estimate the input signal. This example is illustrated in Fig. 2.9.

It is widely assumed that this white noise corresponds to a Gaussian noise, specially in image processing. It means that all the elements of $\epsilon \in \mathbb{R}^d$, where d is the dimension of the X, come from the same Gaussian distribution with zero mean and standard deviation σ. However, the noise assumption is selected based on the context. For instance, the noise associated to MRI is commonly

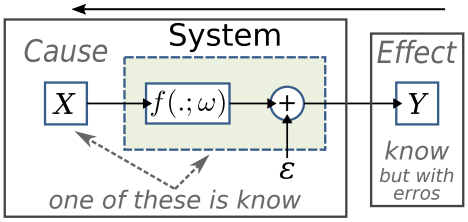


Figure 2.9: Inverse problem scheme with white noise. In the scheme, X is the input, $f(.;\omega)$ represents a transformation and ϵ is the noise injected and Y is the output. As an inverse problem, the idea is to obtain the input X, that could be represented by a noiseless signal, or the system with parameters ω, given the other two quantities. The data observed Y is always known, but it contains errors.

indicated as a Rician noise [112] and in audio signals is the impulse noise [113]. In images, particularly in Synthetic Aperture Radar (SAR) and medical ultrasound images, speckle noise is a prevalent form of multiplicative noise [114].

It is essential to recognize that assuming white noise can be a strong simplification of real-world scenarios. In practice, noise is rarely isolated and may have complex characteristics. Nevertheless, the white noise assumption serves as a mathematical abstraction that simplifies the analysis, making problem-solving more tractable [110]. As a result, this approximation is considered suitable for many practical applications.

As a result, the inverse problem is characterized because the input or the system is unknown. For instance, in the context of the restoration problem, it is indeed an inverse problem where we have knowledge of the noisy signal (the effect), and the goal is to obtain the original signal without noise (the cause). This is a classic scenario in inverse problems and the Linear Fredholm integral equation of the first kind plays a fundamental role in image restoration [115, 110], including medical imaging.

However, the aforementioned description did not consider another scenario where inverse problems are defined: the determination of the parameters ω in the function $f(.;\omega)$. Generally, while a forward problem involves determining the data produced by particular model parameters, an inverse problem involves estimating the parameters that produced the observed data, the effect [116, 110]. Figure 2.9 illustrates the inverse problem where, from the noisy effects, it is necessary to determine the cause or the system, given the other two quantities.

Machine Learning is an Inverse Problem

As recently mentioned, obtaining the parameters of a system based on observations is an inverse problem. This process is precisely what is carried out in ML optimization, it is the learning process.

In supervised learning, the target Y (the effect) and the input X (the cause) are known. The learning process involves updating the model parameters based on a loss function, typically using methods such as SGD in DL (see Section 2.2.2). As illustrated in Fig. 2.9, this scenario corresponds to knowing both the cause and the effect and only needing to determine the model parameters. For instance, Tarantola's book on inverse problems [116] illustrates how to estimate model parameters and perform data fitting using methods mentioned in Sections 2.1 and 2.2, demonstrating the relationship between inverse problems and ML.

Additionally, inverse problems such as obtaining a clean signal from a noisy observation can be approached using ML. In DL, architectures based on AEs are commonly used for these tasks [68]. A notable example is the Deep Image Prior (DIP) [117], which addresses various inverse problems such as super-resolution, inpainting, and denoising using an AE architecture.

2.4.1 Ill-posed Problems

Typically, several inverse problems, as described by Hadamard [118], are considered ill-posed. Hadamard's perspective suggested that these problems do not accurately represent real-world scenarios, considering that it is a consequence of the problem being not properly presented. However, it has become evident that he was mistaken in this regard [110].

The well-posed problems are those problems that describe systems of equations whose solutions behave as it is (heuristically) expected from a physical system. According to Hadamard's definition, the well-posed problems are linear problems that satisfy the following requirements:

- Existence: the problem must have a solution. It means that the function $f(.)$ is surjective.
- Uniqueness: it has a unique solution, i.e., $f(.)$ is bijective.
- Stability: the solution must depend continuously on the data. If this property is satisfied, it is referred as *well-conditioned*.

In simpler terms, for a problem to be well-posed, it needs to have a unique solution and the solution should vary continuously with changes in the input data. Conversely, if a problem violates any of these requirements, such as being *ill-conditioned* due to lack of stability, then it is considered ill-posed. In general, many inverse problems are often ill-posed.

Machine Learning for Ill-posed Problem

The properties necessary for obtaining a well-posed problem are essential for developing models that generalize properly. The learning process in ML should be well-posed to converge to an optimal solution to the problem.

In Section 2.1.3, three main sources of uncertainty that ML must handle were introduced. For instance, with incomplete observation, a deterministic system cannot be well-modeled. This lack of data generates an ill-posed problem that must be addressed by the ML algorithm.

In cases where the problem is ill-posed, it is necessary to detail some implicit 'prior belief' to ensure that the ML algorithm can generalize properly when solving the problem. One of these implicit priors is the *hypothesis space* [21], which is implicitly defined by the selected model and is a determinant for expressing the model's capacity. For example, the linear regression algorithm has the set of all linear functions of its input as its hypothesis space. This space can be extended by including polynomials rather than linear functions [21]. Using a 'simpler' hypothesis space can emphasize obtaining a unique solution.

Other features relevant for ensuring proper learning in ML are the *smoothness prior* and the *local constancy prior* [21]. These priors aim to ensure stability, requiring the problem to be well-conditioned. The smoothness and local constancy priors assume that the function to learn must satisfy the condition

$$f(x) \approx f(x + \epsilon), \tag{2.20}$$

assuming that a minimal variation in the input corresponds to a minimal variation in the output. For instance, the smoothness prior is especially relevant for optimization methods based on gradients, such as the SGD used in DL (see Section 2.2.2) and the local constancy is fundamental in methods based on k-nearest neighbors.

Although there are simpler algorithms that rely exclusively on these priors, in general, they are not sufficient for solving the complex statistical challenges that current AI-based methods aim to address. Therefore, to effectively deal with the ill-posed problems commonly associated with AI tasks, it is necessary to apply the regularization technique.

2.4.2 Regularization

Regularization is a technique employed to address ill-posed problems. It was originally introduced by Andrey Tikhonov [119]. It involves introducing constraints that enhance the injectivity and stability of these problems.

Consider an ill-posed problem, and it is desired to obtain a single solution defined by a linear function $y = \Phi\alpha$ by introducing some additional identifying criteria. In this way, a regularization function $J(\alpha)$ that evaluates the desirability of a would-be solution α is introduced. This regularization function, has to be defined for smaller values being preferred. In this way, the general optimization problem by Tikohonov's regularization is defined as follows:

$$\min_{\alpha} J(\alpha) \qquad \textbf{subject to} \quad y = \Phi\alpha. \tag{2.21}$$

An ill-posed inverse problem could be considered an infinity of different paths to a set of solutions given some data. The regularization has the main purpose of selecting a unique or a subset of solution from this infinity set of solutions, recovering the ideal cause x or, at least, finding a x' that is close to this ideal solution.

Bayesian Interpretation

In Section 2.1.3, the problem statement was defined through a Bayesian interpretation. Within this framework, ML models cast the problem as one of minimizing a loss function subject to a regularization-based constraint. This can be mathematically formulated as follows:

$$\underset{\omega}{\operatorname{argmin}} \ \mathcal{L}(\omega; X, Y) \quad \textbf{s.t.} \quad J(\omega), \tag{2.22}$$

where (X, Y) denotes the data and target variable pair in supervised learning, and ω represents the model parameters. To reframe this expression, a Lagrange multiplier λ is introduced:

$$\underset{\omega}{\operatorname{argmin}} \ \mathcal{L}(\omega; X, Y) + \lambda J(\omega). \tag{2.23}$$

Finally, in line with Eq. (2.7), the formulation can be refined to:

$$\underset{\omega}{\operatorname{argmin}} - \log p(Y|X, \omega) - \log p(\omega|\lambda). \tag{2.24}$$

Observing Eq. (2.24), it is possible to find similarities with the objective described in vari-

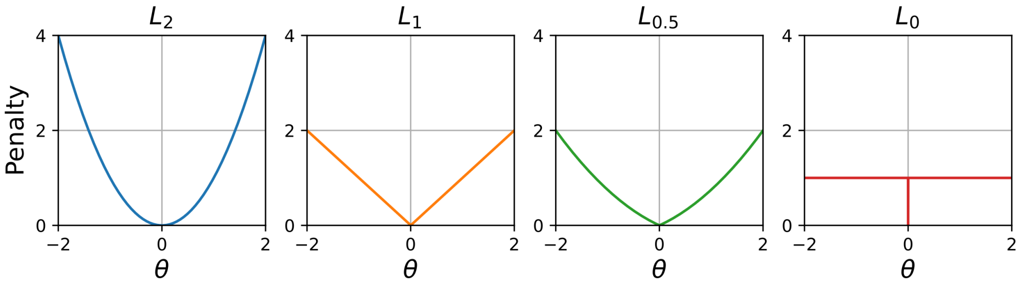


Figure 2.10: L_p-norms with p = 2, 1, the L_p-quasinorm with p = 0.5 and the L_0-pseudonorm.

ational inference in Section 2.1.4. As it can be observed in Eq. (2.12), the KL Divergence, $D_{KL}(q_\phi(\omega|x)||p(\omega))$, is a regularization factor that is forcing to obtain the solution in which the distribution of the parameters $p(\omega)$ approximately follows the easier parametric distribution denoted by $q_\phi(\omega|x)$. In other words, variational inference uses a regularization factor for obtaining a solution that it is close to the ideal one.

2.5 Sparse Representation

Sparse representation is widely utilized across various domains for tasks such as data compression, high-dimensional data analysis, and enhancing the performance of ML algorithms. In the context of ML, as discussed in Section 2.3.2, it considered a measure of interpretability of the model [8] because it provides insight into how variables interact jointly rather than individually.

In essence, sparse representation involves representing a signal or image using only a few elements from a considerably large dictionary [120]. These representations comprise a sparse set of non-zero coefficients, indicating the contribution of each basis vector to the overall approximation. A prominent example of this concept is the JPEG image compression standard [121], which employs the Discrete Cosine Transform (DCT) to transform image blocks into a basis where the representation is sparse. This implies that most of the higher frequency components become zero or near-zero, facilitating efficient compression.

Moreover, the application of sparse representation extends beyond data compression. It can be instrumental in data fusion by extracting the most relevant features from diverse sources. For instance, in medical imaging, sparse representation can aid in obtaining pertinent features from various imaging modalities and subsequently integrating them through fusion techniques for enhanced diagnostic accuracy [122]. There are more examples oriented to Feature Selection methods that will be discussed in Chapter 5.

In this field, the L_p norm plays a fundamental role. Consider a vector $x \in \mathbb{R}^d$, the L_p norm is defined as $||x||_p = (\sum_i |x_i|^p)^{1/p}$. However, this norm is defined for $p \geq 1$. In the case of $0 < p < 1$, $||.||_p$ is known as quasinorm [120]. The L_1 and L_2 are the well-known Manhattan and Euclidean norm, respectively.

In the case of $p < 1$, the aforementioned expression cannot be considered a norm, since it does not satisfy the triangle inequality [120] (see Fig. 2.10):

$$||x + y||_p \leq ||x|| + ||y||. \tag{2.25}$$

This inequality formalizes that the shortest distance among two points is a line.

2.5.1 Sparse Vector

In the field of mathematics and signal processing, a vector is said to be sparse if most of its elements are zero. More formally, a vector is considered sparse if there exists a basis (a set of linearly independent vectors) in which the vector has a sparse representation. In this way, it is important to note that the sparsity of a vector can depend on the basis in which it is represented. Some vectors may not appear sparse in one basis, but may be sparse in another.

In order to measure the sparsity of a vector, the pseudo-norm L_0 is the main approach. It is described as follows:

$$||x||_0 = \sum_i \mathbb{1}(x_i = 0) = \lim_{q \downarrow 0} ||x||_q^q, \tag{2.26}$$

where $\mathbb{1}(.)$ is the indicator function. As a result, this pseudo-norm counts the number of elements that are 0 in the vector x.

This L_0 pseudo norm is a special of L_p-norm in which the absolute homogeneity condition is not satisfied. For this reason, it is commonly known as pseudo-norm. Nevertheless, it is often referred to improperly as L_0-norm in the literature. In this thesis, it will be indicated as L_0-norm.

2.5.2 Sparse Solutions

As inferred, obtaining a sparse representation can be categorized as an ill-posed problem, as outlined in Section 2.4.1. It is intuitively understood that if there are two subsets of variables containing the same information, there is no unique solution. However, sparsity can help address ill-conditioning by improving the stability of the solutions. In the context of ML, it is often assumed that the model will utilize the entirety of the data to resolve the problem. Here, regularization becomes crucial in deriving a sparse solution:

The model's objective, involving the parameters ω is defined as follows:

$$\underset{\omega}{\operatorname{argmin}}\ \mathcal{L}(\omega; X, Y) \quad \textbf{s.t.} \quad ||\omega||_0. \tag{2.27}$$

This problem setup will be further explored in Chapter 5. It is important to note, however, that the L_0-norm is non-convex and its computation can become prohibitively expensive and intractable, particularly with high-dimensional data [120]. This necessitates careful consideration of the computational strategies and approximations used in practical applications to effectively manage the complexity of such models.

2.6 HyperSpectral Imaging

HyperSpectral Imaging (HSI), also known as imaging spectroscopy, refers to a technology that combines traditional digital imaging and spectroscopy methods, providing detailed information about the materials in the captured scene. Indeed, this technology possesses the capability to surpass the limitations of human vision. Although it is a technology that it has been used in remote sensing — typically associated with the use of satellite- or aircraft-based sensor technologies to measure objects without physical contact — for over three decades [123, 124], recently the access to this technology has become more widespread, making it more accessible for use in a wide range of fields.

The sensor measures the electromagnetic radiation reflected, emitted, or transmitted by objects across a range of wavelengths, generating a vector of radiance values for each pixel [125, 126]. This measurement of the absorbed or reflected radiation at a detailed wavelength range results in a pattern, known as a spectral signature [125, 126]. As a result, each pixel in an image corresponds to a unique spectral signature that contains information about the molecular composition and texture of the materials, which can be potentially used to identify any object. This feature corresponds to spectroscopy technology, and at the same time, the digital imaging provides the spatial information in the HSI, which is used for obtaining information about the morphological features [127].

The aforementioned description is similar to another technology known as multispectral imaging, as both technologies share similar characteristics. The primary distinction lies in the number of measurements across the wavelengths. The term ‘hyper’ in hyperspectral suggests an abundance, or “too many”, referring to the extensive number of wavelength bands captured [123]. The ‘bands’ are a narrow wavelength range of the electromagnetic spectrum. Hyperspectral sensors are engineered to detect radiation across a broad range of the electromagnetic spectrum. Although the wavelength range of HSI varies depending on the system and the application, the sensors commonly capture information from the Visual and Near-Infrared (VNIR) region to the Short-Wavelength Infrared (SWIR) region, with each band having a bandwidth of around 5-20 nm [126, 128]. Additionally, some hyperspectral sensors are capable of covering the electromagnetic spectrum up to 12 microns [127]. In summary, HSI systems are characterized by a more significant number of spectral bands compared to multispectral sensors, allowing for more detailed and precise data collection. Figure 2.11 illustrates an example of a hyperspectral image, depicted the image obtained in a spectral band and the spectral signature of a pixel, for the case of a human brain.

2.6.1 Hyperspectral Data Structure

As previously discussed, HSI can be regarded as an imaging modality that delivers substantial information about the interaction of light with objects across various wavelengths. Like any imaging modality, the data from HSI is structured, specifically into what is known as a hyperspectral cube.

The hyperspectral cube is essentially a three-dimensional array composed of a sequence of images, each representing data captured at specific spectral bands or channels, as illustrated in Fig. 2.11. Each band is dedicated to capturing light interactions within a distinct, narrow range of wavelengths. Consequently, for a hyperspectral image with spatial dimensions $W \times H$ and L spectral bands, the data can be represented as $X \in \mathbb{R}^{W \times H \times L}$. However, depending on the application, this structure may be altered to discard spatial information.

Although spatial data is crucial in some research, such as the study by Wei *et al.* [130], which explores spectral-spatial features for HSI classification, in remote sensing applications, HSI is often characterized by high spectral resolution and low spatial resolution due to each pixel covering extensive ground areas. In this field, the spatial resolution is one of the most challenging features to achieve in this technology due to the associated trade-offs between spatial resolution, spectral resolution, and signal-to-noise ratio, since high-spatial-resolution hyperspectral imaging can be difficult to achieve [131, 132]. As a result, in some scenarios, the spatial information is considered less critical and is removed by reshaping the hyperspectral cube into a matrix $X \in \mathbb{R}^{N \times L}$ where $N = W \cdot H$.

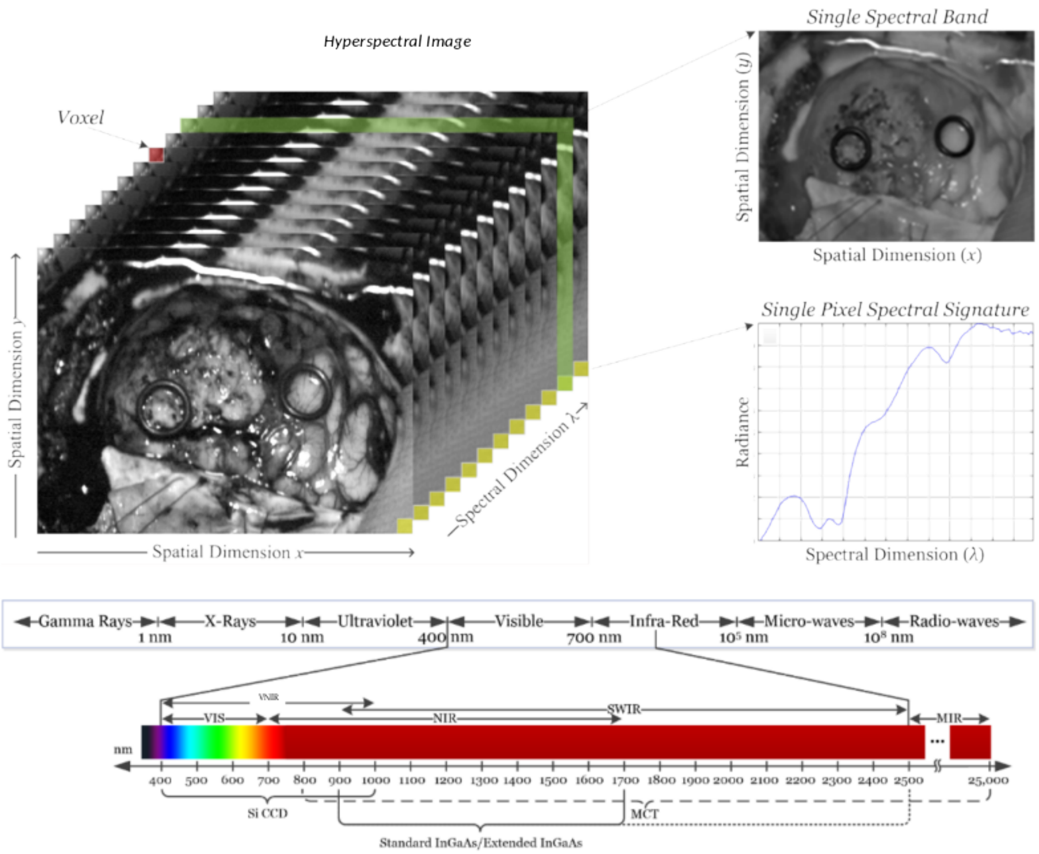


Figure 2.11: Basic structure of a hyperspectral image, single band representation at a certain narrow wavelength and spectral signature of a single pixel. Image obtained from [129].

This reshaping facilitates analyses focused primarily on spectral data, optimizing processing for applications where spatial context is secondary.

2.6.2 Medical Hyperspectral Imaging

This technology emerged in the field of earth observation and space exploration since the beginning of this millennium [123, 133]. It has found widespread applications in various geological studies, including lithological and mineralogical mapping, ore exploration, and environmental geology [134, 124]. Additionally, this technology is being utilized in other domains such as disease diagnosis and surgical guidance [135], forensic science for evidence analysis [136], food quality evaluation [137], plastic identification for the recycling industry [138], or art conservation [139], among others.

This section will specifically delve into the applications of HSI within the healthcare sector, examining how this technology aids in advancing medical practices and improving patient care outcomes. One of the key attributes that makes HSI particularly valuable in healthcare is its non-invasive nature [129]. This feature allows medical professionals to obtain crucial physiological and biochemical information without direct contact or disruption to tissue, offering a significant advantage in both diagnostics and therapeutic monitoring.

The study of light propagation through biological tissues is instrumental in identifying various diseases [140]. This process involves three key photophysical phenomena: refraction, scattering, and absorption. Those phenomena serve as distinctive markers of the molecules' responses to light, providing critical diagnostic information [140]. These properties of the interaction between light and biological tissue motivate the use of technologies that exploit the information on light propagation through tissues to develop tools for diagnosis support.

While various technologies are available for measuring light, such as Raman spectroscopy [141], HSI stands out as an optical spectroscopy imaging modality that directly measures the incoming radiance spectra of light [127]. However, it is important to note that the use of HSI systems in medical applications has not yet become standardized, which is a critical factor to consider in its broader adoption and implementation. This lack of standardization can affect the comparability of results across different studies and applications in healthcare scenarios.

The application of HSI in the medical field has been gaining traction recently. While HSI has been utilized for specific cases, such as assessing neurocognitive disorders [142] and analyzing diseases in the eye through retina oximetry [143], its use in cancer analysis has been particularly noteworthy [129]. In 2016, Halicek *et al.* [144] highlighted the limitations of HSI in cancer detection, offering valuable insights into its applications and suggesting directions for future enhancements of this technology. Current research includes diverse studies targeting various types of cancer, such as gastrointestinal stromal tumors [145], tumor tissues in the head and neck [146], and brain cancer [147].

Although the potential of HSI for cancer analysis is increasingly recognized, the lack of standardization in this technology complicates the development of CV algorithms tailored for processing these specific tasks. Nevertheless, progress is being made; recently, a benchmark dataset focused on brain cancer detection was released [148]. This dataset, developed from various acquisition campaigns using a customized intraoperative hyperspectral acquisition system, represents a significant step toward standardizing data and enhancing the reliability and effectiveness of HSI in medical diagnostics.

Chapter 3

Information-Theoretical Framework

Information is the resolution of uncertainty.

Claude Shannon

Even though DL models demonstrate exceptional performance, their integration into certain domains faces challenges due to their ‘black box’ nature and their multilayer nonlinear structure. This is particularly important for this thesis, which seeks to provide insights into the internal mechanisms of such models to facilitate their integration as support tools in the field of medicine.

In conventional practice, ML models undergo validation using designated datasets, with performance assessment based on specific metrics, such as ‘accuracy’ or the ‘F1-Score,’ particularly in classification tasks. Nevertheless, after validation, the critical concern shifts towards establishing trust in the model’s real-world applicability. This is particularly vital in sensitive domains like medicine, where the emphasis must be placed on ensuring fairness and reliability in the deployment of such models. Nevertheless, these qualities can be challenging to ascertain in models where their predictions are not interpretable by humans. In such cases, would it be possible to trust in the ‘black box’?

Understanding the model’s problem-solving mechanism is essential for gaining confidence in its performance. One potential solution is to establish a clear connection between input features and the model’s predictions [91], the ‘Explainable ML’. In the previous chapter, Section 2.3.3, various algorithms were discussed with the aim of identifying the features that characterize the obtained result, that is, the prediction. However, as it was mentioned, employing these methods is resource-intensive, and achieving generalizability is complex because each sample must be analyzed independently. This is due to the fact that the explanation provided by these methods relies on subjective interpretability, which is shaped by the observer’s perspective. This scenario is exemplified in the study that we conducted about how a specific attention mechanism works for skin cancer detection [102], where Grad-CAM is employed to assess how the attention mechanism impacts skin cancer monitoring. As a result, the study aimed to identify patterns commonly used by medical doctors in their clinical decisions in order to validate the attention mechanism employed.

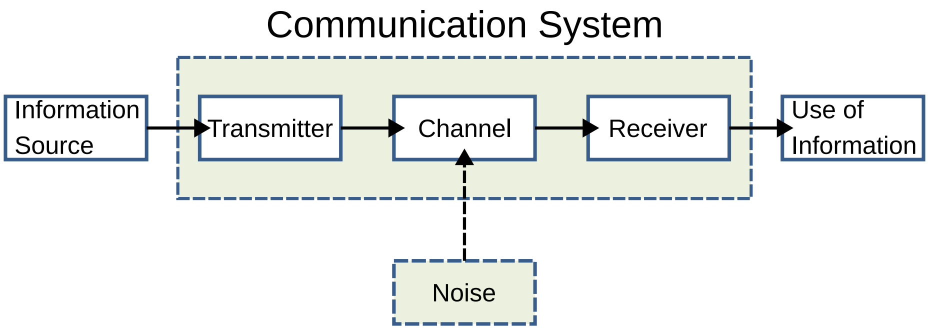


Figure 3.1: This block diagram illustrates the components of a communication system. The transmitter's primary functions include message compression and encoding to facilitate error-free transmission. At the receiving end, the transmitted message is decoded and decompressed to retrieve the original information.

Currently, the DL architectures are getting deeper and, as a consequence, the number of parameters is increasing. For instance, the well-known architecture VGG-19 [149], which was proposed in 2014, has 19 nonlinear layers, and it contains about 144 million parameters. As you would notice, this number of parameters is a clear indication that the interpretability or explainability of this model is really complex. A simpler model is easier to understand at the cost of losing accuracy in its predictions. In a critical task, such as those found in the field of medicine, the trade-off between accuracy and interpretability of the model is especially relevant [91].

In this way, the idea of finding a more abstract perspective that allows to understand the mechanism of the proposed model is interesting to improve the interpretability of these models. Here, the 'information' plays a fundamental role, since it is an interdisciplinary concept which it applied by almost every scientific discipline within its context and regarding specific phenomena [9].

A fundamental problem in formalizing our intuitive ideas about information is to provide a quantitative notion of "meaningful" or "relevant" information. These issues were intentionally left out of Information Theory (IT) in its original formulation by Shannon, who declared he was not going to be interested in meaning at all [10]. Shannon focused attention on the problem of transmitting information rather than judging its value to the recipient, understanding information as something transmitted from one point to another [11]:

> *It might, for example, be a random sequence of digits, or it might be information for a guided missile or a television signal.*

The IT is a theory of communication that operates within a probabilistic framework, where messages are assigned probabilities. This branch of applied mathematics and computer science was developed to address theoretical issues related to optimally encoding messages based on their statistical structure [150, 151]. It differs from Wiener's work, which focused more on prediction and control, proposing a strategy to optimally filter noise from a received message. In contrast, Shannon's IT proposes a scheme in which the message is initially coded to withstand minimal degradation in the transmission channel and then decoded upon reception to obtain the original message [151], see Fig. 3.1.

Shannon's formulation of IT offers a unique perspective that extends beyond communication systems, the processing of information can describe a machine and its memory. It provides a framework that can even be extrapolated to the functioning of the human brain. For instance, a psychologist cannot fail to consider a scenario where the external world serves as the source of messages, and the mind, with its intricate network of neurons, acts as the receiver. In this analogy, our sensory organs such as eyes and ears play the role of communication channels [11].

This perspective aligns with ideas argued by Wiener in the 1940s, suggesting that the brain can be interpreted, at least in part, as a logical machine. Within the brain, intelligent behavior emerges from complex interactions of feedback loops involving neurons [152, 11]. While it's important to acknowledge that the human brain is the most complex organ in the body, IT offers a valuable method for enhancing our understanding of how the brain processes information.

It is well-known that DL algorithms, a subset of ML represented by the DNN, were initially conceived as computational models inspired by biological learning processes. These models aimed to replicate, to some extent, the learning mechanisms that take place within the human brain [21], see Section 2.2.

Much like the human brain, the IT framework offers a valuable perspective for comprehending the inner workings of the 'black box' represented by DL models. In this thesis, the various algorithms proposed have been evaluated and interpreted using an IT-based approach. This approach justifies the models' performance by tracing the flow of information through the different layers. Consequently, this approach enhances interpretability (Section 2.3) and ensures that model accuracy remains uncompromised.

3.1 Entropy

As mentioned earlier, Shannon did not have an interest in the meaning of a message. Nevertheless, he was interested in preserving the information that a message contains and for this purpose, it is necessary to quantify the amount of information that the message contains. To quantify the amount of information in any source, it is necessary to understand that there is a structure.

The structure is a concept that comes naturally, but defining it with precision is more difficult. Structure is influenced by a multitude of factors, making it nearly impossible to provide a comprehensive definition that encompasses all possible forms [150]. For instance, the distribution of colors in a row of an image may initially appear entirely random. However, when the image is examined with a focus on its underlying structure, discernible patterns emerge, revealing the information encoded within the image. A similar interpretation can be applied to a piece of music, where its structure, which is defined by different features such as the beat, dynamic, or tonality, can convey information about the upcoming note or the genre of the composition.

It is noteworthy that structure plays a significant role regarding information. Structures can exist within sources that initially appear to be random, and when such structures are present, they enhance the predictability of the source. Therefore, to quantitatively assess structure within randomness, the application of probability theory becomes essential. In summary, the fundamental idea behind IT is that discovering the occurrence of an improbable event provides more information than learning about a probable event. Consequently, the information content of a communicated message depends on the degree to which the content of the message is surprising or unexpected, i.e., the uncertainty of the message. As a result, IT built a bridge between information and uncertainty.

In 1928, Hartley proposed that when a symbol is selected in a message from a finite set of possible symbols S, the number of available choices can be considered as a measure of information

[153]. Hartley's definition of the 'amount of information' corresponds to the number of choices, or the cardinality of the set:

$$I_H(S) = \log_b |S|,$$

where $|S|$ represents the cardinality of the set. Essentially, Hartley introduced a measure of uncertainty, quantifying the number of possible distinct messages that could be produced.

The Shannon's IT was heavily influenced in Hartley's work, but he recognized the need to go beyond the cardinality of the message set to account for other fundamental aspects in the transmission of information [154]:

> *I started with information theory, inspired by Hartley's paper, which was a good paper, but it did not take account of things like noise and best encoding and probabilistic aspects.*

Shannon emphasized that the probability of selecting each message is crucial and must be incorporated into the formulation.

Considering the random variable X with possible outcomes $S_X \in \{s_1, \ldots, s_N\}$, having probabilities $p_X \in \{p_1, \ldots, p_N\}$ where $p(x = s_k) = p_k$ and N is the number of elements, Shannon noted that Hartley's information content holds when we have no prior knowledge about the data, i.e., when $p_k = 1/N$ [151]. To fully characterize an element s_k of a set of symbols S_x, occurring with different probabilities p_k, the information amount of an outcome s_k is given by:

$$I_H(s_k) = \log_2 \frac{1}{p_k}. \tag{3.1}$$

It is important to note that the logarithm base can be chosen as needed, but it is often calculated using the base-2 logarithm ($\log_2$) to express the measure in 'bits'. In Shannon's terminology, a 'bit' is defined as the fundamental unit of information, representing the amount of information conveyed by a binary question. Using the natural logarithm (log) is also a valid choice, representing a different scaling of the measured information.

The probabilistic perspective demonstrates that the amount of information associated with a message is a measure of its unpredictability, as it inversely depends on its probability. For a given message s_k, if $p_k = 1$, the information content is zero (indicating perfect knowledge). However, if p_k is small, the message is more unexpected, resulting in a higher information content. This aligns with Hartley's earlier formulation of information content.

Equation (3.1) represents the amount of information, but it does not express the information provided by the random variable X defined over the set of messages S_X. Shannon introduced the concept of **entropy** as a measure of uncertainty, where the information content is weighted by the probability of each message, yielding the average amount of information contained in the set of messages. The entropy of a random variable X is given by:

$$\mathcal{H}(X) = -\sum_{i=1}^{N} p(x_i) \log_2 p(x_i) = -E[\log_2 p(X)], \tag{3.2}$$

where $X \in \{x_1, ..., x_N\}$ composed by N number of samples and the probability mass function, $p(.)$, assigns a probability to each outcome in the samples space X, $p : X \to [0, 1]$. According to Eq. (3.2), the Hartley's amount of information proposed in 1928 [153] is satisfied when all individual samples in the sample space have the same probability of occurrence, as Shannon pointed out.

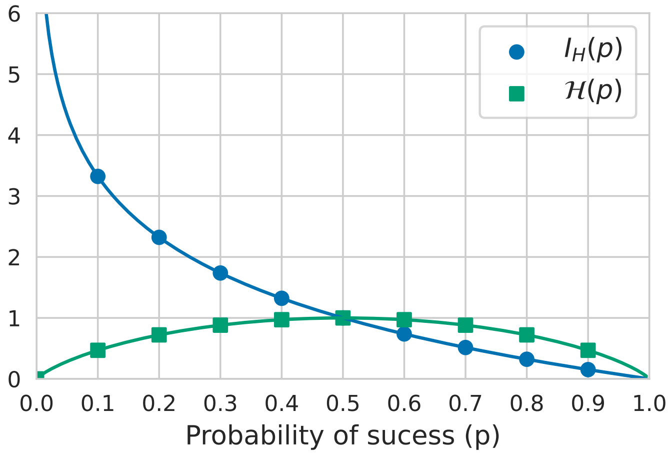


Figure 3.2: Information content ($I_H(p)$) and entropy ($\mathcal{H}(p)$) for a Bernoulli random variable measured in bits.

The entropy expression, as shown in Eq. (3.2), where the information content is weighted by probability, implies that rare events, which inherently contain a high amount of information, are undervalued due to their infrequency. Conversely, common events are underestimated due to their lower information content. This is illustrated in Fig. 3.2, which shows an example of the information content and Shannon's entropy using a Bernoulli random variable. This distribution is defined by a single probability, represented as p. Observing the information content, when $p \approx 0$, the event conveys more information, but the entropy undervalued this information due to its low frequency. In this specific case, for a Bernoulli random variable, the information content ($I_H(\cdot)$) and the entropy ($\mathcal{H}(\cdot)$) coincide at $p = 0.5$, which is the point of maximum information — 1 bit.

For a continuous variable, Shannon proposed that the correct continuous analogue of discrete entropy, referred to as *differential entropy*, corresponds to using the Probability Density Function (PDF) and replacing summations with integrals. However, this continuous approach has been the subject of debate because it is not invariant regarding changes in scale or changes in coordinates in general [150]. This differential entropy is typically expressed using the natural logarithm.

The entropy, which name is formed from Greek to mean "transformation content" [11], originally emerged in the field of thermodynamics in 1865 by Rudolf Clausius, building on the significant contributions of Sadi Carnot in 1824. However, it was Ludwig Boltzmann who later expressed entropy from a probability standpoint [11]. Before the IT, the entropy was associated to disorder or chaos, but both are crude and subjective concepts. However, Shannon's work in IT provided a rigorous foundation for understanding entropy as a measure of uncertainty, and ultimately, as a measure of information. The impact of information sciences extends far beyond the scope of this thesis, permeating various branches of science and leading to profound insights, as elaborated in books like [155].

3.1.1 Joint and Conditional Entropy

As previously mentioned, information theory operates within a probabilistic framework. It extends the concept of entropy to encompass more than a single sample space, introducing the notion of a joint probability distribution. This distribution characterizes the probability of events defined in terms of multiple random variables and represents the likelihood of various combinations of outcomes for these variables.

Consider two random variables, X and Y, with their joint distribution typically denoted as $p(x, y) = p(X = x \cap Y = y)$. This joint distribution provides information about the probability of X taking on a specific value and Y taking on a specific value simultaneously. The estimation of the entropy for this joint probability distribution results in the **joint entropy**, denoted as $\mathcal{H}(X, Y)$, which is calculated using the joint probabilities $p(x, y)$ in Eq. 3.2.

$$\mathcal{H}(X, Y) = -\sum_{x \in X} \sum_{y \in Y} p(x, y) \log_2 p(x, y) = -E_{X,Y}[\log_2 p(X, Y)]. \tag{3.3}$$

A fundamental property of joint entropy is that it is always less than or equal to the sum of the entropies of the individual random variables:

$$\mathcal{H}(X, Y) \leq \mathcal{H}(X) + \mathcal{H}(Y). \tag{3.4}$$

This inequality holds true unless the marginal distributions of X and Y are independent, in which case equality is established [150, 151].

Beyond joint entropy, IT introduces the concept of **conditional entropy** to assess the uncertainty of one random variable given knowledge of another. Conditional entropy, denoted as $\mathcal{H}(Y|X)$, quantifies the remaining uncertainty in Y when X is known. It can be calculated using conditional probabilities, providing valuable insights into the interdependence of random variables in a probabilistic system.

$$\mathcal{H}(Y|X) = -\sum_{x \in X} \sum_{y \in Y} p(x, y) \log_2 p(y|x) = -E_{X,Y}[\log_2 p(Y|X)]. \tag{3.5}$$

Building upon the concept of conditional entropy, it becomes possible to compute the joint entropy using the following relationship:

$$\mathcal{H}(X, Y) = \mathcal{H}(X) + \mathcal{H}(Y|X) = \mathcal{H}(Y) + \mathcal{H}(X|Y). \tag{3.6}$$

This relationship highlights the interplay between individual entropy and conditional entropy when assessing the overall uncertainty in a system characterized by multiple random variables.

Furthermore, drawing from Eq. (3.6) and Eq. (3.4), it can be established that [150]:

$$\mathcal{H}(Y|X) \leq \mathcal{H}(Y), \tag{3.7}$$

This inequality signifies that the conditional entropy of Y given X is always less than or equal to the entropy of Y itself. The equality in this relationship holds true if and only if X and Y are independent, emphasizing the fundamental role of independence in the context of conditional entropy.

3.1.2 Relative Entropy

Another crucial concept in IT is **relative entropy**, often referred to as KL Divergence [39] or information divergence. Relative entropy is a measure of how one probability distribution differs from another. The KL Divergence between two PDF, denoted as $p(x)$ and $q(x)$, is as follows:

$$D_{KL}(p||q) = \sum_{x \in X} p(x) \log \left(\frac{p(x)}{q(x)} \right). \tag{3.8}$$

KL Divergence indeed quantifies dissimilarity between probability distributions, which conceptually resembles a distance metric. However, it does not qualify as a true distance metric because it adheres to only one of the fundamental properties of distance metrics. Specifically, KL Divergence is non-negative, but it lacks symmetry, meaning that $D_{KL}(p||q) \neq D_{KL}(q||p)$ [151].

Indeed, relative entropy, or KL Divergence, provides a valuable measure of inefficiency when approximating one probability distribution with another. It quantifies the information lost during this approximation process.

While the asymmetry of KL Divergence can be advantageous in scenarios where directional coupling is a concern, there are situations where a symmetric information dissimilarity measure is preferred [151]. For such cases, there are other expressions of divergences with the same purpose, to capture the information lost when it is using an approximation, but with symmetric properties, such as the Jeffrey's divergence (J divergence) [156, 157] or the Jensen-Shannon divergence [158].

Relation with the Cross-Entropy

The CE, described by Eq. (2.5), is commonly used for classification tasks in ML. However, this expression can be decomposed as follows:

$$CE(p,q) = -\sum_{x} p(x) \log q(x) = \mathcal{H}(p) + D_{KL}(p||q). \tag{3.9}$$

Minimizing the CE corresponds to a minimization of the $D_{KL}(p||q)$.

This decomposition makes it more intuitive to understand what the cross-entropy measures. As previously mentioned, entropy is a measure of uncertainty, and cross-entropy corresponds to the uncertainty of the events described by $p(x)$ by observing $q(x)$. Consequently, reducing the cross-entropy implies that $q(x)$ is better modeling the events of $p(x)$.

3.1.3 Mutual Information

Mutual Information (MI) is a fundamental concept that quantifies the amount of information shared between two random variables. It quantifies the reduction in uncertainty about one variable given knowledge of the other, thereby capturing the probabilistic dependence between the two variables.

In Section 3.1.1, it was described that the conditional entropy, $\mathcal{H}(X|Y)$ (Eq. 3.5), quantifies the remaining uncertainty in X when Y is known. In IT, when information is transmitted through a channel (see Fig. 3.1), the conditional entropy is referred to as the 'equivocation' of the input, X, to the channel [150]. This represents the remaining uncertainty about X after observing the channel output Y. In conclusion, $\mathcal{H}(X|Y)$ can be considered as a measure of the noise injected in the channel in Fig. 3.1, since it contributes directly to the uncertainty of the input. In the context of

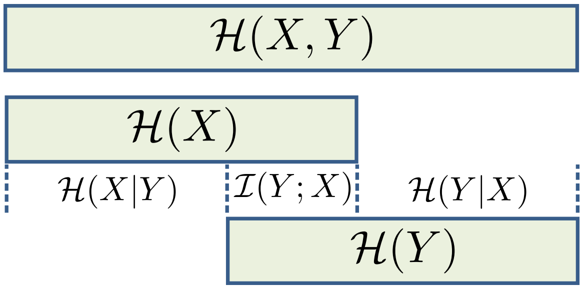


Figure 3.3: Diagram showing additive and subtractive relationships of the different information measures associated with correlated random variables X and Y.

this example, when $\mathcal{H}(X|Y) = 0$, it signifies that the channel is noise-free, whereas $\mathcal{H}(X|Y) = \mathcal{H}(X)$ implies that the noise completely obscures the transmitted input, there is no information transmitted through the channel.

Mutual Information quantifies the information obtained about an input variable after accounting for any degradation due to uncertainty, which is quantified by the conditional entropy:

$$\begin{aligned}\mathcal{I}(X;Y) = \mathcal{H}(X) - \mathcal{H}(X|Y) = \mathcal{H}(Y) - \mathcal{H}(Y|X) = \\ = \mathcal{H}(X) + \mathcal{H}(Y) - \mathcal{H}(X,Y).\end{aligned} \tag{3.10}$$

As it is observed in the last expression, MI can also be measured by taking the sum of the individual uncertainties $\mathcal{H}(X)$ and $\mathcal{H}(Y)$, and then subtracting the uncertainty of the joint distribution, represented by the joint entropy $H(X,Y)$. All the expressions that can be used to calculate MI are summarized in Fig. 3.3.

An alternative, more concise expression for MI is:

$$\mathcal{I}(X;Y) = \sum_{x\in X}\sum_{y\in Y} p(x,y)\log\frac{p(x|y)}{p(x)} = \sum_{x\in X}\sum_{y\in Y} p(x,y)\log\frac{p(x,y)}{p(x)p(y)}, \tag{3.11}$$

Observing the last expression in Eq. (3.11), it is evident that mutual information is related to the Kullback-Leibler Divergence (Eq. (3.8)). MI quantifies the disparity between the joint distribution of two random variables and what it would be if the two variables were independent, represented by the product of their marginal distributions:

$$\mathcal{I}(X;Y) = D_{KL}\left[p(x,y)||p(x)p(y)\right]. \tag{3.12}$$

Properties of Mutual Information

The MI is a non-negative quantity, represented as:

$$\mathcal{I}(X;Y) \geq 0, \tag{3.13}$$

where $\mathcal{I}(X;Y) = 0$ if and only if $p(x,y) = p(x)p(y)$, indicating that both random variables are statistically independent.

Furthermore, the expression $\mathcal{I}(X;Y)$ is symmetric with respect to the variables x and y. Thus,

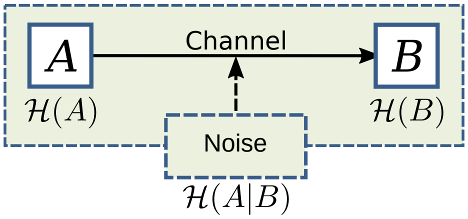


Figure 3.4: Entropy in an information channel.

by exchanging X and Y, we obtain the following result:

$$\mathcal{I}(X;Y) = \mathcal{I}(Y;X) \geq 0. \tag{3.14}$$

This property implies that the information the channel provides about X when observing Y is the same as the information the channel provides about Y when recognizing that X was sent. This symmetry can be intuitively deduced by examining Fig. 3.3.

3.2 Information Channel

The **information channel** serves as the medium through which information is transmitted from a source to a receiver. It is important to clarify that the channel itself does not introduce additional information into the transmitted data; instead, it functions as a conduit for the existing information [157, 150].

Ideally, the entropy associated with the channel should match the entropy of the message being transmitted through it, thereby accurately representing the amount of information carried by the message. However, in practical communication systems, the presence of noise in the channel is a common occurrence. Noise refers to unwanted signals or effects that accompany the desired signal carrying the information. This noise can introduce uncertainty and variability into the transmitted data, thereby increasing the entropy and reducing the amount of information that can be reliably extracted at the receiver (as illustrated in Fig. 3.1).

As discussed in Section 3.1.1, conditional entropy quantifies the remaining uncertainty of a random variable when another variable is observed. With this definition in mind, it becomes intuitive to think of conditional entropy as a measure of the 'noise' in the channels. Taking this into consideration, quantifying the MI is a crucial step in assessing the impact of noise on the channel.

Figure 3.4 provides an illustrative example of the transmission of information from a random variable A to another variable B through a channel. The channel can be described considering the following cases [150]:

- **noise-free**: This is the ideal scenario, where the remaining uncertainty should indeed be $\mathcal{H}(A|B) = 0$.

- **noisy**: The presence of noise in the channel, reducing the information carried by A, it is depicted when $0 < \mathcal{H}(A|B) < \mathcal{H}(A)$. This is a consequence of the increasing in the uncertainty provided by the noise.

- **ambiguous**: It means that the information of A is not effectively transmitted to B, and it

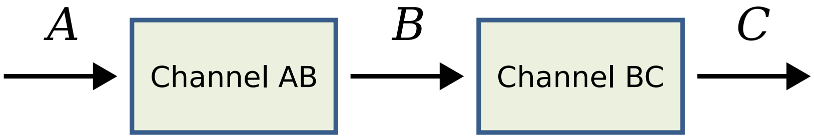


Figure 3.5: Cascade of two channel

remains completely obscured. In this case, the $\mathcal{H}(A|B) = \mathcal{H}(A)$.

The concept of MI can be useful in quantifying the amount of information carried by the channel (Eq. 3.10). For instance, in the case of a noise-free channel, $\mathcal{I}(A; B) = \mathcal{H}(A) - 0 = \mathcal{H}(A)$, demonstrating that the mutual information represents the total information carried from A to B without any loss due to noise.

3.2.1 Cascaded channel

It is common to transmit information through a cascade of different information channels rather than trough just one channel. In this case, intuitively, we would expect additive loss of information arising from the cumulative effect of uncertainty (or equivocation) from each channel in the cascade.

If we consider a channel that goes from A to B (Channel AB) and another channel that goes from B to C (Channel BC), as it is illustrated in the Fig. 3.5; thus we can define a cascade of channels as occurring when the following condition holds:

$$p(c_k|a_i, b_j) = p(c_k|b_j) \ \ \forall i, j, k. \tag{3.15}$$

This description corresponds to a Markov chain, a special type of discrete stochastic process in probability theory in which the probability of an event occurring depends uniquely on the immediately preceding event.

For the aforementioned example, the cascade of channel AB with channel BC, it is true that:

$$\mathcal{H}(A|C) - \mathcal{H}(A|B) \geq 0, \tag{3.16}$$

This inequation indicates that the remaining uncertainty of A regarding C, the conditional entropy, is greater or equal than the remaining uncertainty of A with respect to B. For the equality condition corresponds to the ideal scenario, a noise-free cascade channel.

Data processing inequality

As mentioned earlier, the noise level in a channel is quantified using MI. Considering the Eq. (3.16), it is commonly expressed by the Data Processing Inequality (DPI) [157]. The DPI represents the hierarchy of information flow through a system consisting of N cascaded channels. It is a series of inequalities expressed as:

$$\mathcal{I}(A; T_1) \geq \mathcal{I}(A; T_2) \geq ... \geq \mathcal{I}(A; T_N), \tag{3.17}$$

where each T_i corresponds to a specific channel within the system.

3.3 Information Source

In IT, the information source and the information channel are pivotal components, playing a fundamental role in the examination of information transmission and processing. Information sources refer to the devices or the underlying processes that generate random sequences of data.

Depending on the specific context and field of study, these sequences generated by the information source may be referred to as time series (in statistics), stochastic processes (in mathematics), or signals (in engineering) [150]. As illustrated in Fig. 1.1, an image can be treated as a type of signal, and it is considered that it contains information which is provided by the information source, the device and "situation" that generates the random sequence.

3.3.1 Memoryless Information Source

Memoryless information sources are a simplified model of information sources where each symbol or data point generated by the source is considered to be independent of any previous or future symbols. In other words, the occurrence of each symbol is not influenced by what happened before or what will happen next, and the probability of each symbol's occurrence remains constant over time.

In general, memoryless information sources are considered a naive approach to interpreting data sources, as they assume that there are no correlations between the source's outputs at different times. For instance, in natural language processing, the probability of a word appearing in a sentence can depend on the words that came before it. However, this approach can be useful in certain contexts, such as channel coding, where the aim is to transmit information over a noisy channel. In this case, using a memoryless source can simplify the encoding and decoding process [159].

3.3.2 Information Source with Memory: the Markov Source

On the other hand, an information source with memory, specifically a Markov Source, is a type of source in which there are correlations between the outputs of the source at different times. In a Markov Source, the generation of a symbol is conditioned by the current state of the source, and this state changes with each emitted symbol.

In a Markov Source, each symbol is emitted independently, but the probability of a specific symbol occurring depends on the current state of the source. This dependence is commonly modeled by a transition matrix, denoted Π, which describes how the source states evolve over time. The probability distribution of each symbol is determined solely by the current state of the source and is not influenced by past emitted symbols:

$$W^t = \Pi W^{t-1}, \tag{3.18}$$

where W^t is the state in a state t and Π is the $N \times N$ transition matrix. N is the number of possible states.

3.4 Information-Bottleneck Principle

The Information Bottleneck (IB) [160] is an information theoretical principle which describes the extraction of relevant information that an input random variable X contains about an output random variable Y. This principle formalized the problem, such as finding a short code for X that preserves the maximum information about Y. The IB, as claimed the author [160], may be considered a generalization of the rate distortion theory developed by Shannon [161].

Rate distortion theory, a branch of IT, provides the theoretical foundation for lossy data compression. Its main goal is to find an optimal compressed representation, denoted as Z, of a random variable X, which eliminates unnecessary details and redundancy. Rate distortion theory addresses the minimum MI required between the input and output of the channel to ensure that the distortion in the channel does not exceed a specified threshold D. In this approach, a distortion measure, noted as $d(x, z)$, has to be defined where distortion or 'cost' is defined to penalize the system based on the input of the channel $x \in X$ and its output $z \in Z$. This distortion measure can be defined using a more generic approach as a squared error (L_2 distance) or a Manhattan distance (L_1 distance) [162]. However, this distortion measure is usually adapted based on the purpose, for instance, in image compression the distortion is commonly measured using the Structural SIMilarity (SSIM) [163] which is based on the human perception [162, 164]. The average distortion is defined as:

$$\overline{d} = \sum_{x \in X} \sum_{z \in Z} p(z|x)p(x)d(x, z) = E_{X,Z}[d(X, Z)], \tag{3.19}$$

where $p(z|x)$ represents the channel.

Considering this, the goal of rate distortion theory is to identify the optimal channel in which the average distortion is acceptable, denoted as $p(z|x) : \overline{d} \leq D$. To achieve this, the theory aims to minimize the MI (transmitting less information) under the given constraint on the expected distortion. As a result, the rate distortion function that describes the minimum information rate required for the given average distortion D is defined as follows:

$$R(D) = \min_{\{p(z|x):\overline{d} \leq D\}} \mathcal{I}(X; Z). \tag{3.20}$$

It is important to note that there exists a monotonic trade-off between the rate of quantization and the expected distortion: as the rate increases, the achievable distortion decreases. Intuitively, tolerating a larger distortion allows for the use of lower information rates (more compression), and conversely, increasing the information rate enables lower distortion [150].

As described, the rate distortion theory depends on the distortion measure, which is often not readily available or difficult to define. The distortion measure quantifies the relevant information with respect to another random variable. For example, defining a distortion measure to identify a specific element in an image can be challenging. The concept of IB addresses this by defining the relevant information in X based on Y, leveraging access to the joint distribution $p(x, y)$ as part of the problem setup.

Similar to rate distortion theory, the goal of the IB framework is to capture the relevant information, but in this case, it is about capturing the relevant information of X with respect to Y. Using as reference a cascade of channel XZ with channel ZY, similar to the observed in Fig. 3.5, we aim to understand how much information about Y is contained in Z, where Z represents the compression of X. The amount of information about Y present in Z can be quantified as follows:

$$\mathcal{I}(Z;Y) = \sum_{y}\sum_{z} p(y,z)\log\frac{p(y,z)}{p(y)p(z)} \leq \mathcal{I}(X;Y), \tag{3.21}$$

Certainly, it is intuitive to understand that lossy compression cannot transmit more information than the original data. Just like the trade-off between rate and distortion in rate distortion theory, there's a trade-off here between compressing the representation and retaining valuable information. Obviously, there is no a single right solution for this trade-off [160]. The ideal scenario is to retain a fixed amount of valuable information about the relevant signal Y while minimizing the number of bits needed to represent the original signal X. Essentially, we are passing the information that X provides about Y through a 'bottleneck' created by the compact summaries in Z.

The formal definition of the IB for an optimal solution is given by the following equation:

$$\min_{p(z|x)}\ \mathcal{I}(X;Z) + \beta\mathcal{I}(Z;Y)\,, \tag{3.22}$$

where β serves as the Lagrange multiplier associated with the constrained meaningful information, and it ensures the normalization of the mapping $p(z|x)$ for every x. In other words, β controls the trade-off between compression and meaningful information. It plays a crucial role in determining how much emphasis is placed on preserving relevant information about the output variable Y while compressing the input variable X into the representation Z.

Therefore, the IB principle can be considered a rate distortion theory, where the distortion measure is quantified by the divergence of information (Eq. 3.12) [165]. What makes the IB principle particularly intriguing is its ability to offer a unified framework for various information processing tasks, encompassing prediction, filtering, and learning [160]. This versatility allows it to be applied to a wide range of fields and problems in which information plays a crucial role.

3.5 InfoMax Principle

The InfoMax principle, introduced by Linsker in 1988 [166], serves as a foundational concept in ANNs. It aims to describe the behavior of ANNs in terms of maximizing information preservation. The primary objective of this principle is to achieve the highest possible MI between the input variable X and the compressed representation Z, $\mathcal{I}(X;Z)$. This optimization principle is defined as follows:

$$\max_{p(z|x)}\ \mathcal{I}(X;Z) + \beta\mathcal{R}(Z)\,, \tag{3.23}$$

where $p(z|x)$ represents a channel, as it was described in the IB principle. The goal is to find the optimal channel that maximizes the mutual information while considering a regularization term $\beta\mathcal{R}(Z)$. This regularization helps control the complexity of the model, ensuring that the information preservation is balanced with model simplicity.

This principle is used in classical methods such the Independent Components Analysis (ICA) [167], a statistical technique used to extract independent sources from a set of observations that are linear mixtures of those sources. This is done by assuming that at most one subcomponent is a

Gaussian and that the subcomponents are statistically independent of each other.

Unlike the IB principle, which considers supervised settings with knowledge of the joint distribution $p(x, y)$, the InfoMax principle is specifically considered for unsupervised settings where the joint distribution is unknown. Consequently, the InfoMax principle primarily focuses on obtaining the best possible representation in the variable Z but it does not inherently reduce or discard less relevant information since there is not a random variable Y which provides that information.

Due to its unsupervised nature and the lack of joint distribution information, the InfoMax principle can be challenging to use when the goal is to obtain a compressed representation of input X that exclusively preserves the most relevant information while discarding less relevant details. This distinction highlights the differences in the objectives and applicability of the InfoMax principle compared to the IB principle.

In conclusion, the InfoMax principle does not satisfy the principle of minimum redundancy, as it compresses all the information contained in X, even if some features are redundant. Redundancy in information theory refers to information that is predictable and, as a result, does not provide new or additional information. However, In IT, the redundancy has a specific purpose, it provides tolerance to errors.

Therefore, it would not be accurate to consider the InfoMax principle as inherently 'worse' than the IB principle. In contrast, InfoMax principle can be interpreted as a generalization of the IB and both principles have their specific use cases and advantages. The InfoMax principle can be particularly convenient in scenarios where preserving all the information in the input is essential, such as in authentication systems [168].

3.6 Kernel-Based Entropy Estimator

In the 1950s, Alfred Renyi introduced a parametric family of entropies as a mathematical generalization of Shannon's entropy [151]. The Renyi's α-order entropy has the purpose of solving some discrepancies about the entropy estimation based on units that do not correspond to the bit. The Renyi's α-order entropy has been widely applied in ML [151, 169] combined with Parzen window density estimation [170] in order to estimate the PDF. Nevertheless, the estimation of PDF in high-dimensional data is a complicated task.

To address this, Giraldo et al. [171] proposed a framework for entropy estimation that leverages infinitely divisible kernels to construct a Reproducing Kernel Hilbert Space (RKHS). This approach utilizes the axiomatic characterizations of Renyi's entropy to define functionals on normalized positive definite matrices, aligning with these axioms without the prerequisite of known or estimated event probabilities. Such kernel-based estimators are not only mathematically robust but also computationally efficient [172], making them particularly suitable for DNNs where outputs frequently manifest in high-dimensional spaces. It is worth noting that similar to ML models based on RKHS, which guarantee a linear solution in this space, such as Support Vector Machine (SVM) [173], this method follows a quadratic computational complexity, $O(N^2)$.

Given a normalized Gram matrix $A \in \mathbb{R}^{N\times N}$ where N is the number of samples, the Kernel-based

Renyi's α-order entropy, $S_\alpha(.)$, is depicted as follows:

$$S_\alpha(A) = \frac{1}{1-\alpha}\log_2\left(tr(A^\alpha)\right) = \frac{1}{1-\alpha}\left[\sum_{i=1}^{N}\lambda_i(A)^\alpha\right] \tag{3.24}$$

where $\lambda_i(A)$ is the i^{th} eigenvalue of A and tr denotes the trace of the matrix. In the limit of $\alpha \to 1$ is reduced to the Shannon entropy of $\mathcal{H}(X)$.

This normalized Gram matrix is obtained by the gram matrix $K \in \mathbb{R}^{N\times N}$, where N is the number of samples. This matrix K is obtained by a kernel $k(.,.)$, using the random variable X, that can correspond to a Radial Basis Function (RBF). In this way, this matrix using the RBF kernel by the kernel trick [174] is depicted as:

$$\begin{aligned} K_{i,j} &= k(x_i, x_j), \\ k(x_i, x_j; \sigma) &= e^{\frac{1}{\sigma^2}||x_i - x_j||_F^2}, \end{aligned} \tag{3.25}$$

where $||.||_F^2$ denotes the Frobenius norm and σ the kernel width parameter of the RBF kernel. Finally, this normalized Gram matrix A is obtained by the following normalization:

$$A_{ij} = \frac{1}{N}\frac{K_{ij}}{\sqrt{K_{ii}\cdot K_{jj}}}. \tag{3.26}$$

3.6.1 Mutual Information

In order to obtain the entropy associated to two random variables X and Y, the joint entropy, Giraldo et al. [171] applies Hadamard product, which is a element-wise operation where $(A \circ B)_{ij} = A_{ij}B_{ij}$. If B is assumed to be the normalized Gram matrix of Y, the joint entropy between X and Y is defined as:

$$S_\alpha(A,B) = S_\alpha\left(\frac{A\circ B}{tr(A\circ B)}\right) \tag{3.27}$$

Finally, the MI estimation is identical to the Shannon's formulation, see Eq. (3.10), which can be defined as:

$$\mathcal{I}_\alpha(A,B) = S_\alpha(A) + S_\alpha(B) - S_\alpha(A,B) \tag{3.28}$$

3.6.2 Properties of Radial-Basis Kernel

In order to generate the RKHS, it is possible to use different kernels such as linear, bilinear or polynomial kernel [173]. However, in this thesis, the RBF kernel is the applied for the kernel-based entropy estimator, as it is illustrated in Eq. (3.25). This kernel has a hyperparameter to consider, the known kernel-width σ.

To construct the RKHS, various types of kernels can be employed, such as linear, bilinear, or polynomial kernels [173]. For the purposes of this thesis, however, the RBF kernel is utilized for the kernel-based entropy estimator, illustrated in Eq. (3.25). A crucial hyperparameter within this kernel is the kernel width, σ, which significantly influences the performance of the entropy estimator.

The choice of σ in the RBF kernel is pivotal because it determines how effectively the kernel captures the data's underlying structures. The following properties associated with the RBF kernel highlight the sensitivity of the estimator to changes in σ [175]:

$$\begin{aligned}
&\lim_{\sigma\to 0} S_\sigma(A) = \log(N) \\
&\lim_{\sigma\to 0} \mathcal{I}_\sigma(A;B) = \log(N) \\
&\lim_{\sigma\to \infty} S_\sigma(A) = 0 \\
&\lim_{\sigma\to \infty} \mathcal{I}_\sigma(A;B) = 0.
\end{aligned} \tag{3.29}$$

These equations indicate that σ controls the estimator's operating point relative to its bounds. Specifically, if σ is too small, the estimator tends to approach $\log(N)$, reflecting a state of high entropy that suggests minimal effective differentiation between data points. Conversely, a considerable σ leads to an entropy estimate of 0, implying that all data points are treated as virtually identical, hence underestimating the true diversity or variability within the dataset. Properly tuning σ is therefore essential to balance sensitivity and specificity, ensuring that $S_\sigma(A)$ and $\mathcal{I}_\alpha(A;B)$ provide meaningful insights into the structure and relationships in the data.

3.6.3 Radial-Basis Kernel-width Estimation

As it has been observed in the previous section, by using the RBF kernel for entropy estimation, it is necessary to define a hyperparameter known as kernel-width, σ. In this section, it will be indicated some methods commonly used for this purpose.

Silverman's rule of thumb

Silverman's rule of thumb [176] is a widely used heuristic for selecting the bandwidth of the RBF kernel, prized for its computational simplicity. However, it is crucial to exercise caution when employing this rule, as it can sometimes produce significantly inaccurate estimations. The rule is based on assumptions about the data's distribution that may not hold in all cases, particularly in datasets with non-normal distributions or significant outliers. Therefore, while Silverman's rule offers a quick and easy method for setting the kernel width, it is important to validate its effectiveness within the specific context of the data being analyzed to avoid potential inaccuracies in the results.

Silverman's rule of thumb is articulated in the following mathematical form for setting the kernel width σ:

$$\sigma^* = \gamma N^{(-1/(4+d))}, \tag{3.30}$$

where γ is typically derived from the empirical standard deviation of the data, and N is the number of data points, while d represents the dimensionality of the data.

While this method is straightforward and commonly used, it has been refined to better accommodate datasets of varying dimensionalities. Tapia et al. [175] introduced a modification known as the Normalized Silverman's rule, which adjusts the kernel width to account for the dimensionality of the data:

$$\sigma^* = \gamma N^{(-1/(4+d))}\sqrt{d}. \tag{3.31}$$

This adjustment, multiplying by $\sqrt{d}$, aims to mitigate the effect that increasing dimensionality has on the mean square distance between points in the dataset. This consideration is crucial because as dimensionality increases, the standard Silverman's rule may not scale the kernel width appropriately, potentially leading to less effective density estimation and higher error rates in high-dimensional spaces.

Kernel-Alignment Loss

The method of kernel alignment, proposed by Wickstrom *et al.* [169] for kernel-based entropy estimation, introduces an advanced technique for optimizing the kernel width σ in supervised learning settings. This technique is particularly useful when dealing with the Gram matrices of high-dimensional input data K_σ and a lower-dimensional target K_Y that σ can be estimated by the previous Silverman's rule. The kernel alignment $\mathcal{A}(K_\sigma, K_y)$ [177] focuses on maximizing the alignment between these matrices to determine the optimal σ, as described by the following equation:

$$\begin{aligned} \sigma^* &= \underset{\sigma}{\operatorname{argmax}} \, \mathcal{A}(K_\sigma, K_y), \\ \mathcal{A}(K_a, K_b) &= \frac{\langle K_a, K_b \rangle_F}{||K_a||_F \cdot ||K_b||_F}, \end{aligned} \tag{3.32}$$

where $||.||_F$ and $\langle ., . \rangle_F$ denotes the Frobenius norm and inner product, respectively.

This σ estimation corresponds to an iterative process [169] and an exponential moving average is employed:

$$\sigma_t = \beta \sigma_{t-1} + (1 - \beta) \sigma_t^* \tag{3.33}$$

where β is a smoothing factor that balances between the previous value σ_{t-1} and the newly optimized value σ_t^*.

Part II

Research Contribution

Chapter 4

Information-theoretical Evaluation in Deep Learning

> If a machine is expected to be infallible, it cannot also be intelligent.
>
> *Alan Turing*

This chapter presents an interpretation that provides a general intuition about how the DL models works. This interpretation will provide an understanding of the problem-solving mechanism used in these models, which will enable the design of methodologies for validating these models based on this approach. Although this interpretation can be extended to different contexts, throughout this chapter, we will focus on a critical scenario: medical screening supported by AI, where the interpretability of the models is particularly relevant.

The use of medical imaging technologies has become integral to diagnostic decision-making and treatment planning. CV, a subfield of AI, offers significant advantages in enhancing diagnoses, as it enables more accurate and faster analysis of medical images, including CT scans, X-rays, MRI, and mammograms, even sometimes surpassing human capabilities. CV seeks to replicate the human visual system, providing specialized algorithms for tasks like image recognition and object tracking. However, the emergence of DL has effectively replaced numerous of these conventional methods, particularly in tasks such as classification and object detection, drawing considerable attention recently.

At the same time, the rapid increase in radiological imaging data has created a pressing need for greater efficiency in analysis, especially considering the limited number of trained radiologists available. As the workload of radiologists has surged significantly, DL models have emerged as practical clinical decision support systems, capable of handling extensive datasets and improving diagnostic accuracy. To keep up with clinical demands, research by Hosny *et al.* [5] shows that the average radiologist would need to interpret an image every 3–4 seconds.

Deep Learning models possess the capability to process extensive datasets, addressing challenges like disease detection in medical images. Studies, such as Liu *et al.* [6], have demonstrated that DL models deliver performance on par with healthcare professionals. Consequently, DL models offer adaptable and highly accurate outputs, mitigating human bias, reducing associated costs, and reducing the time burden of demanding tasks. Consequently, DL-powered AI can serve as practical clinical decision support systems, aiding radiologists in real-world scenarios. A seamlessly integrated AI component within the medical imaging workflow has the potential to enhance efficiency, reduce errors, and accomplish objectives with minimal manual intervention [5]. This can be achieved by providing radiologists with pre-screened images and identifying relevant features, ultimately streamlining the diagnostic process. In this way, this technology can contribute to reducing the heavy workload of radiologists today.

Despite the advantages provided by the DL models, its implementation and use in healthcare settings have not been completely achieved. More studies are required to consider the integration of these algorithms in the healthcare setting [6, 178]. There are several ethical challenges that have to be addressed. The interpretability of the model is one of the most challenging ethical problems that it can be found in the medical domain. The 'black-boxes', such as the DNN, lack of transparency, making it challenging to provide mechanistic explanations and potentially diminishing their reliability and trustworthiness [179]. In conclusion, understanding the model's problem-solving mechanism is essential for gaining confidence in its performance, specially in clinical scenarios since the clinical decision-making relies heavily on evidence interpretation [179, 180].

In addition, there is another relevant challenge in this domain is the lack of sufficient labeled data [180], known as the *data challenge*. This is caused by the difficulty of performing a systematic collection of data to create large and well-curated datasets for training DL models. As a result, it is possible to obtain poorly representative training data sets, which can introduce biases into the algorithms [179]. For instance, this concern was addressed by Leslie et al. [181] since patterns of systemic health inequity and bias could be embedded in AI systems designed to combat the COVID-19 pandemic.

The ethical challenge of interpretability is compounded by the data limitations, which can result in overfitting when training datasets are not sufficiently large. These challenges are prominent in DL models due to their high number of parameters, which can lead to overfitting if the training dataset is not sufficiently large. As a consequence, the overfitted model will perform well on the training data but poorly on new data. However, there are techniques to mitigate this problem, such as transfer learning [182] or data augmentation [183, 184]. Furthermore, these problems are magnified by the current trend to deeper neural networks [185, 186, 187], where the vanishing gradient problem [188] is highly pervasive. Although skip-connections, which connect the outputs of layers at different levels of depth (different representations), have been proved to overcome this limitation and provide other benefits during the training process [189].

Considering the aforementioned problems, this thesis puts special emphasis on enhancing the interpretability of DL models to improve their reliability through an analysis based on the interpretation. This interpretability allows obtaining a more in-depth understanding of the problem-solving mechanism and establish specific characteristics that a robust model has to satisfy. The approach used for enhancing the interpretability of these models is based on information, describing the internal mechanism of the DNNs based on how the information is propagated along the different layers. In addition, this chapter addresses the use of a limited labeled dataset, highlighting the presence of data challenges in the experimental results described.

This chapter is structured as follows: it begins by introducing the IT perspective of DNNs, which will be evaluated in a case study where the aforementioned data challenge is present. A methodology based on this perspective is proposed for validating the models in the different experiments. In addition, the analysis is extended to self-supervised approaches using the AE architecture. Finally, the effect of the dropout technique on DL models is studied, and an IT interpretation is provided to describe the technique.

4.1 Information Theoretical Perspective for Deep Learning Models

DNNs have been particularly relevant for supervised learning tasks; however, the theoretical understanding of DL models remains unsatisfactory. As previously discussed, this lack of understanding raises several ethical concerns in critical fields, such as healthcare. Enhancing interpretability is essential for the integration of these algorithms into such fields. Improved interpretability can leverage the use of these models, which often outperform simpler and more interpretable ML methods that have significant limitations in handling the complexity and variability of real-world data.

This thesis draws significant inspiration from the work of Tishby and Zaslavsky, who provided crucial insights using information-theoretic concepts [97]. This perspective aims to interpret the DL models as a communication channel, similar to the analogy described in Section 3 where this approach was extrapolated to the human brain. This interpretation, from an engineering-perspective, provides theoretical limits for the optimal transmission of information through a DNN, offering a framework for understanding optimal architectures and features of each layer.

Tishby and Zaslavsky formulated the supervised DL model as an information-theoretic tradeoff between compression and prediction, using the IB framework (see Section 3.4). In other words, DNN layers aim to find a maximally compressed mapping of the input variable that preserves as much as possible the information on the output variable. Therefore, considering the IT-perspective, this approach describes any this DL model by the quantifying the information in the layers, ignoring the 'meaning' of the extracted features (see Section 1.1). This engineering-oriented perspective, distinct from 'human-readable' interpretations used in 'explainable ML' (see Section 2.3), provides a basis for designing optimal DNN architectures and verifying well-fitted models.

In their work [97], Tishby and Zaslavsky describe the structure of DNNs as a Markov cascade, involving intermediate representations between the input and output layers. This structure can be interpreted as a cascade channel, as described in Section 3.2.1. Furthermore, they formalize the aforementioned trade-off between data compression and prediction within DNNs using the IB framework, depicted in Eq. (3.22), and subject to constraints defined by the Markov chain.

Regarding the information characteristics in the different layers, as the workflow of a typical DNN is described by a Markov chain, the input of a layer T_i depends exclusively on the output of the layer T_{i-1}. Here, the i^{th} hidden layer is denoted as a single multivariate variable. In this way, the information propagation is defined by the DPI, as shown in Eq. (3.17).

When considering the IB framework and the Markov chain $X \rightarrow Z \rightarrow \hat{Y}$, we have the inequality $\mathcal{I}(X;Y) \geq \mathcal{I}(\hat{Y};Y)$, where Z represents the bottleneck latent space (a lossy-compression of input X), $\hat{Y}$ is the prediction of output Y based on X. This inequality means that the prediction obtained from $f : Z \rightarrow Y$ cannot be more informative about the class than the original information contained in the image. Implicitly, this indicates that the lossy compression Z not only results in a loss of information from X, but also leads to a loss of information about Y, i.e., $\mathcal{I}(X;Y) \geq \mathcal{I}(Z;Y)$.

The aforementioned inequality in the Markov chain can be extrapolated to a DNN, even though the optimal lossy compression Z may not be explicitly represented in the DNN architecture. Therefore, for a DNN with L hidden layers, the information propagation should satisfy the following inequality:

$$\mathcal{I}(X;T_1) \geq \mathcal{I}(X;T_2) \geq ... \geq \mathcal{I}(X;T_L), \tag{4.1}$$

$$\mathcal{I}(T_1;Y) \leq ... \leq \mathcal{I}(T_{L-1};Y) \leq \mathcal{I}(T_L;Y). \tag{4.2}$$

These inequalities describe the flow of information from the input layer to the bottleneck latent space in the forward direction and from the bottleneck to the output layer in the backward direction within a DNN. This aligns with the concept that as data is processed through the layers of a DNN, the amount of relevant information may decrease (according to the DPI) respect to the input, but the deeper representations in each successive layer become more useful for the ultimate prediction task (aligned with the objective of supervised learning). In addition, these inequalities are describing that the bottleneck has a balance between the information of the input and the output, as it is defined in the IB framework.

The aforementioned description is regarding about the supervised DNNs but the IB could not describe the non-supervised models, such as the AE. AEs are widely recognized for their ability to compress input data X into a latent representation through self-supervised learning (see Section 2.2.3). This process is achieved by minimizing the difference between the input, X, and the reconstructed output data, denoted as $\hat{X}$. This optimization objective can be interpreted as an effort to maximize the MI between X and the reconstruction $\hat{X}$.

As deduced, AEs constitute a special subset of DL architectures, typically forming a Markov chain. However, due to the lack of a labeled dataset that maps the input to an output, the IB framework cannot describe the learning process carried out in these architectures.

In essence, AEs operate in alignment with the InfoMax principle, which is detailed in Section 3.5, while considering constraints established by the Markov chain, to achieve their main purpose of compressing high-dimensional data. By maximizing the MI between the input X and the output $\hat{X}$, AEs effectively learn to extract the most informative features from the input data and employ them for the purpose of reconstructing the original data, $\hat{X}$. However, given that there is no a variable that describes the relevance of the information, there is no guarantee that less relevant information or features will be discarded in this process. In other words, AE aims to comprise the whole information of X but do not prioritize which information should be discarded in this lossy compression. The information loss in the latent space Z is a consequence of the data compression, given that Z is dimensionally smaller than the input.

In [190], Yu *et al.* propose the information theoretic methodology to understand the dynamics of learning and the design of the AEs, providing fundamental insights into the layer-wise flow of information. Due to the symmetric nature of such architectures formed as a Markov chain, authors claimed that AEs must meet specific requirements for the DPI, as it is described in Eq. (4.3) and

Eq. (4.4):

$$\mathcal{I}(X;E_1) \geq \mathcal{I}(X;E_2) \geq ... \geq \mathcal{I}(X;E_{L-1}) \geq \mathcal{I}(X;Z), \tag{4.3}$$

$$\mathcal{I}(\hat{X};D_1) \geq \mathcal{I}(\hat{X};D_2) \geq ... \geq \mathcal{I}(\hat{X};D_{L-1}) \geq \mathcal{I}(\hat{X};Z). \tag{4.4}$$

In the aforementioned inequalities, the encode path comprises the layers E_i, the decode path consists of the layers D_i, and Z represents the bottleneck in the AE, denoting the compressed latent space. The number of layers in the encode path and decode path, including the compressed latent space, is L layers. These properties are established due to the AE's symmetry, where the decoder always aims to "undo" the transformations carried out by the encoder [190].

Based on the two DPI described in Eq. (4.3) and (4.4), there is a second type of DPI associated with layer-wise MI, which indicates that the divergence between the encode and decode layers at the i^{th}-level does not increase compared to the previous layers. In other words, since entropy decreases as you go deeper into the network (due to compression) and the divergence between the corresponding encode and decode layers is minimal, the following inequality must hold, as shown in Eq. (4.5):

$$\mathcal{I}(X;\hat{X}) \geq \mathcal{I}(E_1;D_1) \geq \mathcal{I}(E_2;D_2) \geq ... \geq \mathcal{I}(E_{L-1};D_{L-1}) \geq \mathcal{I}(X;Z). \tag{4.5}$$

Yu and Principe [190] indicate that both encoder and decode enforce the entropy in the bottleneck layer, hence the role of the AE is to maximize the entropy in the hidden layer. In this way, the $\mathcal{I}(E_{L-1};D_{L-1}) \simeq \mathcal{H}(Z)$, which is a consequence of the InfoMax principle since it means that the $H(Z|E_{L_1}) \simeq 0$. It implies that there is no uncertainty about Z given E_{L-1}. As a result, following the InfoMax principle with the mentioned constraint, an AE can be defined as follows:

$$\max_{p(z|x)} \; \mathcal{I}(X;Z) + \beta\mathcal{H}(Z). \tag{4.6}$$

4.1.1 Data Processing Inequality: An intuitive explanation

DNNs are capable of learning optimal representations from the input X to solve tasks described by Y. These representations capture complex and non-linear interactions among the features of the original space, interactions that might not be evident from the original space itself. Consequently, it becomes challenging to understand what these representations describe and whether they are representative for obtaining Y. The DPI illustrated in Eq. (4.1) and (4.2) describes this phenomenon:in the initial layers, the relationship with the input X is more evident, while the output Y becomes more prominent in the later layers.

The explanation methods used in DL, see Section 2.3.3, aim to describe the behavior of the model using the original space X or Y. For instance, methods like DeepLIFT or Grad-CAM aim to identify the relevant features in the original input space that are representative for the task. Figure 2.8 illustrates the use of the Grad-CAM algorithm to evaluate the attention mechanism that we introduced in [102] for skin-cancer screening, by detecting the region of interest in the image used by the DNN. In contrast, explanation methods based on perturbation, such as LIME, aim to describe the behavior of the model in the output Y, analyzing the impact of the input features on the model performance.

The reason behind why explanation methods focus on input or output, beyond the obvious

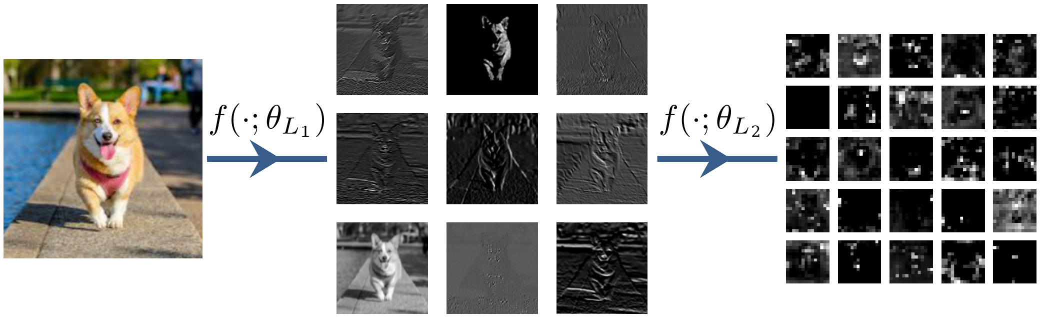


Figure 4.1: Illustration of the features obtained in a convolutional network where the first layer has the parameters θ_{L_1} and the second layer corresponds to the parameters θ_{L_2}. While in the first convolutional features obtained can be deduced some features, such as illumination or borders in the original, others are complicated to deduce. This deduction is more challenging in the second layer.

reasons, can be deduced through DPI. Using Eq. (4.1) as reference, the first layer, T_1, is that one that contains more information about X but less information about Y. However, the T_1 is assumed to have higher entropy, meaning it contains more average information than Y. Therefore, while T_1 has sufficient information, the relationship with Y is not evident in this representation, which is a consequence of the desired property in these models: generalization.

The first layers in DNNs have to deal with higher uncertainty than the latter layers, given that the input has higher variability than the output, which corresponds to a label. As a result, a first layer T_1 has to obtain a more general data structure that contains the maximum information on the input X, extracting features present in this input. As a result, the new structure obtained in T_1 is less evident for describing the target Y. Recall that in IT, the data structure plays a significant role regarding information (see Section 3.1). In subsequent layers, the information in T_1 is 'reorganized' and 'transformed' to make relationships with Y more evident. This transformation process results in an assumed loss of information, reducing the entropy of the following output layers but increasing the MI with the output.

This assertion is illustrated in Fig. 4.1, using an image input (the dog) and there are two convolutional layers (see Section 2.2.1). The figure shows that as we move to deeper layers, there is not only a loss of information but also a reduction in the evident connection to the input X. In the filters of the first convolutional layer, $f(\cdot;\theta_{L_1})$, some features of the original image, such as illumination over the dog or border detections, are evident. Nevertheless, these features do not explicitly describe a dog, for instance, the textures you find on the ground do not have to be relevant to the dog's detection, but this information is contained in the input X. It is a consequence of the expected generalization in the DNN. Regarding the second convolutional layer, $f(\cdot;\theta_{L_2})$, the 25 filters highlight features in the image that are considerably more difficult to describe in X. Additionally, this loss of information can be observed by noting the output of the second layer, which corresponds with a decrease in entropy. Although it is more complicated to illustrate, this second layer is obtaining a description that increases the MI with the output Y.

4.1.2 Information Plane

The Information Plane (IP) [97, 191] is a visualization tool that aims to provide a illustrative understanding of the dynamics of training, learning processes, and the internal representations

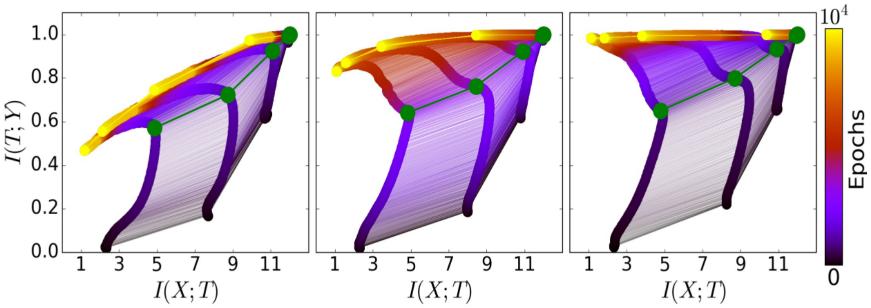


Figure 4.2: Information Plane illustrating the evolution of layers in a DNN across different training data samples. The left panel shows results with 5% of the data, the middle with 45%, and the right with 85%. Image obtained from [191].

during the training process of DL models. Popularized by Shwartz-Ziv and Tishby [191], the IP leverages the fundamental aspects of the IT perspective, such as the IB theory and the DPI, to analyze DNNs.

In the IP, the two key quantities, $\mathcal{I}(X;T)$ and $\mathcal{I}(T;Y)$, are plotted on the coordinate axes, where T represents a layer in a DNN. $\mathcal{I}(X;T)$ quantifies the mutual information between the input data X and the layer T, while $\mathcal{I}(T;Y)$ measures the mutual information between the layer T and the output labels Y. This allows for visualization of whether the DPI is satisfied within the proposed DNN.

Figure 4.2 shows an example extracted from [191] that illustrates the IP estimation of a DNN. The results in Fig. 4.2 were obtained using a simple DNN with a reduced number of neurons and layers, trained on a synthetic dataset. The experiment was conducted with a progressively increasing number of samples. The figure demonstrates that the DPI is violated, particularly Eq. (4.2), except in the final experiment with most of the data. Given the simplicity of the model and data, it is difficult to observe the aforementioned generalization in the early layers, but this will be explored further throughout the thesis.

The analysis using the IP focuses on how the MI on the depicted axes evolves throughout the training process, shedding light on how the model processes and refines information as training progresses. The following section provides a concise overview of the phases involved in the training process.

Information Plane Evolution

The evolution of the IP estimation during the training process encompasses two distinct phases, as outlined by Shwartz-Ziv and Tishby [191]. In the initial phase, known as the *fitting phase*, the layers of the network enhance the information pertaining to the target variable Y (i.e., the MI $I(T;Y)$ increases). Subsequently, the training progresses into the *compression phase*, during which the layers of the network progressively reduce the information associated with the input data X (i.e., the MI $I(X;T)$ decreases).

This compression phase is intricately linked to the concept of generalization, wherein extraneous or irrelevant information is compressed, thereby mitigating the risk of overfitting [191, 169]. However, the precise relationship between compression and generalization remains a topic of ongoing discussion within the research community [192].

4.2 Deep Learning Insights for Classification

As mentioned in Section 4.1, this thesis draws significant inspiration from the work of Tishby and Zaslavsky [97], a perspective that has been validated in various works [191, 169, 192, 190]. In this section, we explore the validation of this interpretation through a case study in which the dataset is limited, meaning the number of labeled samples is scarce.

Here, we propose the use of the kernel-based entropy estimator [171], as described in Section 3.6, which is both computationally efficient and mathematically robust. Using this estimator, we will compute MI estimations to generate the IP, which will be employed to validate the DNN models using the IT-perspective and the conditions depicted in Section 4.1.

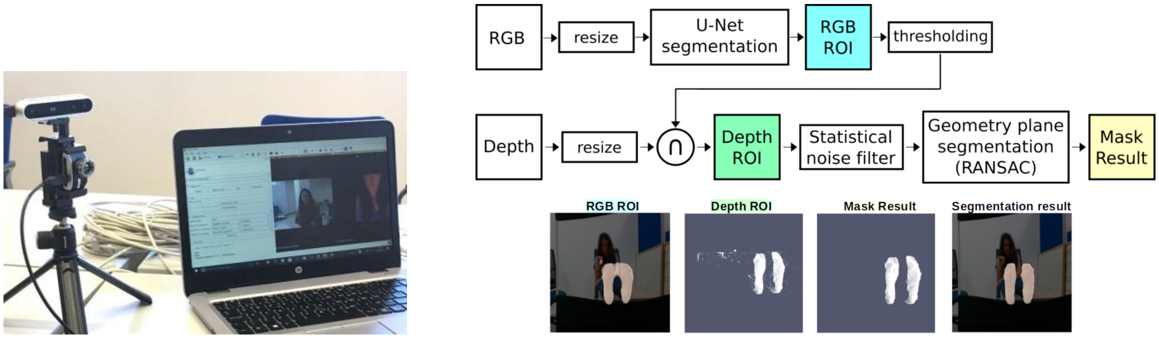
Given the lack of sufficient labeled data, there are several limitations that affect how well a DNN can perform. The trade-off between the number of samples used for training and those used for model validation makes it difficult to assess whether the model is overfitting. In other words, classical classification metrics are not feasible for evaluating the model due to the small test set, and the limited number of training samples can lead to overfitting. However, we will propose various architectural configurations and attempt to identify limitations in these approaches using the IP, with estimations obtained through the kernel-based entropy estimator, as an evaluation tool.

4.2.1 Case Study and Dataset: Diabetic Foot Case with Limited Labeled Samples

Healthy humans keep their inner temperature within a narrow range of 36.8ºC. Consequently, skin temperature measurement is a commonly employed method for diagnosing various diseases, including diabetes mellitus. Diabetes is a chronic condition resulting from elevated glucose levels due to the body's ineffective production or utilization of insulin [193, 194]. The management of this disease necessitates ongoing medical attention to prevent complications, as it has the potential to gradually harm the heart, blood vessels, eyes, kidneys, and nerves over time. [194]. According to the International Diabetes Federation, approximately 537 million adults aged $20 - 79$ were living with diabetes in 2022, and this number is projected to reach 643 million by 2030 [193]. Additionally, it is estimated that 240 million individuals have undiagnosed diabetes worldwide [193]. This underscores the fact that diabetes is emerging as one of the most rapidly growing global health crises of the 21^{st} century.

One of the leading causes of non-traumatic amputations is foot ulcers, a medical complication often experienced by diabetic patients as a result of diabetes mellitus [193]. Referred to as diabetic foot, this condition is frequently associated with neuropathy disorders and peripheral vascular disease affecting the lower limbs. This is because the temperature of extremities is significantly influenced by peripheral blood flow vessels [194, 195]. Abnormal temperature readings can serve as an early indicator of ulcer formation. As a result, temperature monitoring can be employed for the early detection and evaluation of the disease before other symptoms become apparent [194].

Infrared Radiation (IR) thermography, or IR thermal image, is a non-invasive imaging modality that maps the surface temperature of a body [195], and it is an alternative to the medical that uses a contact sensor for measuring the skin temperature. In addition, the appearance of low-cost devices, based on microbolometers, has favored the use of infrared cameras as supplementary diagnostic tools. These affordable devices allow clinical practitioners to measure the superficial temperature



(a) (b)

Figure 4.3: In (a), the acquisition system employed to generate the foot image database is composed by an Intel RealSense D415 camera for visible as well as depth of the pixels (RGB-D), and a Thermal Expert TE-Q1 Plus for the thermal infrared (IR) modality. In (b) is illustrated the proposed algorithm which is implemented for the segmentation of the soles of the feet and subsequently applied to thermal images

under a controlled ambient environment [196] providing multiple recordings at short time intervals. The viability of infrared sensors to analyze anomalous temperature patterns in diabetic diseases has been previously reported [195, 197, 198].

Existing literature has used foot temperature asymmetry as a diagnostic tool and for monitoring the severity of diabetic foot complications [199, 200]. This assessment method involves calculating the difference between the mean plantar temperature of the affected foot and the unaffected foot, following a standard contralateral analysis. Furthermore, this analysis is complemented by an examination of temperature patterns to identify anomalies that could signify potential ulcerations. Basically, if this temperature difference surpasses a certain threshold, an alarming sign is set.

In alignment with this approach, several studies have been conducted as part of this thesis. For instance, in the work [201], an algorithm based on U-Net for segmentation has been developed. This algorithm is applied to extract the feet from thermal images acquired using a custom imaging system, as depicted in Fig. 4.3a, to obtain images of the soles of the feet. Figure 4.3b illustrates the developed algorithm, wherein segmentation is derived from the RGB-D image, generating an image with both visible and depth information for each pixel. This work was extensively compared with the state-of-the-art in [202], where the proposed approach, optimized using spatial information, was found to exhibit the best robustness in terms of spatial overlap, accuracy, and precision. Once the segmentation is obtained, this mask is then applied to the IR image.

Subsequently, a standard morphological model was proposed for a group of healthy subjects through spatial image registration, which can serve as a template for further analysis of temperature patterns in diabetic foot disorders [203]. Nevertheless, determining the threshold in contralateral analysis for Diabetic Foot Ulcer (DFU) detection remains a subject of debate. In patients with diabetic foot infection, a plantar foot temperature asymmetry of 1.35 °C is considered sufficient to prompt urgent treatment [204], whereas a threshold of 2.2 °C is typically used for assessing an inflammatory process and the impending development of a foot ulcer [205]. In our study [203], it was observed that some healthy subjects exhibited a mean temperature difference exceeding this threshold. Consequently, it is crucial to acknowledge the heterogeneity among individuals and advocate for the establishment of personalized thresholds and treatments.

In conclusion, the early-stage detection of DFU [202, 206] constitutes a challenging medical analysis scenario for the classification task. The need for personalized analysis for each patient can be demanding, but recent advancements in AI may offer a potential solution by leveraging temperature patterns as a more generalized approach compared to traditional contralateral analysis. While DL models have exhibited potential in wound classification [207, 208], and specifically in diabetic ulcer identification [209, 210], the majority of existing research concentrates on ulcers that are already visible. The use of thermal imaging for DFU detection is an emerging area of research, where AI can play a fundamental role in detecting ulcers in an early stage. However, it has to consider the ethical problems aforementioned, the mechanism of the DL model and the lack of sufficient data.

In this case study, our goal is to provide an analysis based on the IT perspective described in Section 4.1, while addressing the significant challenge we face in this work: the data limitation. This data-related challenge will be detailed in the following section.

Dataset

Currently, one of the largest datasets oriented towards DFU detection is the Diabetic Foot Ulcers Grand Challenge (DFUC 2020) [211]. This dataset primarily focuses on identifying ulcers that are already visible and comprises 4000 visible images, offering close-up views of the foot. The dataset is evenly split into training and testing subsets, with 2000 images allocated to each one. It has been extensively utilized and evaluated using various state-of-the-art models [212], with the model presented in [213] achieving the highest performance (an F1-Score of 74.3%). Additionally, another publicly available dataset includes 754 visible images of healthy and diabetic ulcer skin from different patients [214]. This dataset has undergone comprehensive evaluation, achieving an F1-Score exceeding 97% in the best case [215]. Nevertheless, the differences in imaging modality and the scenarios captured make these datasets unsuitable for our intended application, which is the early-stage detection of DFU.

As previously mentioned, foot temperature analysis has been employed as a monitoring tool for diabetic foot complications. In the realm of state-of-the-art research, there exists a dataset known as the INAOE (Instituto Nacional de Astrofísica, Óptica y Electrónica) thermogram dataset [206]. This dataset consists of thermographic samples obtained from 167 volunteers, with 122 being diabetic subjects and 45 being non-diabetic individuals. Among the volunteers, 105 are female and 62 are male, with a mean age of 27.76 ± 8.09 in the control group and 55.98 ± 10.57 from the diabetic group. Originally, this dataset was created to investigate temperature distribution in the plantar region among diabetic and non-diabetic subjects and measure these differences. However, the INAOE thermogram database has found extensive use in the early-stage detection of DFU. For the acquisition, the authors used two different infrared cameras (FLIR E60 and FLIR E6). As stated by the authors [206], the dataset is slightly unbalanced towards diabetic cases that almost tripled those from the control group.

Despite the limited sample size and evident class imbalance, both ML and DL models have been applied to this dataset. The typical workflow involves an initial segmentation step, followed by feature extraction to generate a feature vector used as input [216]. Remarkably, reported accuracy levels neared 100% when employing more complex models that had been pre-trained with other datasets, the technique commonly known as transfer learning. Moreover, the enhancement of images for diabetic foot detection has been explored using various state-of-the-art CNNs [217]. Notably, an

Table 4.1: DFU dataset Summary

Dataset	Control Images	Diabetic Images	Image size
INAOE	45	122	64 × 64
IACTEC	74	0	64 × 64

F1-Score of 95% was achieved using MobileNetV2. Additionally, feature extraction from temperature maps was conducted for classification tasks using machine learning models. In this case, an F1-Score of 97% was reported as the best result when utilizing AdaBoost with a set of 10 features. Furthermore, three advanced deep learning architectures were investigated for classifying subjects with diabetes [218]. However, it was acknowledged that these complex models require substantial amounts of training data to optimize the thousands of model parameters. Since the INAOE dataset is not sufficiently large for training such models, the authors proposed an augmentation technique based on the Fourier transform, achieving accuracy values above 95%, and even a perfect score of 100% with ResNetV2. However, as we have been described in this chapter, it is complicated to trust in those models and there is no previous work trying to validate these results.

With the purpose of testing our approach in a binary classification task, DFU detection indicating diabetic sample or not, a dataset composed by images acquired by infrared thermography has been generated by the integration and normalization of existing available datasets. These datasets contain thermal feet images from diabetic and non-diabetic subjects, as summarized in Table 4.1. The process of creating the custom dataset and merging it with the INAOE dataset is described in the following subsections.

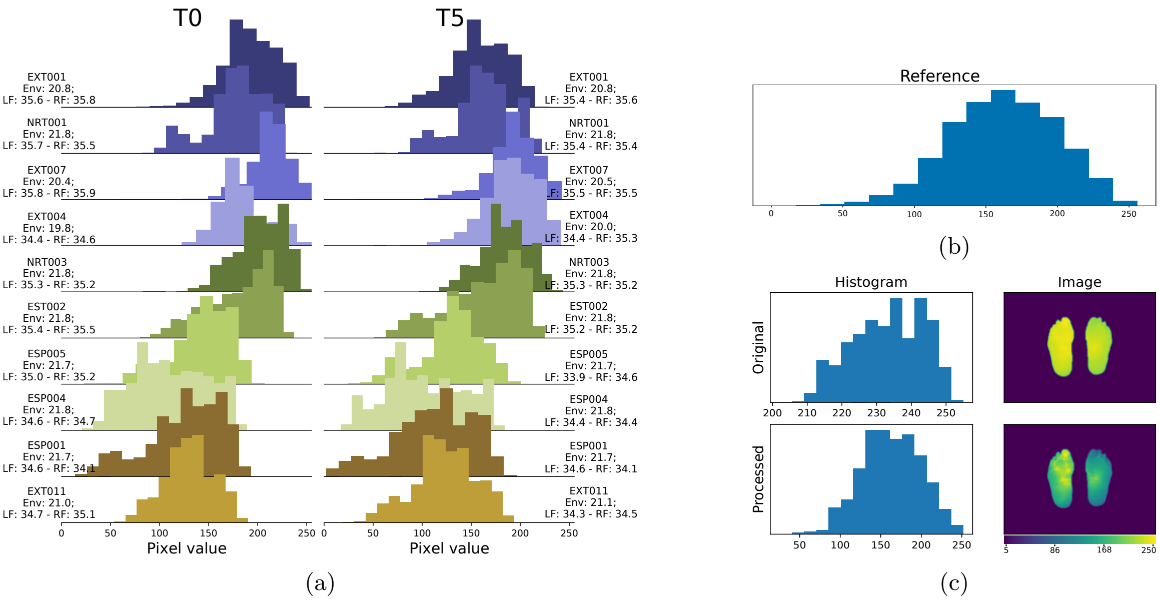
Internal dataset

In order to balance the number of samples per class in the INOAE dataset [206], we integrated a second custom dataset, generated by IACTEC[1], the technology center associated to the Astrophysical Research Institute of the Canary Islands (Instituto de Astrofísica de Canarias, IAC)[2]. This dataset [202] contains 74 infrared thermal images, captured from 37 non-diabetic volunteers, 15 female and 22 male, with a mean age of 40 ± 8 in a range between 24 and 60 years old. This dataset was acquired using the system illustrated in Fig. 4.3a, which has a TE-Q1 Plus thermal camera from Thermal Expert™ (i3system Inc., Daejeon, Republic of Korea). Images were saved using 16-Bit PNG format with a spatial resolution of 384 × 288 pixels. The acquisition campaign was carried out in November 2020, acquiring two sets of images per subject. The first image (T0) was captured immediately after the person becomes barefoot and sits with legs extended forward or lies down in a supine position with the feet off the ground. The second image was taken five minutes later (T5), meanwhile the subject was at the same resting position.

With the purpose of testing our approach in a binary classification task – detecting DFU and indicating whether a sample is diabetic or not – a dataset composed by images acquired by infrared thermography has been generated by the integration and normalization of existing available datasets. These datasets contain thermal feet images from diabetic and non-diabetic subjects, as summarized in Table 4.1.

[1] https://www.iac.es/es/observatorios-de-canarias/iactec

[2] https://www.iac.es/en



(b)

(a)

(c)

Figure 4.4: (a) Histograms of a pair of images (T0 and T5) from the same subject extracted from the IACTEC dataset. The temperature of each foot, left (LF) and right (RF) foot, and environment (Env) is detailed on the sides in degrees Celsius. (b) Reference histogram. (c) Example of the histogram matching processing for a single subject.

Data merging

Since both datasets were acquired under different ambient conditions and using different devices, it was necessary to standardize them in a meaningful way. For this purpose, a histogram matching process was used over all images, so that all match a reference histogram [219]. Histogram matching is a useful technique when the contrast level of a group of images has to be unified. As we aim to analyze spatial features rather than temperature values, histogram matching will not distort the information contained in the images for the purpose of our analysis.

In this way, the IACTEC dataset was established as a reference, since it offers a well-known acquisition protocol [202]. Figure 4.4a illustrates histograms of T0 and T5 from the IACTEC dataset. As observed, the data distributions are similar at T0 and T5. Nevertheless, the images at T5 tend to have a better qualitative representation of the temperature pattern in the feet, being more visible to the naked eye. For this reason, the reference histogram was computed using the T5 samples from various subjects. These samples were selected after a qualitative visual inspection of the complete dataset, selecting 6 initial samples. Then, the histogram distributions were analyzed to obtain the reference, using the skewness ($Skew$) and kurtosis ($Kurt$) statistics. As can be seen in Fig. 4.4a, a high-rate of T5 data distributions are negatively skewed (or left-skewed). Analyzing the initial selected samples, the optimal $Skew$ ranges from -0.05 to -0.4, while $Kurt$ achieved a maximum value of -0.85. Thus, 12 images that fulfilled those requirements were selected as references. The average histogram, $\hat{h}$, from those images was obtained as follows:

$$\hat{h}_i = \frac{1}{N}\sum_{j=1}^{N} h_{ij}$$

where N is the number of samples and h_i represents the value of the i^{th} bin of the original histogram. In this experiment, the number of bins for histogram computation was set to 15. Figure 4.4b illustrates the reference histogram, while Fig. 4.4c shows the distribution of the pixel values and the examples images before and after performing the histogram matching. The processed histogram was quite similar to the reference histogram, offering an improvement in the visual interpretation of the temperature patterns.

As expected, the image contrast increases by applying the histogram matching, as it is illustrated in Fig. 4.4c. Thus, temperature patterns in both datasets were more visible, having the entire dataset similar contrast. Finally, histogram matching was applied to the IACTEC images to obtain the same contrast that in the processed INAOE images, so both datasets were modified. In the IACTEC dataset, the changes are subtle, but a qualitative improvement was observed in the samples. In conclusion, data normalization performed to improve temperature patterns, which is acceptable for our application, looks promising and allows the unification of independent datasets.

4.2.2 Experimental Setup

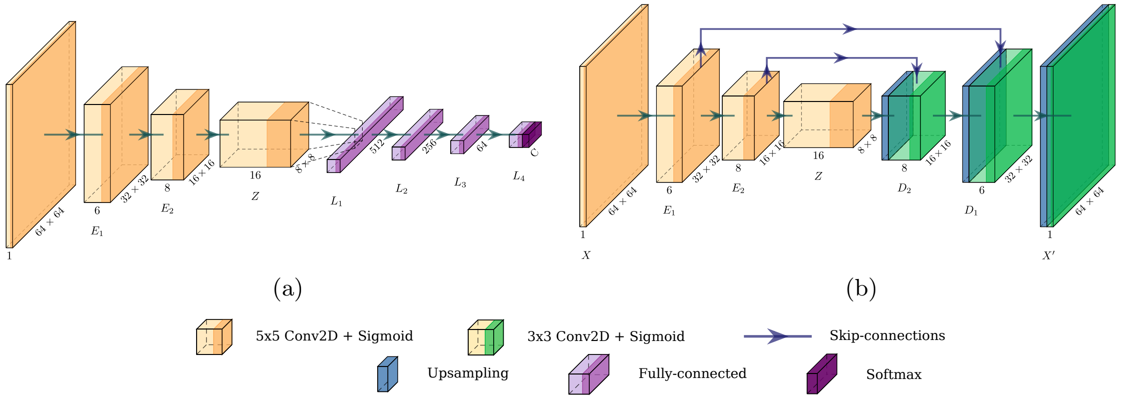
This chapter is dedicated to the analysis of various DNNs from an information-theoretical perspective (as discussed in Chapter 4.1). The goal is to identify promising models that demonstrate robust performance when confronted with a scarce dataset. To draw meaningful conclusions, several experiments will be conducted, encompassing diverse architectural designs and various techniques for addressing overfitting. This section provides a detailed outline of the experimental procedure.

Proposed Network Architecture

Most popular network architectures for image classification tend to use convolutional layers in the first steps of the process (e.g., ResNet [220], VGG [149]) to make a representation and reduce the amount of information, generating a latent space Z with lower dimensionality than the input space X. It may be described as an encoding phase, where the convolutional layers try to extract features of the image, and classification in which the last layers have the purpose of identifying the image. Reducing dimensionality of the input decreases the number of parameters to be used when training the network, simplifying the overall classification process. Following this motivation, an architecture based on a first stage of *encoding*, $T_E \in \{E_1, E_2, Z\}$, followed by a *classifier*, $T_{Cl} \in \{L_1, L_2, L_3, L_4\}$, was designed. Fig. 4.5a illustrates an example of such architecture. Since the final goal is to classify images, the training objective, see Eq. (2.7), is defined by the CE (Eq. 2.5).

While the proposed architecture may not be as deep as those found in the current state-of-the-art for general-purpose image processing, it is important to note that even with this number of layers, the model will contain millions of parameters. It is crucial to keep in mind that this analysis is conducted under conditions where the number of available samples is limited. In total, after the preprocessing step and the merging of both datasets, as described in Section 4.2.1, the dataset comprises only 211 images (see Table 4.1), which is notably low for the utilization of complex models such as DNNs.

As mentioned earlier, when working with small datasets, transfer learning can significantly enhance the model's classification performance. This approach involves using a pre-trained model and fine-tuning it to adapt to the specific classification task with the available samples [221]. In this



(a)

(b)

Figure 4.5: Proposed network architectures for DFU classification. (a) Architecture proposed for classification where the convolutional layers are provided by the pre-trained AE. (b) Convolutional AE proposed where the skip connections are optional depending on the experiment.

study, an AE architecture [67], depicted in Fig. 4.5b, has been used to apply transfer learning from a pre-trained AE to the first layers of our classification architecture Fig. 4.5a.

AutoEncoders for Fine-Tuning

Transfer learning [221, 182] is a technique that involves using a model initially trained for a specific task as a starting point for training on a second task. The initial state of the parameters of the model, θ_0, is generated from a training process with another dataset, which is usually larger and more complete. This approach is commonly employed to address overfitting concerns and ensure that the trained model performs well on new data. As you will observe in the subsequent section, we have incorporated experiments where transfer learning is applied.

To implement transfer learning, it has been utilized an AE [67]. As explained in Section 2.2.3, the primary purpose of an AE is to learn a self-supervised efficient representation (encoding) for a given dataset, typically for dimensionality reduction. In essence, the AE compress the essential information within the input data. In this case, the path layers were used as T_E in the proposed model to classify the DFU dataset. Considering that an AE, from an information-theoretical perspective, performs based on the InfoMax principle, the use of AEs is the best option to obtain a general compression of the input, without discarding irrelevant features based on a task.

In addition, the AEs allows obtaining different configuration of the latent space. In specific, it will be analyzed different configuration based on convolutional or fully-connected AEs. In addition, it will be evaluated the compressed latent space obtained by an AE applying the skip-connection technique. A more detailed focus on the AEs will be included in Section 4.3.

Considering this approach, AEs were trained to reconstruct an input X from a compressed representation Z. Thus, applying image reconstruction is straightforward, as labeled samples are not necessary. Given an output $\hat{X} = \hat{f}(X; \theta)$, representing the reconstructed image, the model is evaluated by comparing $\hat{X}$ with the original one X, using the MSE (Eq. 2.18).

This training process has been carried out in two steps: a first training using a dataset with numerous samples (i.e., the reference dataset) and a second fitting step where the AE was trained using the DFU dataset. Transfer learning should be applied in the same domain of the target dataset [222]. However, since the INAOE dataset [206] is the only public thermogram dataset for

Table 4.2: Summary of the proposed experiment configurations.

Experiment	Pretrained Encoder	Bottleneck Size
PCAE	Yes	$16 \times 8 \times 8$
PCAES	Yes	$16 \times 8 \times 8$
PFCAE	Yes	256
NPCE	No	$16 \times 8 \times 8$
NPNE	No	-

DFU, currently, it is not possible to obtain another dataset in the same domain. Therefore, Fashion-MNIST (FMNIST) [223] was selected as the reference dataset for pretraining. It consists of a large dataset of grayscale images with a black background and a normalized histogram, being a dataset with similar features to our preprocessed dataset. Remember that the first layers of the proposed architecture, Section 4.2.2, has the purpose of extracting useful features from the input. Even when the topic of the dataset (i.e., fashion-related elements) is not associated to our data, the amount, and quality of its samples has demonstrated in our experiments to be robust for obtaining coherent feature extraction filters in Z.

4.2.3 Experiment Descriptions

In this section, we provide a comprehensive description of each experiment, outlining the model architecture and crucial hyperparameters. We conclude with precise instructions for replicating the training process to enhance the transparency and reproducibility of our research.

PCAE (Pretrained Convolutional AE without skip-connections)

A Pretrained Convolutional AE ($PCAE$) without skip-connections was used to generate the encoder for DFU classification. The architecture of this convolutional AE is illustrated in Fig. 4.5b, excluding the skip-connections. The encoder contains the convolutional layers, with convolutional kernel size (K) of 5×5 with stride and padding of 2. The decoder contains upsample layers, using 2D nearest neighbor interpolation, followed by convolutional layers with K of 3×3, with stride and padding of 1. The number of filters in each layer of the encoder path corresponds to 6, 8, and 16 respectively.

PCAES (Pretrained Convolutional AE with Skip-connections)

A Pretrained Convolutional AE with Skip-connections ($PCAES$), similar to the previous case but including the skip-conecctions, was employed. This experiment uses the architecture illustrated in Fig. 4.5a and 4.5b, applying skip-connections. The configuration of the hyperparameters is the same as in PCAE.

PFCAE (Pretrained Fully-connected AE)

A Pretrained Fully-connected AE ($PFCAE$) was defined for the DFU classification encoder. The encoder of this AE is conformed by four fully-connected layers $T_E \in \{64 \times 64, 1024, 512, 256\}$. Thus, the first layer corresponds to the input layer, defined by the image size of the dataset (see Table 4.1) and, in this case, $Z \in \mathbb{R}^{256}$.

NPCE (Non-Pretrained Convolutional Encoder)

This classification model architecture is similar to those obtained in PCAE and PCAES experiments. However, the classification model uses a Non-Pretrained Convolutional Encoder (*NPCE*), meaning there is no transfer learning process.

NPNE (Non-Pretrained No Encoder)

This is the baseline approach in which there is no latent space generated from an encoder, and the original image is the input of the classifier. This experiment will be referred to as *NPNE*.

Dataset Split

The dataset, consisting of 241 images, is divided into two subsets: a test set containing 32 samples and a training set containing the remaining samples. The test set was carefully selected following an analysis to ensure that the samples are as representative as possible of the original dataset, and it is balanced, comprising 16 samples from the control group and 16 samples from the diabetic subjects. For a more comprehensive overview of the dataset and its characteristics, please see Section 4.2.1.

Model Training

For consistency and clarity, the model training details for all experiments are summarized below:

- **Batch size**: A batch size of 128 samples was used for training.
- **Optimizer**: The ADAM optimizer [65] was employed for DNN training
  - **Initial Learning Rate**: The initial learning rate (lr) was set to $5 \cdot 10^{-4}$.
  - **Exponential Decay Rates**: Parameters β_1 and β_2 for moment estimation were set to 0.9 and 0.999, respectively.
- **Learning Rate Decay**: The learning rate decays by κ every iteration t, with κ set to 0.999.

$$lr_t = \kappa * lr_{t-1},$$

4.2.4 Experimental Results

Initially, to assess the classification models, it is employed the classical classification metrics, including sensitivity, specificity, precision, and accuracy. It is important to note that, given the limited size of our test dataset, the reliability of these metrics may be subject to scrutiny. In Table 4.3, it is presented a concise summary of the classification outcomes for the diverse experimental configurations. As indicated by these conventional metrics, all models appear to be promising at first glance in addressing the designated task.

However, the effectiveness of these models has been examined from an information-theoretical perspective, based on the IB principle and the constraints of the architecture (see Section 4.1). This rigorous assessment helps us identify the models with the highest classification abilities.

To address the limitations of our small test dataset, which comprises only 32 samples (as described in Section 4.2.3), the IT metrics were estimated using the kernel-based estimator, as discussed in

Table 4.3: Classification metrics results for each of the experiments.

Experiment	Sensitivity	Specificity	Precision	Accuracy
PCAE	0.937	0.937	0.937	**0.937**
PCAES	**1.000**	0.687	0.762	0.844
PFCAE	0.812	**1.000**	**1.000**	0.906
NPCE	0.937	0.87	0.882	0.906
NPNE	0.875	0.937	0.933	0.906

Section 3.6. This approach, focusing on the analysis of the IP, allows us to obtain a more precise and credible assessment of the models' performance, while it is validated the use of kernel-based entropy estimator for scarce dataset.

DFU Classification Analysis

The analysis of the classification task aimed to determine whether the models generated in the proposed configurations held the same promise as indicated by the classification metrics presented in Table 4.3. In experiments where a pretrained AE, summarized in Table 4.2, was employed, the encoder in the classification model was initialized using the encoder path and bottleneck. Consequently, layers $T_E \in \{E_1, E_2, Z\}$ were initialized based on the AE configuration, as illustrated in Fig. 4.5b.

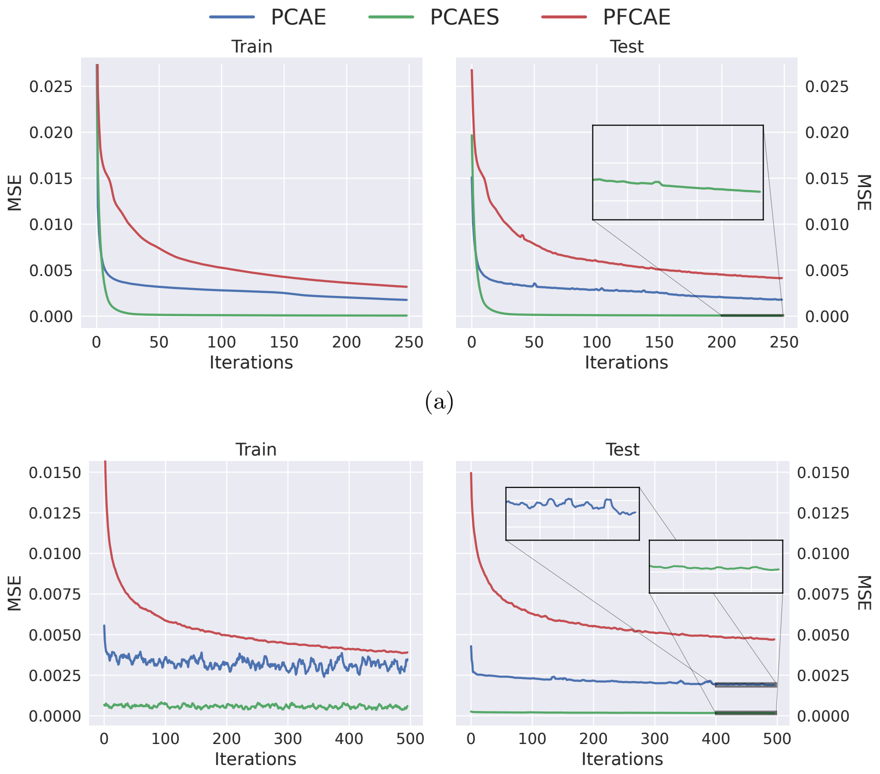
In experiments utilizing a pretrained encoder (i.e., PCAE, PCAES, and PFCAE), the AEs demonstrate robust performance, as evident from the evolution of the loss values shown in Fig. 4.6b. The MSE loss, as defined by Eq. (2.18), representing the error between the desired output X and the reconstruction $\hat{X}$, consistently decreases in all configurations. This reduction in loss occurs during both training and testing, indicating the absence of overfitting in the AEs. The qualitative results of $\hat{X}$ shown in Fig. 4.7 also confirm the satisfactory final output for various configurations, albeit with slightly inferior results in the AE based on fully-connected layers (PFCAE), as anticipated. Based on these findings, models employing pretrained encoders are expected to exhibit good performance.

Subsequently, the classification model was evaluated as previously mentioned, and this analysis was supported by IP estimation. It is important to note that our test set was limited in size and the IP has been estimated by the kernel-based entropy estimator.

Regarding the kernel-based entropy estimation, the RBF kernel was applied. In these experiments, the kernel width (σ) selection was carried out by maximizing the kernel-alignment loss (as detailed in Section 3.6.3). The kernel-alignment loss, Eq. (3.32) was performed on each epoch t using 200 σ values from 0.1 to 10 times the mean distance between the samples in one mini-batch, as Wickstrom et al. proposed [169]. To stabilize the σ values across mini batches, an exponential moving average has been used.

Figure 4.8 and Fig. 4.9 provide insight into the model's performance during the training process. Figure 4.8 illustrates the evolution of the CE loss value across different iterations (t). Most models exhibit a convergence towards an optimal solution, except for NPNE.

In the NPNE experiment, which is the simplest scenario where the encoder is eliminated, the training and testing loss values exhibit a distinctive pattern of overfitting. This behavior is evident from the training loss value decreasing while the testing loss value increases over iterations, as



(a)

(b)

Figure 4.6: AutoEncoder training for Transfer Learning: Loss values (MSE) during training (left) and testing (right) on (a) the FMNIST dataset and (b) the DFU dataset.

depicted in Fig. 4.8e.

The IP trajectories shown in Fig. 4.9 were used just in the classifier layers $T_{Cl} \in \{L_1, L_2, L_3, L_4\}$, discarding T_E (see Fig. 4.5a). In the experiments where a pretrained AE was used, the input X for computing MI in IP trajectories is Z, i.e., the output of the encode path as it is illustrated in Fig. 4.5a. As a reference, the entropy of these compressed representations obtained by the different AEs is around 3.2 bits for PCAE, 1.2 bits for PCAES, and 3.5 bits for PFCAE. More details about the AEs applied will be found in the next section 4.3. In the NPCE and NPNE experiments, the input is the original X, with an entropy of around 5 bits.

Note that the output Y is the test set which, as mentioned before, is balanced for both classes, $N_{C_1} = N_{C_2}$. Thus, the theoretical maximum value in $\mathcal{I}(T;Y)$ is given by $log_2(C) = log_2(2) = 1$. In addition, the IP trajectories contain the information on the figure at the left, since a minimization of the CE is translated as an increase in the $\mathcal{I}(L_4;Y)$.

When analyzing the non-pretrained experiments (NPCE and NPNE), several noteworthy observations emerge from the loss progression in Figures 4.8d and 4.8e, as well as from the IP estimation depicted in Figures 4.9d and 4.9e.

For NPCE, it is evident that there are numerous iterations where the training CE loss struggles to converge, although it eventually stabilizes (see Fig. 4.8d). The substantial divergence between training and testing losses, which gradually increases over time, is indicative of overfitting and

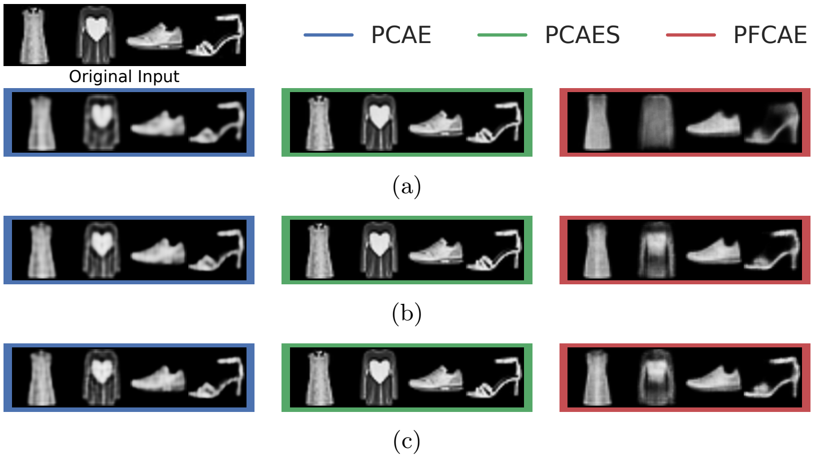


Figure 4.7: Qualitative results in different iterations during FMNIST training: (a) 24 iterations, (b) 170 iterations and (c) 250 iterations.

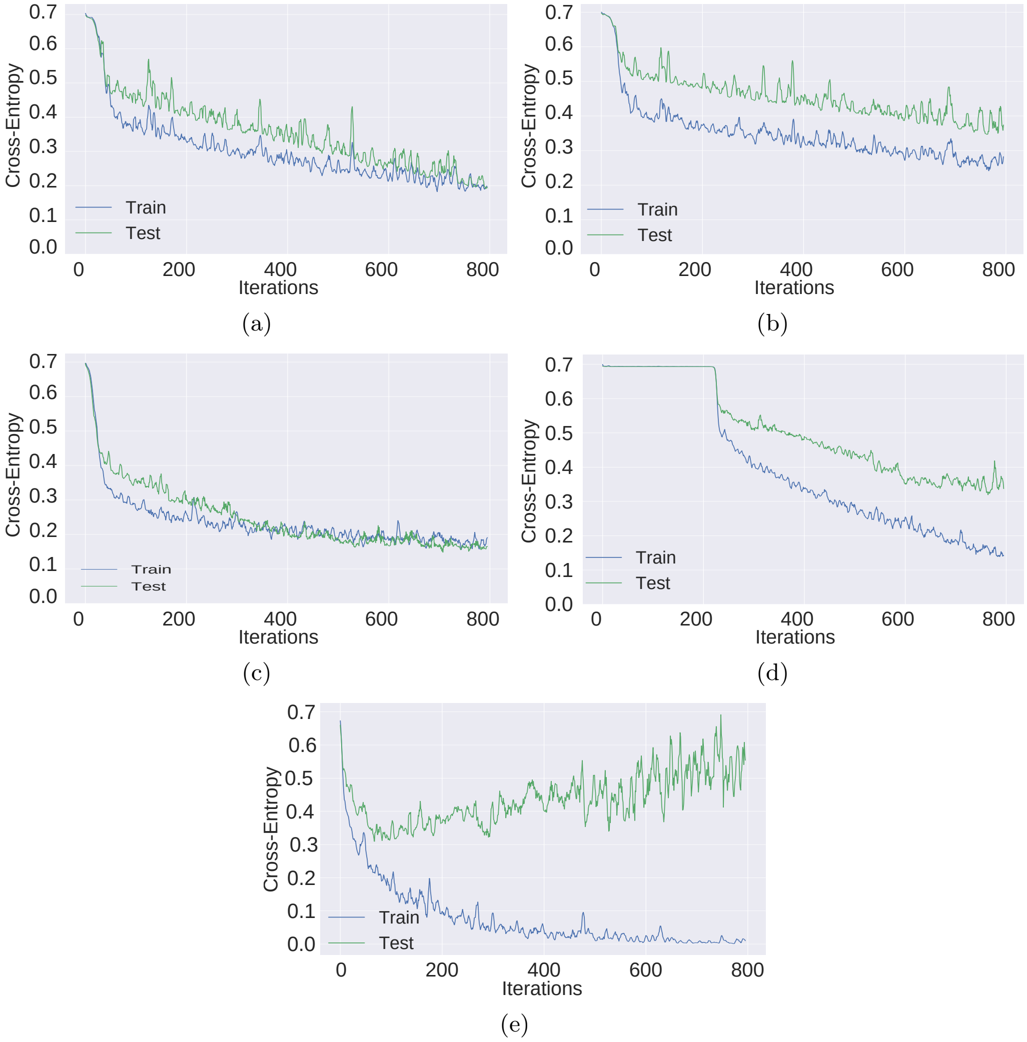
possibly underfitting. This imbalance between training and testing losses results in limited progress and decreased generalization on labeled data. While the issue of overfitting is not as pronounced when considering the training loss alone (as observed in Fig. 4.8d), it becomes more apparent in the MI estimation, particularly after the compression phase, see Section 4.1.2. Notably, the MI estimation in Fig. 4.9d reveals a violation of the DPI principle, see Eq. (4.1), where $I(X, L_2) < I(X, L_3)$. This DPI violation is likely linked to overfitting during the compression phase, as suggested by Wickstrom et al. [169]. Furthermore, the MI estimation $\mathcal{I}(T; Y)$ falls far short of its theoretical maximum value, indicating suboptimal model performance.

As for NPNE, a clear case of overfitting is observed, where the test CE loss value increases while the training CE loss value consistently decreases (Fig. 4.8e). The MI trajectories suggest that the estimation experiences an fitting phase. However, over time, it has become stuck within a narrow range, fluctuating without showing a discernible increasing or decreasing pattern. This behavior suggests a potential issue of underfitting, implying that the number of parameters is insufficient to adequately characterize the data and capture the intricacies of the input-target relationships. Additionally, there is a DPI violation, see Eq. 4.2, where $I(L_3; Y) > I(L_4; Y)$.

In conclusion, it is evident that the non-pretrained models (NPCE and NPNE) do not exhibit the same level of promise as suggested by the performance metrics presented in Table 4.3. In both cases, there are observed violations of DPI, and the values of $\mathcal{I}(L_4; Y)$ are significantly distant from their maximum theoretical counterparts.

Among the pretrained approaches using transfer learning, it is evident that the models achieve a decreasing test CE loss value, as illustrated in Fig. 4.8, and consistently achieve better progression compared to the non-pretrained approaches. Additionally, examining the IP trajectories of PCAE, PCAES, and PFCAE (Figures 4.9a, 4.9b, and 4.9c), it is observed a phase of constant fitting in most layers, with the exception of L_1 in PCAE and PCAES. In these cases, it is noticed an initial decreasing trend in $I(X; T)$ during the early iterations, followed by a fitting phase.

Importantly, there are no observed violations of DPI in the different pretrained approaches. However, in PCAE and PCAES, the IP trajectories of L_2 and L_3 are closely aligned, overlapping



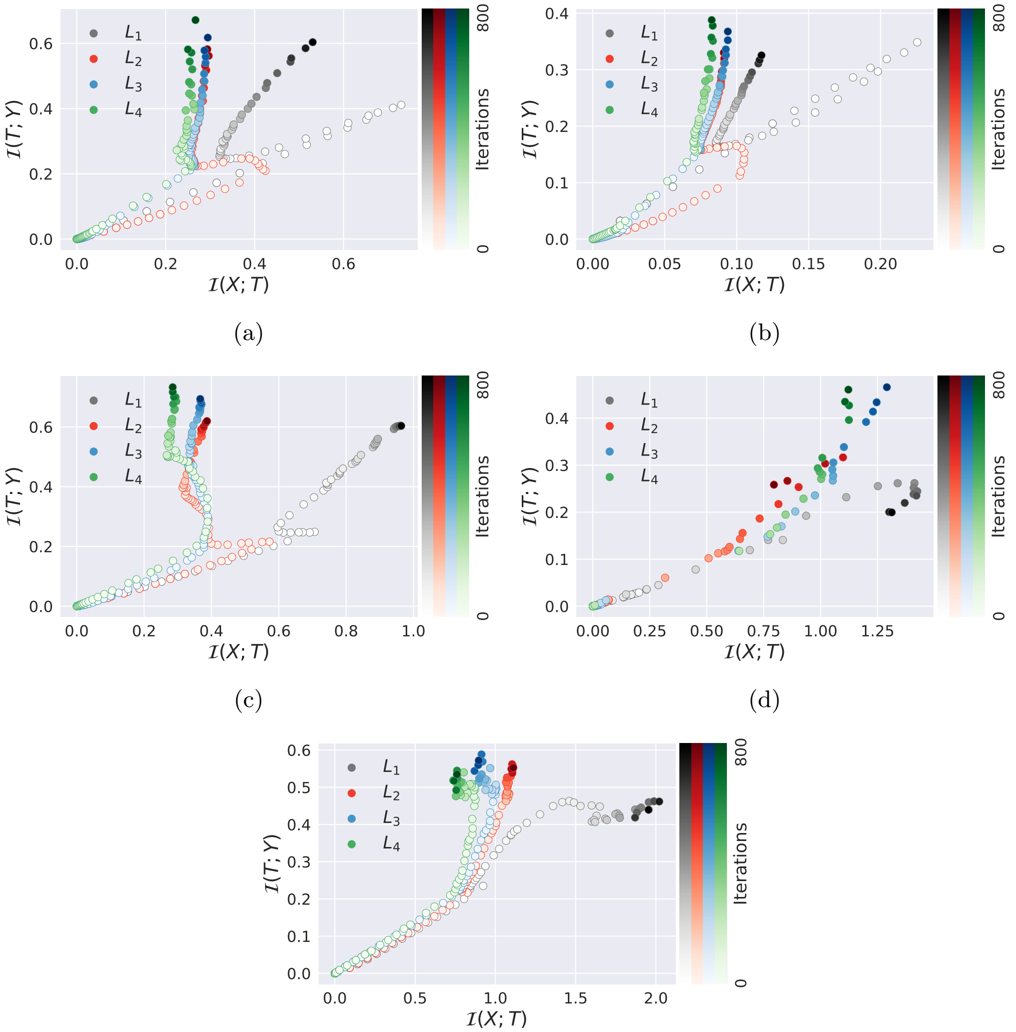
(a) (b) (c) (d) (e)

Figure 4.8: Loss error in the different experiments: (a) PCAE, (b) PCAES, (c) PFCAE, (d) NPCE and (e) NPNE.

each other. This overlap may suggest that these layers are similar, potentially indicating that the model could be further reduced in complexity, i.e., it is possible to reduce the number of parameters.

It is worth noting that PCAES encounters a similar problem to the non-pretrained approach, where $\mathcal{I}(T;Y)$ is far from its theoretical maximum value, as depicted in Fig. 4.9b. Additionally, the MI with respect to the input in PCAES, which corresponds to the output of a pretrained encoder, suggests that PCAES may be overfitting. Conversely, PCAE and PFCAE demonstrate values closest to the maximum theoretical value of $\mathcal{I}(T;Y)$.

In conclusion, transfer learning turns out to be effective even when the reference dataset does not belong to the same domain as the target dataset, as is the case with the DFU dataset. Furthermore, the IP trajectories identified theoretical characteristics, which were previously discussed in the state-of-the-art, in the diverse experiments in a manner that validated the analysis for small datasets using the aforementioned kernel-based entropy estimator (see Section 3.6).



(a)

(b)

(c)

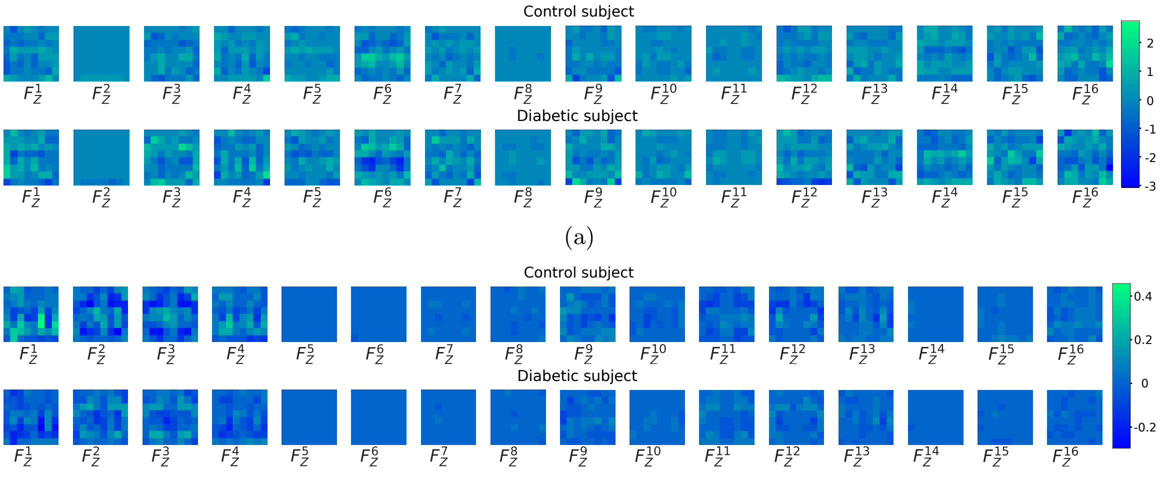
(d)

(e)

Figure 4.9: Information Plane estimation in the different experiments: (a) PCAE, (b) PCAES, (c) PFCAE, (d) NPCE and (e) NPNE.

4.3 AutoEncoder Insights

In the previous section, an analysis of the classification model performance was conducted from an information-theoretical perspective. Overall, the characteristics observed in the previous section align with those discussed in the state-of-the-art, validating this analysis using a scarce dataset through the kernel-based entropy estimator. This approach has proven valuable for identifying biases caused by model overfitting, particularly when the dataset size is limited and traditional classification metrics may not provide reliable insights. While some conclusions were drawn regarding the use of AEs, this section delves deeper into understanding the specific reasons behind the varying performance of different models.



(a)

(b)

Figure 4.10: Saliency maps with features contribution in F_Z^i filters for DFU classification by DeepLIFT. The figure (a) corresponds to PCAE and (b) belongs to PCAES.

4.3.1 Skip-Connection Effects in Compressed Representation

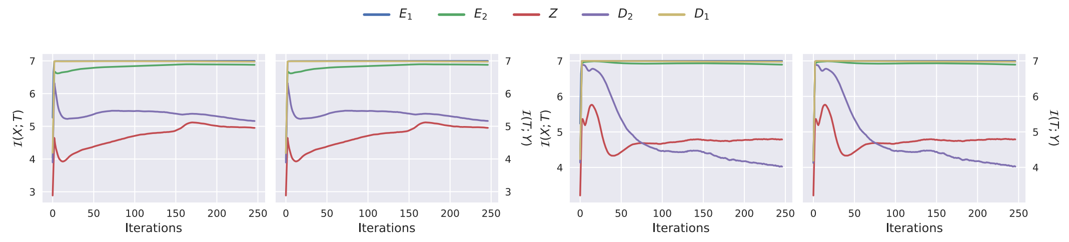
Based on the results from the Section 4.2.4, it is evident that PCAES exhibited the poorest performance among the models where transfer learning was applied. In this section, we will conduct a detailed comparison between PCAE and PCAES to uncover the reasons behind the negative impact of skip-connections on the resulting compressed latent space Z. Both autoencoders were trained under identical conditions, ensuring a fair comparison between the two experiments, with skip-connections being the only distinguishing factor.

Leveraging the models' ability to accurately classify samples, as previously discussed in the preceding section, we employed DeepLIFT [100] to elucidate which features from Z are influencing the classifier's decisions. Essentially, DeepLIFT generates a saliency map in Z, computing an importance score based on a reference image, enhancing the explainability of the DL model (see Section 2.3.3). The choice of the reference image is a crucial factor that significantly impacts the results, as it determines what aspects of the input are considered relevant. For the DFU dataset, it was employed an all-zeros input as the reference, essentially representing the background of the image, encompassing everything except the feet, as illustrated in Fig. 4.4c.

Figure 4.10a and Fig. 4.10b represent the resulting saliency maps for PCAE and PCAES configurations, respectively. A binning-clustering process was applied to help interpret Fig. 4.10, dividing the saliency scores into 15 equidistant bins. This binning approach is particularly helpful in the case of PCAES, where the scale of values is relatively narrow.

These findings reveal that the feature contributions in both the control and diabetic groups exhibit strikingly similar patterns. The saliency maps display comparable patterns and conjugated values, with what appears as “hot” values (green color) in the control group often corresponding to “cold” values (blue color) in the diabetic subjects.

In the case of PCAE (Fig. 4.10a), most filters exhibit spatially irregular patterns in their feature contribution estimations. However, filters F_Z^2 and F_Z^8 demonstrate limited activation, indicating low feature contribution.



(a)

(b)

Figure 4.11: IP estimation in the FMNIST pretraining step: (a) PCAE, (b) PCAES.

On the other hand, PCAES (Fig. 4.10b) reveals only a few relevant filters, specifically F_Z^{1-4}. However, the scale of the feature contributions, as indicated by the colorbar in Fig. 4.10a, differs significantly between PCAE and PCAES, with PCAES showing a reduction by a factor of 10.

Additionally, the IP estimation based on the kernel-based entropy estimator was computed at two different stages of the experiment. The first estimation was conducted during the pretraining process using the FMNIST dataset (Fig. 4.11), and the second estimation was performed during the fitting process using the DFU dataset (Fig. 4.12). While both figures convey the same information, the representation has been adjusted to highlight specific details in the IP estimation at different experiment stages.

As expected, the IP trajectories in the pretraining stage (Fig. 4.11) in the different AEs, shows that $I(X,T) \approx I(T,\hat{X})$. This is because the desired output is roughly similar to the input. Those symmetrical trajectories are highly presented because of the high-quality reconstruction in few iterations. To verify that these estimations are correct, the DPI principle should be satisfied on the encoder, Eq. (4.3), and decoder layers, Eq. (4.4) [190].

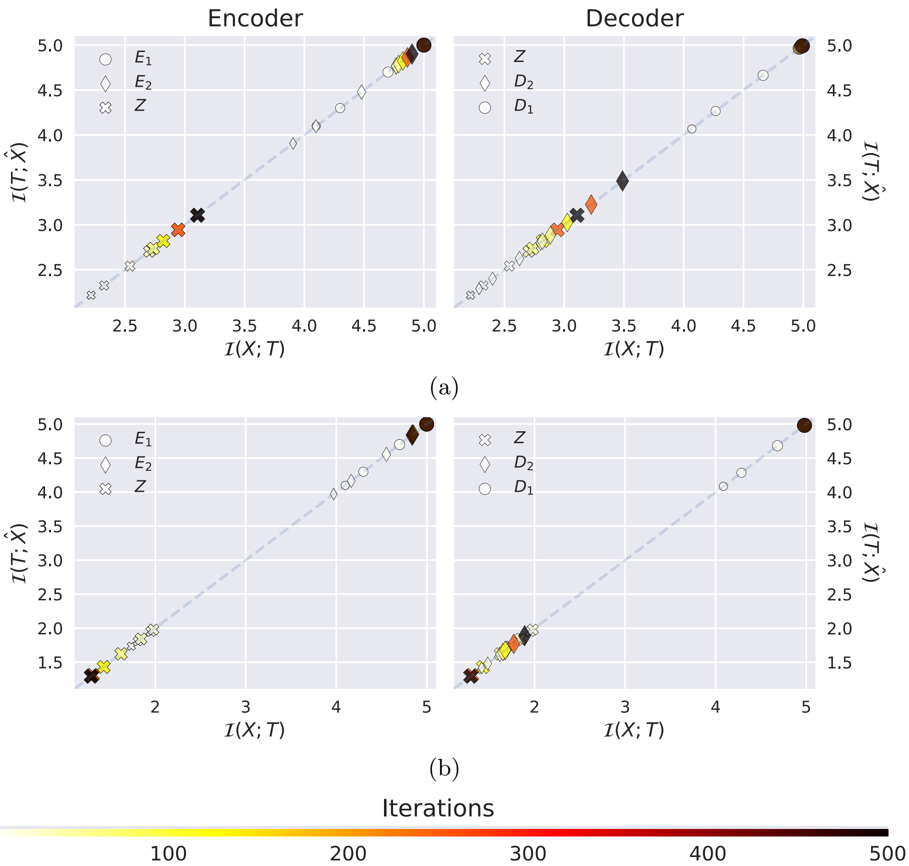
Observing Fig. 4.11b, a violation of the DPI principle occurred in PCAES, as $I(Z,Y) > I(D_2,Y)$. However, the DPI principle assumes that a DNN can be interpreted as a Markov chain, but skip-connections link the encode and decode path (see Fig. 4.5b), as

$$D_i = f_{D_i}([D_{i-1}, E_i]; \theta_{D_i}),$$

which is a violation of the Markov property. As a result, it is not possible to guarantee that Z is the best compressed representation of the input X in PCAES.

Regarding the DFU fitting stage, where the AE is trained to fit the pretrained model to the DFU dataset, it is observed symmetry in the IP estimation, as shown in Fig. 4.12. This symmetry represents a good behavior of an autoencoder, following the diagonal line in Fig. 4.12. However, when we focus on PCAE (Fig. 4.12a), it is evident that the model is consistently in a fitting phase, with different layers progressively increasing $\mathcal{I}(X;T)$. In contrast, PCAES exhibits a different behavior, constantly losing information in the latent space Z as depicted in Fig. 4.12b. This behavior results in the resolution of the DPI violation observed in the initial state provided by the FMNIST pretraining (Fig. 4.11b). It can be concluded that the latent space Z of PCAES is gradually becoming less relevant for reconstruction, even though the reconstruction remains optimal, as evidenced by the MI estimations aligning with the diagonal.

The IP estimation of the PCAES experiment reveals that the incorporation of skip-connections impacts the compression of the input X. Skip-connections attempt to merge information across dif-



(a)

(b)

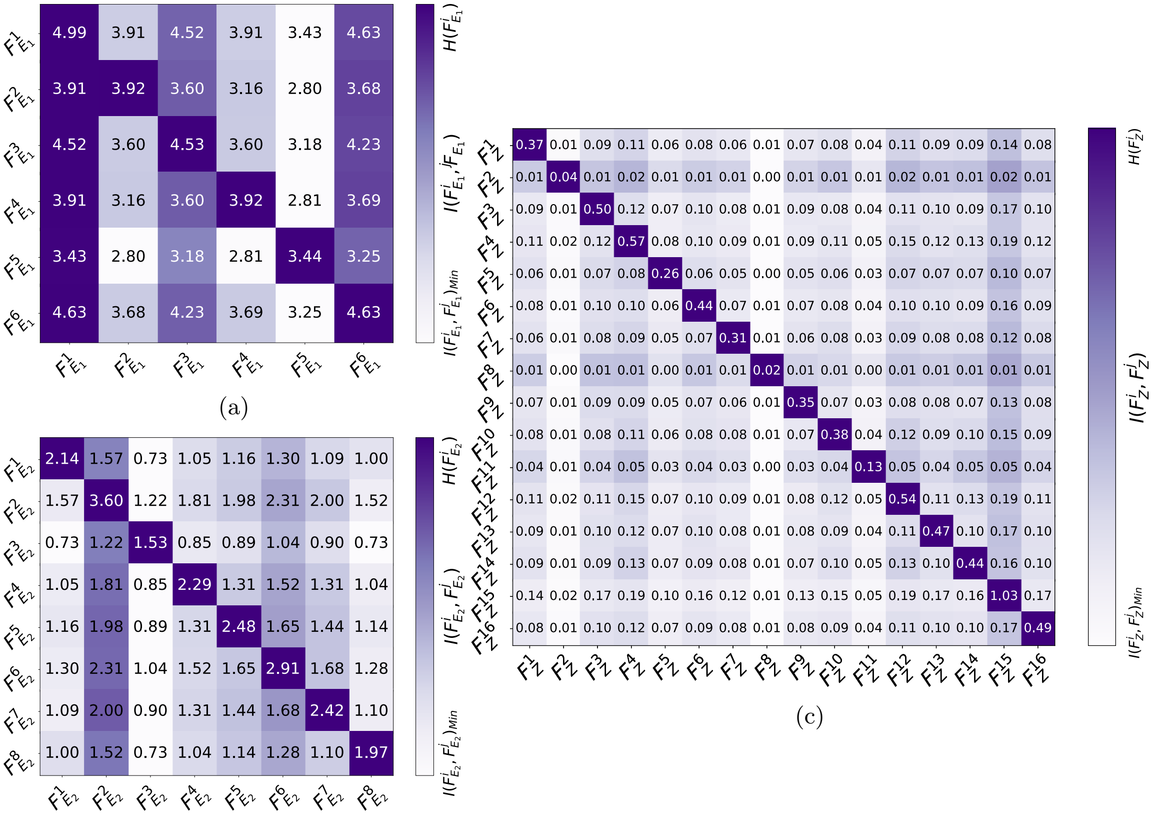
Figure 4.12: IP estimation in the DFU training: (a) PCAE, (b) PCAES.

ferent spaces by combining features from a higher-frequency space with those from a lower-frequency space. It means that the features obtained from an image with more details, the high-frequency space, are combined with the features of an image with fewer details, having smooth changes in intensity, such as areas of uniform color or smooth gradients. However, as inferred from the IP estimation of PCAES, it is not guaranteed that the most compressed space will perform well, as the lower-frequency space may prove ineffective for the reconstruction task, as observed in this experiment, being more important the information provided by the "detailed" features.

Convolutional Filters

In this section, the kernel-based entropy estimator will be used to have a better insight about what is happening in the convolutional filter in PCAE and PCAES. To assess the performance of the different filters in the convolutional layers, a visualization tool is designed to estimate the similarity between filters in a given convolutional layer. The purpose of the proposed visualization tool is to identify the loss of information and additional higher redundancy among filters, providing an idea about whether the different filters are obtaining different patterns. A higher redundancy is represented by the higher MI estimation. In essence, each filter should extract specific information that complements the other filters, so a higher degree of redundancy among filters may serve as an indicator of worse model performance.

In the case of PCAE (Fig. 4.13), a discernible pattern emerges as we delve deeper into the layers



	$F^1_{E_1}$	$F^2_{E_1}$	$F^3_{E_1}$	$F^4_{E_1}$	$F^5_{E_1}$	$F^6_{E_1}$
$F^1_{E_1}$	4.99	3.91	4.52	3.91	3.43	4.63
$F^2_{E_1}$	3.91	3.92	3.60	3.16	2.80	3.68
$F^3_{E_1}$	4.52	3.60	4.53	3.60	3.18	4.23
$F^4_{E_1}$	3.91	3.16	3.60	3.92	2.81	3.69
$F^5_{E_1}$	3.43	2.80	3.18	2.81	3.44	3.25
$F^6_{E_1}$	4.63	3.68	4.23	3.69	3.25	4.63

(a)

	$F^1_{E_2}$	$F^2_{E_2}$	$F^3_{E_2}$	$F^4_{E_2}$	$F^5_{E_2}$	$F^6_{E_2}$	$F^7_{E_2}$	$F^8_{E_2}$
$F^1_{E_2}$	2.14	1.57	0.73	1.05	1.16	1.30	1.09	1.00
$F^2_{E_2}$	1.57	3.60	1.22	1.81	1.98	2.31	2.00	1.52
$F^3_{E_2}$	0.73	1.22	1.53	0.85	0.89	1.04	0.90	0.73
$F^4_{E_2}$	1.05	1.81	0.85	2.29	1.31	1.52	1.31	1.04
$F^5_{E_2}$	1.16	1.98	0.89	1.31	2.48	1.65	1.44	1.14
$F^6_{E_2}$	1.30	2.31	1.04	1.52	1.65	2.91	1.68	1.28
$F^7_{E_2}$	1.09	2.00	0.90	1.31	1.44	1.68	2.42	1.10
$F^8_{E_2}$	1.00	1.52	0.73	1.04	1.14	1.28	1.10	1.97

(b)

	F^1_Z	F^2_Z	F^3_Z	F^4_Z	F^5_Z	F^6_Z	F^7_Z	F^8_Z	F^9_Z	F^{10}_Z	F^{11}_Z	F^{12}_Z	F^{13}_Z	F^{14}_Z	F^{15}_Z	F^{16}_Z
F^1_Z	0.37	0.01	0.09	0.11	0.06	0.08	0.06	0.01	0.07	0.08	0.04	0.11	0.09	0.09	0.14	0.08
F^2_Z	0.01	0.04	0.01	0.02	0.01	0.01	0.01	0.00	0.01	0.01	0.01	0.02	0.01	0.01	0.02	0.01
F^3_Z	0.09	0.01	0.50	0.12	0.07	0.10	0.08	0.01	0.09	0.08	0.04	0.11	0.10	0.09	0.17	0.10
F^4_Z	0.11	0.02	0.12	0.57	0.08	0.10	0.09	0.01	0.09	0.11	0.05	0.15	0.12	0.13	0.19	0.12
F^5_Z	0.06	0.01	0.07	0.08	0.26	0.06	0.05	0.00	0.05	0.06	0.03	0.07	0.07	0.07	0.10	0.07
F^6_Z	0.08	0.01	0.10	0.10	0.06	0.44	0.07	0.01	0.07	0.08	0.04	0.10	0.10	0.09	0.16	0.09
F^7_Z	0.06	0.01	0.08	0.09	0.05	0.07	0.31	0.01	0.06	0.08	0.03	0.09	0.08	0.08	0.12	0.08
F^8_Z	0.01	0.00	0.01	0.01	0.00	0.01	0.01	0.02	0.01	0.01	0.00	0.01	0.01	0.01	0.01	0.01
F^9_Z	0.07	0.01	0.09	0.09	0.05	0.07	0.06	0.01	0.35	0.07	0.03	0.08	0.08	0.07	0.13	0.08
F^{10}_Z	0.08	0.01	0.08	0.11	0.06	0.08	0.08	0.01	0.07	0.38	0.04	0.12	0.09	0.10	0.15	0.09
F^{11}_Z	0.04	0.01	0.04	0.05	0.03	0.04	0.03	0.00	0.03	0.04	0.13	0.05	0.04	0.05	0.05	0.04
F^{12}_Z	0.11	0.02	0.11	0.15	0.07	0.10	0.09	0.01	0.08	0.12	0.05	0.54	0.11	0.13	0.19	0.11
F^{13}_Z	0.09	0.01	0.10	0.12	0.07	0.10	0.08	0.01	0.08	0.09	0.04	0.11	0.47	0.10	0.17	0.10
F^{14}_Z	0.09	0.01	0.09	0.13	0.07	0.09	0.08	0.01	0.07	0.10	0.05	0.13	0.10	0.44	0.16	0.10
F^{15}_Z	0.14	0.02	0.17	0.19	0.10	0.16	0.12	0.01	0.13	0.15	0.05	0.19	0.17	0.16	1.03	0.17
F^{16}_Z	0.08	0.01	0.10	0.12	0.07	0.09	0.08	0.01	0.08	0.09	0.04	0.11	0.10	0.10	0.17	0.49

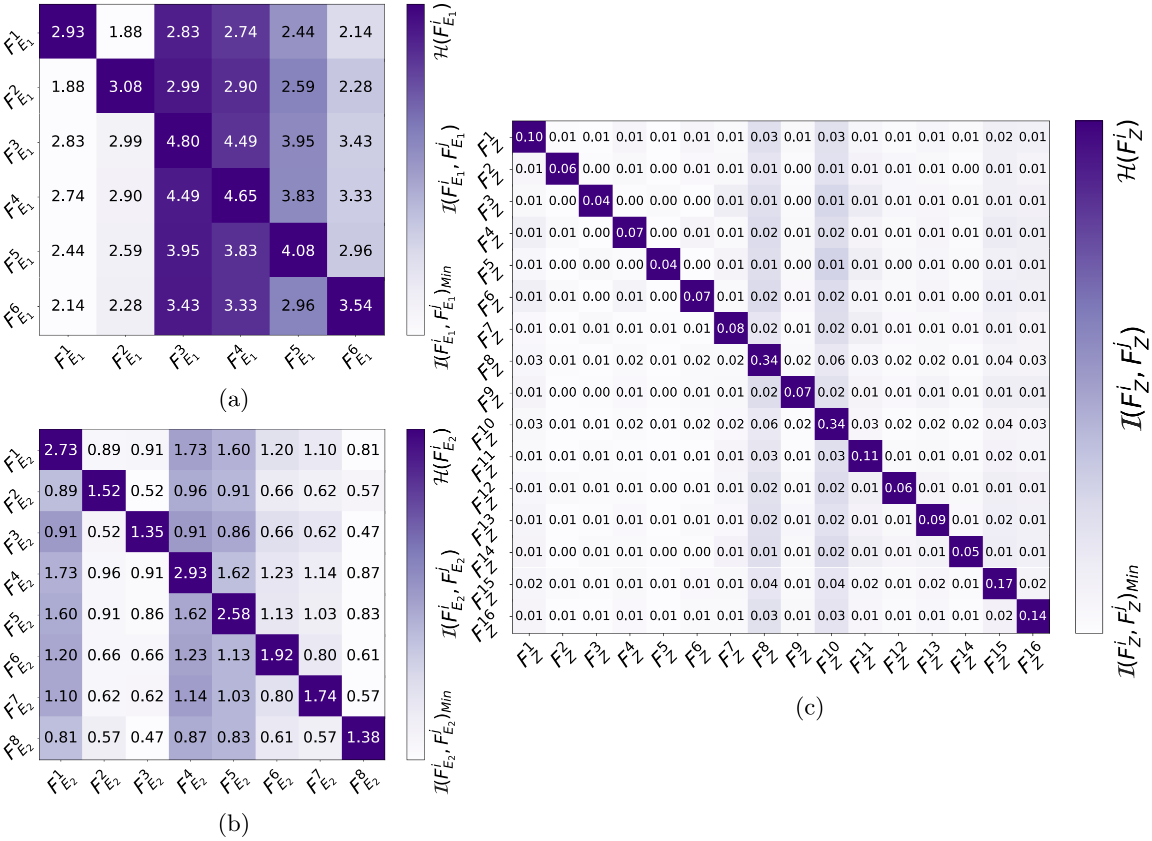
(c)

Figure 4.13: Mutual information estimation between different filters in the encode path of PCAE, F^i_k where i represents the index of the filter and k the encoder layer. The encode path is composed by (a) E_1, (b) E_2 and (c) Z. The trace of the heat-map represents $\mathcal{I}(F^i_k; F^i_k) = \mathcal{H}(F^i_k)$. The colorbar interpretation is applied individually to each row.

of the network. This pattern becomes apparent upon examination of the diagonal elements of the subfigures of Fig. 4.13, where it becomes evident that the entropy of the filters gradually decreases. This pattern aligns with what was observed in the IP estimation (Fig. 4.12).

When analyzing the latent space generated by the first convolutional layer, denoted as E_1, Fig. 4.13a illustrates that certain filters retain entropy levels close to $\mathcal{H}(X) = 5$ bits. As expected from the IP estimation (Fig. 4.12), E_1 serves as a compression of the input X with minimal information loss. In this specific case, the $F^1_{E_1}$ (Fig. 4.13a) shows that contains most information from X, as indicated by its entropy of approximately 5 bits. However, it is worth noting that $F^1_{E_1}$ appears to contain redundant information, as evidenced by the intense color in the first column of the filter's mutual information matrix, suggesting that its information is also present in other filters. In contrast, $F^5_{E_1}$ appears to be the filter with the least redundant information, offering relevant data that complements other filters.

Moving on to the latent space E_2 (Fig. 4.13b), it is observable that the redundancy of information has been decreased, so it means that each filter extracts information which can be complementary for the other filters. In this case, the filter $F^2_{E_2}$, which is the filter with the highest entropy, looks like the most redundant filter, but the MI estimation of the different filters shows that the information



	$F^1_{E_1}$	$F^2_{E_1}$	$F^3_{E_1}$	$F^4_{E_1}$	$F^5_{E_1}$	$F^6_{E_1}$
$F^1_{E_1}$	2.93	1.88	2.83	2.74	2.44	2.14
$F^2_{E_1}$	1.88	3.08	2.99	2.90	2.59	2.28
$F^3_{E_1}$	2.83	2.99	4.80	4.49	3.95	3.43
$F^4_{E_1}$	2.74	2.90	4.49	4.65	3.83	3.33
$F^5_{E_1}$	2.44	2.59	3.95	3.83	4.08	2.96
$F^6_{E_1}$	2.14	2.28	3.43	3.33	2.96	3.54

(a)

	$F^1_{E_2}$	$F^2_{E_2}$	$F^3_{E_2}$	$F^4_{E_2}$	$F^5_{E_2}$	$F^6_{E_2}$	$F^7_{E_2}$	$F^8_{E_2}$
$F^1_{E_2}$	2.73	0.89	0.91	1.73	1.60	1.20	1.10	0.81
$F^2_{E_2}$	0.89	1.52	0.52	0.96	0.91	0.66	0.62	0.57
$F^3_{E_2}$	0.91	0.52	1.35	0.91	0.86	0.66	0.62	0.47
$F^4_{E_2}$	1.73	0.96	0.91	2.93	1.62	1.23	1.14	0.87
$F^5_{E_2}$	1.60	0.91	0.86	1.62	2.58	1.13	1.03	0.83
$F^6_{E_2}$	1.20	0.66	0.66	1.23	1.13	1.92	0.80	0.61
$F^7_{E_2}$	1.10	0.62	0.62	1.14	1.03	0.80	1.74	0.57
$F^8_{E_2}$	0.81	0.57	0.47	0.87	0.83	0.61	0.57	1.38

(b)

	F^1_Z	F^2_Z	F^3_Z	F^4_Z	F^5_Z	F^6_Z	F^7_Z	F^8_Z	F^9_Z	F^{10}_Z	F^{11}_Z	F^{12}_Z	F^{13}_Z	F^{14}_Z	F^{15}_Z	F^{16}_Z
F^1_Z	0.10	0.01	0.01	0.01	0.01	0.01	0.01	0.03	0.01	0.03	0.01	0.01	0.01	0.01	0.02	0.01
F^2_Z	0.01	0.06	0.00	0.01	0.00	0.01	0.01	0.01	0.00	0.01	0.01	0.01	0.01	0.00	0.01	0.01
F^3_Z	0.01	0.00	0.04	0.00	0.00	0.00	0.01	0.01	0.00	0.01	0.01	0.01	0.01	0.01	0.01	0.01
F^4_Z	0.01	0.01	0.00	0.07	0.00	0.01	0.01	0.02	0.01	0.02	0.01	0.01	0.01	0.01	0.01	0.01
F^5_Z	0.01	0.00	0.00	0.00	0.04	0.00	0.01	0.01	0.00	0.01	0.01	0.00	0.01	0.00	0.01	0.01
F^6_Z	0.01	0.01	0.00	0.01	0.00	0.07	0.01	0.02	0.01	0.02	0.01	0.01	0.01	0.00	0.01	0.01
F^7_Z	0.01	0.01	0.01	0.01	0.01	0.01	0.08	0.02	0.01	0.02	0.01	0.01	0.01	0.01	0.01	0.01
F^8_Z	0.03	0.01	0.01	0.02	0.01	0.02	0.02	0.34	0.02	0.06	0.03	0.02	0.02	0.01	0.04	0.03
F^9_Z	0.01	0.00	0.00	0.01	0.00	0.01	0.01	0.02	0.07	0.02	0.01	0.01	0.01	0.01	0.01	0.01
F^{10}_Z	0.03	0.01	0.01	0.02	0.01	0.02	0.02	0.06	0.02	0.34	0.03	0.02	0.02	0.02	0.04	0.03
F^{11}_Z	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.03	0.01	0.03	0.11	0.01	0.01	0.01	0.02	0.01
F^{12}_Z	0.01	0.01	0.01	0.01	0.00	0.01	0.01	0.02	0.01	0.02	0.01	0.06	0.01	0.01	0.01	0.01
F^{13}_Z	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.02	0.01	0.02	0.01	0.01	0.09	0.01	0.02	0.01
F^{14}_Z	0.01	0.00	0.01	0.01	0.00	0.00	0.01	0.01	0.01	0.02	0.01	0.01	0.01	0.05	0.01	0.01
F^{15}_Z	0.02	0.01	0.01	0.01	0.01	0.01	0.01	0.04	0.01	0.04	0.02	0.01	0.02	0.01	0.17	0.02
F^{16}_Z	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.03	0.01	0.03	0.01	0.01	0.01	0.01	0.02	0.14

(c)

Figure 4.14: Mutual information estimation between different filters in the encode path of PCAES, F^i_k where i represents the index of the filter and k the encoder layer. The encode path is composed by (a) E_1, (b) E_2 and (c) Z. The trace of the heat-map represents $\mathcal{I}(F^i_k; F^i_k) = \mathcal{H}(F^i_k)$. The colorbar interpretation is applied individually to each row.

of the filters is not totally contained in $F^2_{E_2}$. Therefore, it can be concluded that in this space, where the number of filters has increased, the information has been distributed in a way that does not indicate any significant issues. While there is still redundancy among filters, the information in each filter is not identical, filters are not completely redundant, and each filter transmits different information.

In the final latent space Z, which serves as the bottleneck of the AE and represents the optimal-compressed representation of X, minimal redundant information is retained, and the entropy is low, as depicted in Fig. 4.13c. Given the specific characteristics of the DFU dataset (see Section 4.2.1), where the background uniformly appears as black, and the images exhibit similarities, it's reasonable that the entropy is low in low-resolution images, as likely values are undervalued. In addition, it may also be influenced by the saturation effect introduced by the sigmoid function, which bounds values between 0 and 1.

Specifically, in the case of Z in PCAE, as shown in Fig. 4.13c, two particular filters, F^2_Z and F^8_Z, exhibit notably lower entropy values compared to other filters. This is likely because the outputs of these filters produce images with uniform values within a narrow range, indicating reduced information content. The saliency maps (Fig. 4.10a) further support the findings derived from the analysis of Fig. 4.13c. Consequently, it can be inferred that these filters could be removed without significantly impacting the classifier's performance, since those do not contain relevant information. However, there is low redundancy in Z, so the impact of the remaining filters is much more significant compared to a latent space with higher redundancy, where information may be distributed across different filters.

Turning our attention to PCAES, as illustrated in Fig. 4.14, a similar pattern to PCAE emerges, which is attributed to the DPI phenomenon where information progressively diminishes through the layers. However, as previously discussed, the DPI violation subsides over time, as evidenced by the gradual loss of information in the latent space Z during training (refer to Fig. 4.12b, where $\mathcal{I}(X; Z)$ reaches its minimum value at the end of training).

Considering the latent space E_1 from PCAES (Fig. 4.14a), most filters share a high MI estimation with $F^3_{E_1}$ and $F^4_{E_1}$. It corresponds to the filters with the highest entropies. Unlike the previous case, it is observed that the distribution along the filters does not seem so gradual. However, in general, it looks like a good compression of the input X with minimal information loss as in PCAE.

In the case of the latent space E_2 in PCAES (Fig. 4.14b), a trend similar to PCAE is observed. Information in the different filters is less redundant, leading to decreased mutual information estimates among different filters. However, unlike PCAE, where a few filters clearly dominate in terms of information content (Fig. 4.13b), PCAES lacks a singular predominant filter. Nevertheless, in general, it can be inferred that this representation has less distinctive filter roles compared to PCAE. In both cases, it appears that E_2 effectively transmits information without one filter significantly outweighing the others in importance.

The primary distinction between PCAE and PCAES lies in their respective latent spaces, specifically in Z. As concluded in the previous section, the use of skip-connections in PCAES disrupts the Markov property, making it uncertain that the bottleneck provides an effective compression of the input X. Consequently, during the training of PCAES, the $\mathcal{I}(X; Z)$ continually decreases, reaching its minimum value by the end of training (Fig. 4.12b). As a result, Fig. 4.14c demonstrates that most filters in Z exhibit low entropy, signifying that their activations are concentrated in limited regions, resulting in a homogeneous output. Furthermore, the more intricate features correspond to higher values in MI estimation, such as $F^{8,10}_Z$ (Fig. 4.14c), although this is not clearly discernible in the saliency map presented in Fig. 4.10b, mainly due to the reduced impact of these features, as indicated by the colorbar in the figure.

In conclusion, this section has provided insights into the impact of skip-connections on information propagation within DNN, particularly in the context of transfer learning with deep neural networks. This analysis contributes to a more in-depth understanding of how DNN operate, shedding light on their inner workings and the implications of skip-connections on information flow.

4.3.2 Dropout effect in AutoEncoders

In our quest to understand and validate DL models, it has been conducted various experiments employing simple models to minimize potential biases in the results. However, one aspect that has not been thoroughly explored is the impact of the dropout technique within DNN since it was not considered for the experiments in 4.2.2. This section introduces an analysis of an ad-hoc convolutional AE that has been designed to incorporate the dropout technique.

As previously introduced in Section 2.2.4, dropout is a popular regularization technique frequently used in DL models to mitigate overfitting [73]. It achieves this by applying multiplicative

Bernoulli noise to each hidden unit within the deep neural network. In our analysis, it is proposed a modification in the AE to stabilize the effect of the dropout technique.

The AE utilized in this section is an adaptation of the AE discussed in earlier sections, as depicted in Fig. 4.5b. Consequently, the number of filters in the convolutional layer remains the same, but it has been introduced a batch normalization [224] layer after each convolutional layer. This technique standardizes the inputs to one layer for each mini-batch, which has the effect of stabilizing the learning process and dramatically reducing the number of training epochs required to train DNN. Additionally, the original activation function, which was a sigmoid, has been replaced with a ReLU. Finally, the dropout technique is applied at the end of each layer of the encode path.

In this section, it is considered three different configurations based on the dropout rate ρ_d depicted in Eq. (2.19), which represents the probability of an element being set to zero, in the AE. The three configurations correspond to $\rho_d = \{0.5, 0.2, 0\}$. The last case, with $\rho_d = 0$, is equivalent to not applying dropout at all. This results in a modified AE for each configuration, allowing us to assess the impact of different ρ_d on the AE's performance and ability to reduce overfitting.

Figure 4.15 displays the IP estimation during the training process on the FMNIST dataset. You can find the IP estimation during the fitting phase on the DFU dataset in Appendix A.1.

An essential consideration is that the DPI, as defined for AE in the decode path (Eq. (4.4)), is not met in any case. However, this does not necessarily indicate a problem with the AE, contrary to the original proposal by Yu et al. [190]. While the AE still have the goal of obtaining an optimal lossy compressed representation by following the InfoMax principle (see Section 4.1), as described in Eq. (4.6), the introduction of noise through the dropout technique alters the AE's performance.

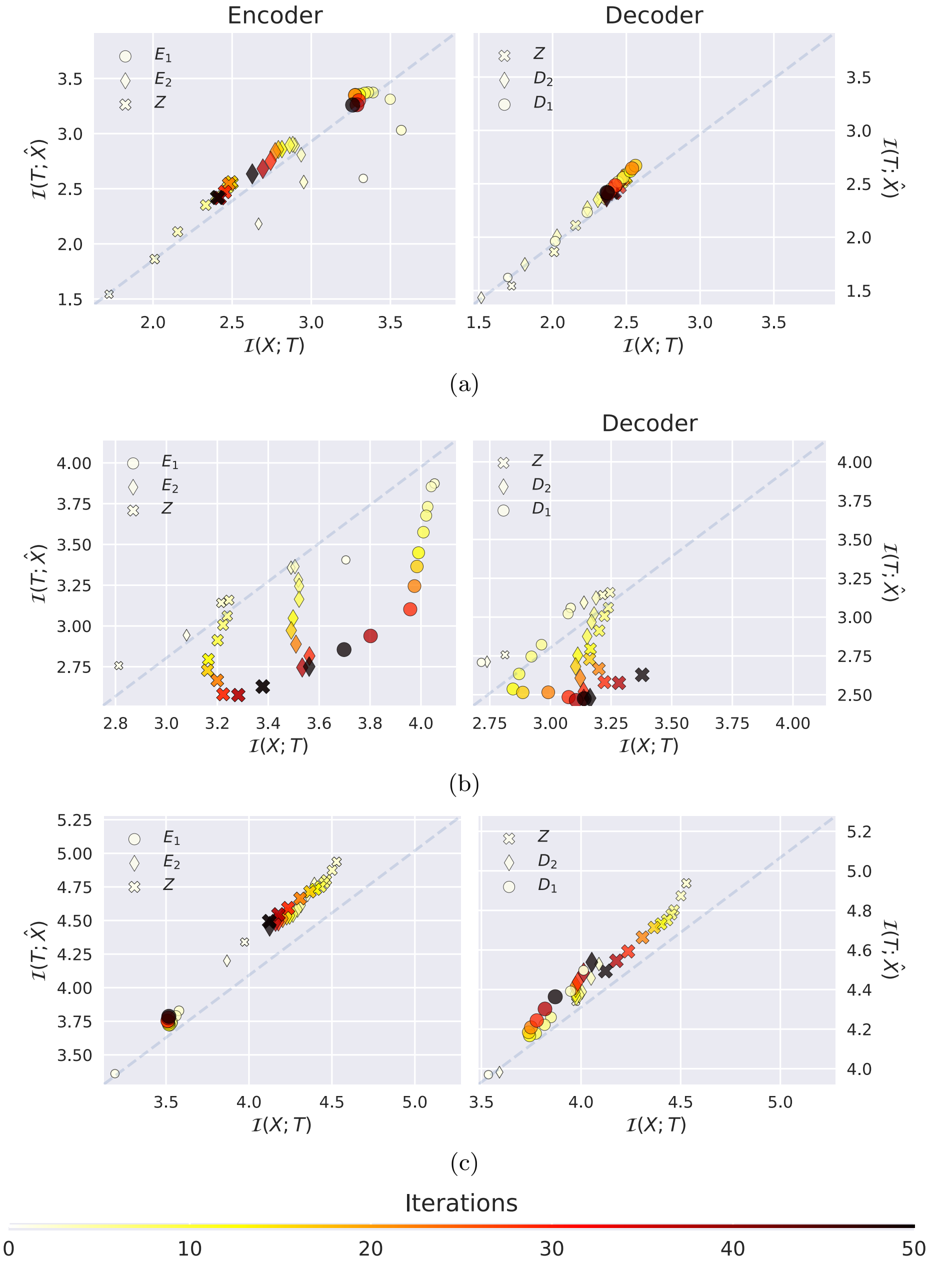
The usage of the dropout technique forces the AE to exhibit tolerance towards noise. In essence, the AE is now effectively addressing the problem described in the rate distortion theory, as defined by Eq. (3.20) (see Section 3.4). Consequently, the DPI must meet the same conditions observed in the IB (Eq. (3.17)). In this context, the ideal AE should satisfy the following condition [175]:

$$\mathcal{I}(X; \hat{X}) = \mathcal{I}(Z; \hat{X}) = \min\{\eta,\, \mathcal{I}(X; X')\}, \tag{4.7}$$

where η is the maximum amount of information that the bottleneck, Z, contains. Consequently, the bottleneck determines the maximum amount of information that can be conveyed in the decode path. Unlike the AEs analyzed in Section 4.3.1, in this case, the decode path must converge to $\mathcal{I}(X; Z)$. In conclusion, it will never be satisfied the Eq. (4.4) when dropout technique is applied in AEs.

This effect can be intuitively understood by considering that the information rate, which is the amount of information transmitted, is influenced by the distortion introduced by noise. To mitigate this distortion, redundancy in information is necessary. In the context of a DNN, different neurons must propagate the same information to account for potential distortions in the channel. The dropout technique deactivates a neuron randomly, the model has to deal with this case having a redundant way to propagate this information. Consequently, as described by the rate distortion theory, the amount of information in the layer is compressed due to the noise, more redundant information leads to a lower information rate.

As a result, the only way for an ideal AE to satisfy Eq. (4.7) is to create a latent space Z large enough to ensure the propagation of all information even through different redundant pathways. This approach compensates for the potential distortions introduced by the dropout technique by



(a)

(b)

(c)

Figure 4.15: IP estimation using dropout technique in the AE among different dropout rates (ρ_d) training with FMNIST dataset: (a) $\rho_d = 0.5$, (b) $\rho_d = 0.2$ and (c) No dropout ($\rho_d = 0$).

maintaining sufficient redundancy in the latent space.

As observed in Fig. 4.15a and Fig. 4.15b, the application of the dropout technique with different ρ_d conditions the amount of information. Specifically, in Fig. 4.15a, where $\rho_d = 0.5$, $\mathcal{I}(X;Z)$ remains around 2.5 bits. In contrast, Fig. 4.15b demonstrates an increase in $\mathcal{I}(X;Z)$ to approximately 3.4 bits. Finally, Fig. 4.15c, where the dropout technique is not applied, illustrates that $\mathcal{I}(X;Z) \simeq 4.15$ bits.

The IP estimation in Fig. 4.15 confirms that the AE's performance by using dropout technique can be described by the rate distortion theory. To validate the model, it needs to be considered the DPI. A DPI violation is only observed when the dropout technique is not applied, as shown in Fig. 4.15c. In this case, a clear violation in the encode path (left in Fig. 4.15c) is evident, $\mathcal{I}(X;Z) > \mathcal{I}(X;E_1)$. Additionally, the various convolutional layers tend to approximate the output $(\hat{X})$ more closely than the input (X). This behavior is why the MI evolution is observed above the diagonal that represents the optimal reconstruction of the AE, where the $\hat{X}$ closely matches the original input X. In general, there are clear indications that the modified AE has overfitting issues, which are a consequence of the uses of batch normalization in the convolution layers. As it was previously mentioned, this normalization allows reducing the number of training epochs required to train DNN and it was applied to stabilize the effect of the dropout technique.

Considering both cases where dropout is applied, it can be concluded that both AE models work properly. As mentioned earlier, based on the rate distortion theory, there should be a constant loss of information during the propagation due to the distortion introduced by the noise produced by the dropout technique in the encode path. Since the decode path does not include dropout, the information in the different decode layers must converge to a point conditioned by Z. This means that there should not be a significant loss of information during the decode path, so $\mathcal{I}(X;D_2) \approx \mathcal{I}(X;D_1)$.

In conclusion, both Fig. 4.15a and Fig. 4.15b provide evidence that both AE models work properly. In the case of $\rho_d = 0.5$, the different convolutional layers of the decoder converge to the information in Z, as seen on the right side of Fig. 4.15a. In contrast, when $\rho_d = 0.2$, the decode path converges to a point where there is a slight loss of information considering the bottleneck, $\mathcal{I}(X;D_1) \approx \mathcal{I}(X;D_2) > \mathcal{I}(X;Z)$, but it still satisfies the DPI condition established by Eq. 4.7.

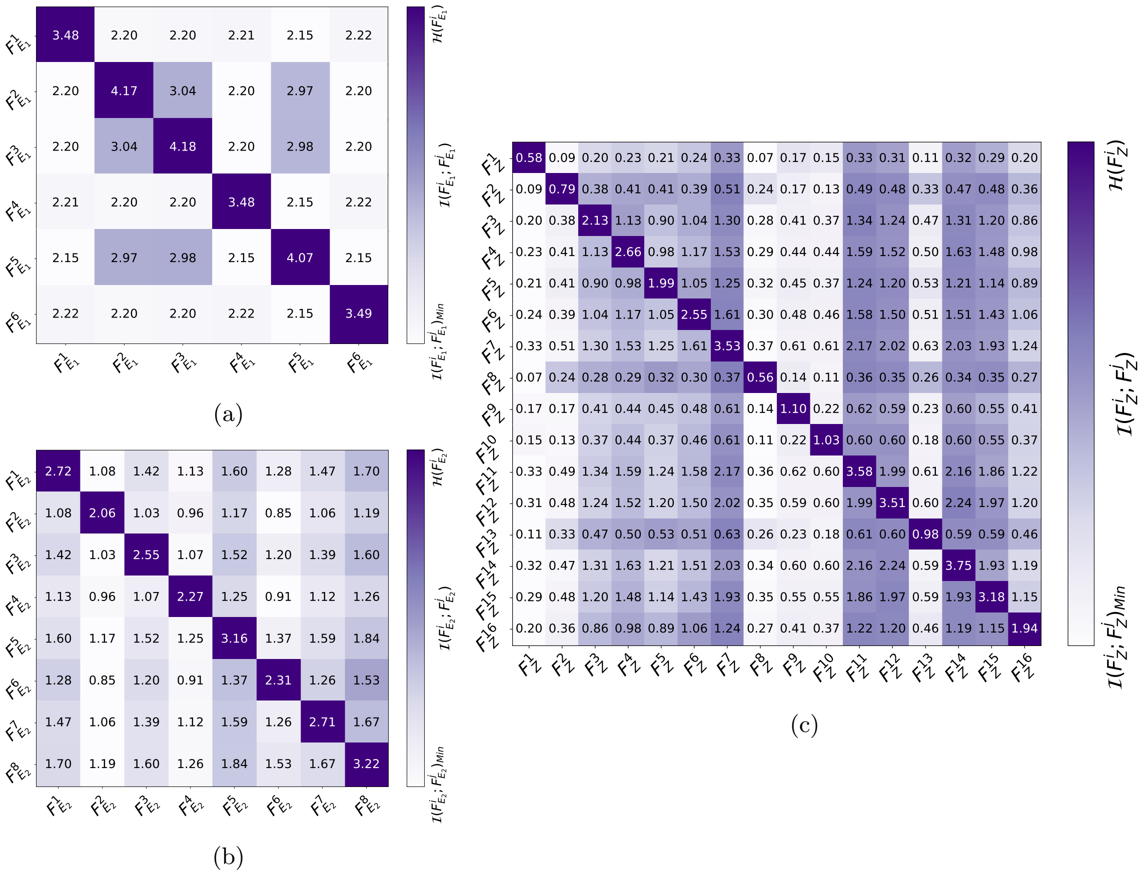
Finally, Fig. 4.15b illustrates a consistent decrease in $\mathcal{I}(T;\hat{X})$ in both cases, for ρ_d equal to 0.5 and 0.2. However, it is evident that $\mathcal{I}(Z;\hat{X})$ converges to approximately 2.5 bits in both cases. This suggests that both $\hat{X}$ produced by different configurations are similar, conditioned by the decode path, but the second configuration contains more information related to X. To address this, a more complex AE is needed, involving an increase in the number of layers and parameters.

As it has been done in previous sections, these AEs were fitted using the DFU dataset. The conclusion drawn from the IP estimation align with those discussed for the FMNIST dataset, and it can be found in Appendix A.1.

Convolutional Filters

As it was done in the previous Section 4.3.1, the convolutional filters produced in the aforementioned AEs will be analyzed by the kernel-based entropy estimator to gain further insights into their performance.

Figure 4.16 illustrates the example where the ρ_d is 0.5 and Fig. 4.17 represents the AE in which the ρ_d is set to 0.2. Firstly, it is noticed that the ρ_d parameter affects the redundancy in the filters.



(a)

(b)

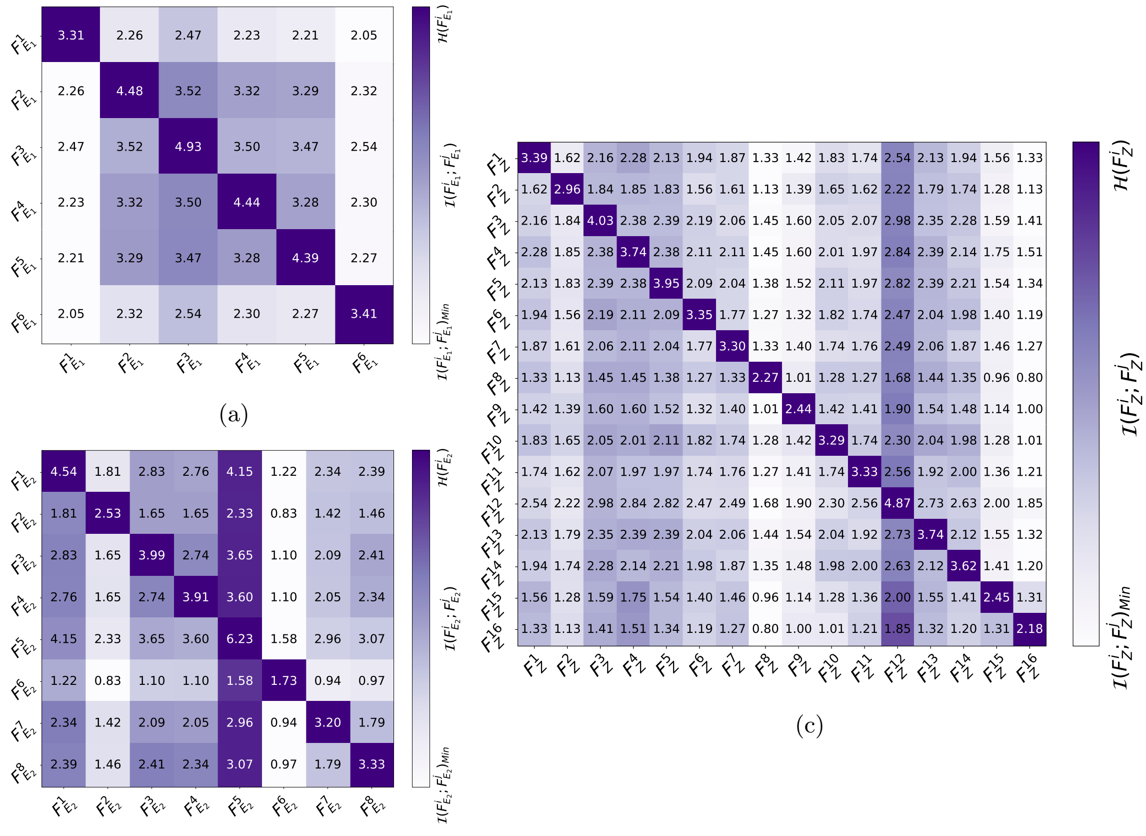
(c)

Figure 4.16: Mutual information estimation between different filters in the encode path of the AE with a dropout rate of 0.5 at the end of each layer, denoted as F^i_k where i represents the index of the filter and k the encoder layer. The encode path is composed by (a) E_1, (b) E_2 and (c) Z. The trace of the heat-map represents $\mathcal{I}(F^i_k; F^i_k) = \mathcal{H}(F^i_k)$. The colorbar interpretation is applied individually to each row.

Increasing the ρ_d the entropy of each filter is reduced, it is specially relevant in the bottleneck as it is illustrated in Fig. 4.16c and Fig. 4.17c where the entropy of filters is lower with a higher ρ_d. This behavior is observable in the different layers, including in E_1 and E_2. In addition, the MI estimation among the different filters is reduced when the dropout rate, ρ_d, is higher.

Comparing the MI estimation in E_2 for both cases, as shown in Fig. 4.16b and Figure 4.17b, it is apparent that the MI is lower with a dropout rate of 0.5. Specifically, $F^5_{E_2}$ exhibits an entropy of 6.23 bits, which is the filter with the highest MI. This could imply that this filter doesn't handle the noise produced by the dropout technique effectively, resulting in higher uncertainty. However, it is worth noting that the information contained in this filter might be found in other filters.

For instance, by examining $F^4_{E_2}$ in Figure 4.17b, it is evident that the conditional entropy, which represents the remaining uncertainty when observing another variable, $\mathcal{H}(F^4_{E_2}|F^5_{E_2})$, is only 0.31 bits. This means that when you observe $F^5_{E_2}$, $F^4_{E_2}$ becomes almost entirely predictable. The reduced redundancy in filters in this case, due to less noise, results in more information being contained within the filters. This makes sense because filters contain more information, increasing the likelihood that this information is present in another filter.



(a)

(c)

(b)

Figure 4.17: Mutual information estimation between different filters in the encode path of the AE with a dropout rate of 0.2 at the end of each layer, denoted as F_k^i where i represents the index of the filter and k the encoder layer. The encode path is composed by (a) E_1, (b) E_2 and (c) Z. The trace of the heat-map represents $\mathcal{I}(F_k^i; F_k^i) = \mathcal{H}(F_k^i)$.

In contrast, Figure 4.16b shows that in the encoder layer E_2 of the AE with $\rho_d = 0.5$, the highest entropy value is found in $F_{E_2}^8$ which is considerably lower than in the other case observed in Figure 4.17b.

From a qualitative perspective, both figures illustrate the redundancy of information based on the uniformity of colors. Using as reference the second encoder layer E_2, Fig. 4.16b and Fig. 4.17b show this property. In Fig. 4.16b, the colors have a more uniform distribution, indicating minimal differences in MI among filters. This uniformity suggests a higher redundancy of information, as it was previously discussed. Conversely, Fig. 4.17b displays less uniform colors, indicating greater variability in MI among filters.

In conclusion, both figures support the expected results discussed in Figure 4.15 and align with the principles of the rate distortion theory. An increase in the dropout rate, denoted as ρ_d, leads to a loss of information as the system must compensate by increasing redundancy to address the noise. Consequently, the MI values between filters are more similar to each other, resulting in a more homogeneous color distribution, as shown in Fig. 4.16 compared to Fig. 4.17.

One noticeable difference in the bottleneck Z, when compared to the results described in Figure 4.13, is the higher entropy values in these experiments. This discrepancy is primarily attributed to the activation function used. As mentioned earlier, in order to mitigate the dropout effect in the AE, a batch normalization layer followed by a ReLU activation function was employed. The ReLU function lacks an upper limit, while the sigmoid function is bound between 0 and 1. Consequently, the sigmoid function exhibits a saturation effect in the compressed latent space Z, where most values approach 0 or 1. This saturation results in lower entropy estimations, as likely values are underestimated

For the reader's reference, the results of the AE without the dropout technique are provided in Appendix A.1.2. However, it should be noted that this AE has previously been demonstrated to not perform adequately.

4.4 Discussion

Artifical Inteligence is a powerful tool for practical clinical decision support systems, offering, for instance, assistance to radiologists in decision-making and helping to alleviate the workload issues discussed earlier in this chapter [5]. However, the lack of sufficient labeled data presents significant limitations for the use of supervised AI algorithms. There is a trade-off between the number of samples used for training and those used to validate the model, making it difficult to determine whether the model is overfitted. Additionally, clinical decision-making relies heavily on evidence interpretation, which is further complicated by the opaque nature of DNN models.

In this chapter, we have explored an emerging area of research related to the early-stage detection of DFU using DNN models as a case of study. Our focus has been on validating these models from a theoretical perspective based on IT. It is important to note that this research area faces challenges related to data availability. The dataset consists of thermogram images, and the experiments have been conducted with a limited test set of only 32 samples, which may not be representative enough to thoroughly evaluate the models.

Additionally, the training dataset comprises only 209 images, which is a relatively small number for DL models. In cases where the dataset size is insufficient, transfer learning is a common approach to address this limitation. Therefore, in three of our experiments, we applied transfer learning by using AEs pretrained on the Fashion-MNIST (FMNIST) dataset. Despite the fact that the FMNIST dataset focuses on fashion-related elements, which is not associated with our data, its quantity and quality of samples have proven to be robust for feature extraction in our experiments. This is particularly noteworthy because both datasets share certain characteristics, such as grayscale images with a uniform background.

Initially, the experiments in this work, as outlined in Section 4.2.2, appeared to yield remarkable metrics, as evidenced in Table 4.3, which included some perfect scores. At first glance, it might have been concluded that all the experiments exhibited exceptional performance, with accuracy scores consistently exceeding 90%.

However, as this chapter has demonstrated, such a conclusion does not hold true when considering the information transmission in a more nuanced way. Upon closer inspection, it became evident that two experiments, specifically those involving transfer learning, displayed superior information transmission capabilities compared to the others. This reveals that while traditional classification metrics may suggest high performance, a more in-depth analysis focused on the IB principle can uncover important distinctions in model performance and information processing. This analysis, primarily based on the IP, has been validated using the kernel-based entropy estimator for a scarce dataset. Theoretical characteristics, already discussed in the state-of-the-art with popular datasets [97, 169], were identified in our analysis. This was particularly evident in experiments where transfer learning was applied, providing validation for the analysis, especially in the context of small datasets.

On one hand, in the experiments where transfer learning has not been applied, specifically NPCE and NPNE, both show instances of DPI violations. In the case of NPCE (Fig. 4.9d), the DPI violation is evident between L_2 and L_3, where $\mathcal{I}(X; L_2) > \mathcal{I}(X; L_3)$. In NPNE (Fig. 4.9e), the DPI violation occurs between L_3 and L_4, with $\mathcal{I}(L_3; Y) > \mathcal{I}(L_4; Y)$. These DPI violations are likely indicative of overfitting.

Moreover, NPNE exhibits a pattern that suggests a potential issue with underfitting, implying that the number of parameters is insufficient to adequately characterize the data and capture the intricacies of the input-target relationships. The MI estimation in NPNE becomes trapped within a narrow range, fluctuating without a discernible increasing or decreasing trend in the information plane (Fig. 4.9e).

On the other hand, the experiments where the transfer learning from AEs has been applied, specifically PCAE, PCAES, and PFCAE, show a consistent phase of fitting in most layers and there are no observed DPI violations in the different IPs. In the specific case of PCAES (Fig. 4.9b), the $\mathcal{I}(T; Y)$ is around 0.4, which is far from the theoretical maximum value. In this case, the theoretical maximum value is 1, since the limited test set of only 32 samples is balanced so the $\mathcal{H}(Y) = 1$. In PCAE and PFCAE (Fig. 4.9a and 4.9c), the $\mathcal{I}(T; Y)$ reaches a maximum value around 0.7. However, this detail is also observable from the objective function (Fig. 4.8).

The issues observed in the pretrained experiment PCAES are primarily attributed to the use of skip-connections in the AE during the transfer learning process. To gain a more in-depth understanding of the impact of skip-connections, a comparison was made between the convolutional AE in both experiments, PCAE and PCAES.

As discussed in Section 4.3.1, the introduction of skip-connections in PCAES disrupts the Markov property, making it uncertain whether the bottleneck provides an effective compression of the input X. This disruption led to a DPI, Eq. 4.4, violation during the pretraining of the AE using the FMNIST dataset (Fig. 4.11b). However, during the fitting phase with the DFU dataset, the DPI violation disappears, as a consequence that the latent space Z is constantly loosing information (see Fig. 4.12b).

A more detailed examination reveals that the inadequate compression of information in Z for PCAES becomes evident when analyzing the individual convolutional filters. Figures 4.13 and 4.14 illustrate the amount of information and redundancy among filters in each latent space of the encoding path: E_1, E_2, and Z. Focusing on the compression in Z, Fig. 4.14c demonstrates that most filters in Z of PCAES exhibit low entropy, indicating that the number of activations in this layer is limited. In contrast, PCAE (Fig. 4.13c) shows higher entropy values and lower redundancy among filters, suggesting that the information in these channels is complementary. These findings are further supported by the saliency maps estimated using the post-hoc method DeepLIFT (Fig. 4.10).

Considering the different theoretical characteristics, which also can be found in the state-of-the-art, it can be concluded that PCAE and PFCAE are the most promising models and the other models are not as good as the Table 4.3 suggested. In this way, the proposed kernel-based entropy estimator has been validated in those case where the number of samples is limited, since all this analysis is obtained using only 32 samples.

In this chapter, a description of the DNN's problem-solving mechanism is provided in this chapter based on the information-theoretical perspective. However, considering that AE is a self-supervised DL architecture that has a remarkable similarity with a transmission channel, as described in Section 3.2, we decided to extend this work to analysis another popular technique in DL modes, the dropout technique.

The outcomes reveal that an AE using dropout cannot strictly adhere to the InfoMax principle, so it cannot satisfy the description of Section 4.1. Instead, the architecture employing dropout appears to align more closely with the rate distortion theory, which can be considered a specific case of the IB, as elaborated in Section 3.5, and it is well-defined for supervised DL models. Considering that the dropout rate indicates the amount of distortion that the AE has to tolerate, an increase of this parameter leads to a lower information rate.

The results in Section 4.3.2 have illustrated this approach based on the distortion rate theory, aligning with the description of AEs described by Tapia et al. [175]. In essence, it was observed that the decode information flow is limited by the bottleneck Z. For this reason, the decode path DPI defined in Eq. (4.4) cannot be satisfied. Moreover, the convolutional filters presented in Fig. 4.16 and Fig. 4.17 effectively portray how increasing the dropout rate intensifies the input compression. Consequently, this heightened compression results in reduced redundancy among filters due to the increased redundancy within each self-filter, as indicated by the lower entropy value.

Finally, the analysis presented in this chapter is scalable and can be applied to DNNs with a larger number of layers. While this study focused on validating the method using custom DNNs with a controlled number of layers, it can certainly be extended to more complex DNN architectures with deeper layers.

The significance of this analysis becomes particularly relevant in cases where the dataset is small. While it may be less critical for large datasets, understanding the flow of information among different layers of a neural network can serve as a valuable metric to ensure the model's effectiveness and interpretability. In the case of a scarse dataset, it has been proved that the uses of the Kernel-Based entropy estimator works properly.

In the context of healthcare, integrating this algorithm for the analysis of radiological imaging data holds great promise. It has the potential to mitigate human bias and reduce the time burden on radiologists by providing a more transparent and interpretable approach to image analysis. As a result, this technology could contribute to alleviating the heavy workload faced by radiologists today. Furthermore, the proposed analysis from a theoretical perspective not only enhances the interpretability of the 'black box' but also underscores the crucial role these advancements could play in shaping the future of medical diagnostics and patient care.

4.5 Conclusions

Deep Learning models, characterized by their multilayer nonlinear structure, have earned significant attention due to their impressive performance in various tasks. However, these models are considered 'black boxes' due to their complex relationships between layers and vast parameter space. Despite their opacity, DL models have rapidly replaced many traditional techniques in CV for image processing. They hold the potential to be valuable tools in aiding radiologists and healthcare professionals in clinical decision-making.

Nonetheless, challenges related to transparency of these models and data availability must be addressed when integrating DL algorithms into the medical domain. The need for large labeled datasets to allow these models to generalize effectively is crucial. Typically, validation of DL models is performed using classical classification metrics on a separate test dataset not seen during training. However, in fields like medicine, data collection can be time-consuming and challenging.

This chapter builds upon our work presented in [172], where different DL models for medical image classification were evaluated using the theoretical framework of the IB principle, providing an approach to understand the problem-solving mechanism of the DNN. As described in this chapter, various DL architectures were implemented, in which the training dataset is rather small in terms of samples. This analysis helps determine whether the promising performance metrics seen in traditional classification evaluations are due to overfitting, or if the models indeed exhibit robust generalization capabilities.

Furthermore, the AE, which is a self-supervised deep learning architecture that shares similarities with a transmission channel, was explored in the context of transfer learning in the experiments discussed earlier. The analysis of AEs aimed to understand the effects of different techniques on their performance. The results indicated that utilizing AEs for transfer learning aligns with the InfoMax principle, signifying that these models effectively learn to maximize MI between input and output. However, when skip-connections are introduced in AEs, it cannot be guaranteed that it satisfied the InfoMax principle.

In addition, it was demonstrated that applying the dropout technique transforms AEs into models described by distortion rate theory. By increasing the dropout rate, these AEs achieve a more compressed representation of the input data, X, at the expense of introducing distortion. This insight sheds light on the behavior of DL models, as layers in a DNN can be characterized by the trade-off between compression and predictability. In this context, the dropout technique emphasizes data compression. Furthermore, this approach provides valuable insights into the behavior of AEs and their suitability for specific applications.

The analysis presented in this work offers a perspective for opening the ‘black-box’ of DL models, providing a more in-depth understanding of their inner workings. This increased transparency and interpretability can be invaluable in the integration of AI in healthcare scenarios. It has the potential to mitigate human bias and reduce the time burden on radiologists by offering a more transparent and interpretable approach to image analysis. This enhanced understanding can lead to more reliable and trustworthy AI systems, particularly in critical applications such as medical imaging and diagnosis. It is important to emphasize that this demonstration is particularly relevant in scenarios with limited data, such as the case study presented in this chapter, where only 32 samples were available for testing. The methods and insights provided here are especially valuable in situations where data scarcity can be a significant challenge.

Chapter 5

Feature Selection with Deep Learning: Hyperspectral Band Selection

> Once you understand the way broadly, you can see it in all things.
>
> *Miyamoto Musashi*

Chapter 4 introduced ethical challenges related to the integration of complex AI models in healthcare settings, focusing on issues such as model transparency and the scarcity of data in specific contexts. The second challenge arises from the fact that supervised ML and DL algorithms typically require a large number of samples to prevent overfitting and enhance their generalization capacity. To address these challenges, a perspective based on IT was proposed for describing these models by examining the information flow, ultimately improving model interpretability. The analysis based on this perspective helps identify issues in algorithms with limited data samples, as demonstrated by the methodology employed.

Furthermore, this approach can guide the design of DL models. In this chapter, a FS method based on DL will be introduced, interpreted through the IT perspective, and applied to select relevant wavelengths in HSI — a process known as band selection — for biological tissue identification. Band selection treats the spectral signature obtained from HSI as a feature vector and aims to obtain a sparse representation (see Section 2.5), offering a more interpretable compressed representation compared to dimensionality reduction methods [225]. This method is not limited to HSI data; it can also be applied to other datasets, as shown in Appendix B, where the proposed DL-based FS method is used, among others, to select the most relevant features extracted from the thermogram dataset for DFU detection used in Chapter 4 [226, 227].

In healthcare, interpreting high-dimensional data is a major challenge, not only for HSI but across various clinical domains. Although FS methods, like band selection in HSI, can help improve interpretability, many datasets still suffer from the 'curse of dimensionality' [228]. This expression, attributed to Richard Bellman, refers to the problems with high-dimensional data such as, for instance, optimizing any function with too many input random variables. To overcome the 'curse of dimensionality', it is necessary to apply the process of 'dimensionality reduction', which consists of transforming a high-dimensional data into a low-dimensional space [229]. Following the main branch of this thesis, the dimensionality reduction can be intuitively expressed as a compressed representation of the data. As it was demonstrated in the Chapter 4, DL models are good at this task.

Given that the understanding of the problem associated with a clinical case is considered a fundamental task, and nowadays, it is commonly associated with high-dimensional data, this task is computationally and statistically challenging. The algorithms of dimensional reduction, such as Principal Component Analysis (PCA) [230] or ICA, can be used to reduce the cost of analyzing the data. However, it is important to note that while these techniques effectively reduce dimensionality and capture the majority of data variance, they may not necessarily improve the interpretability of the data. For instance, while PCA is a beneficial technique for reducing dimensionality and capturing the majority of variance in the data, it does not necessarily enhance the interpretability of the data in the sense that the new dimensions generated by PCA are linear combinations of the original features, representing an abstraction of the original feature vector, and therefore may pose a challenge to interpret in terms of the original variables.

In the context of healthcare, where comprehensive data interpretation is essential, the curse of dimensionality presents a formidable obstacle [231]. A bad interpretation of the data can result in inconsistencies among experts when diagnosing a disease, leading to increased variability in clinical decision-making [232]. A clear example of this challenge can be observed in the diagnosis of retinopathy of prematurity [233], a condition that affects infants with low birth weight, leading to abnormal blood vessel development in the eye's retina and the potential for blindness [233]. For the diagnosis, CV techniques are employed to extract features from retinal images, and medical experts base their clinical decisions on these features. However, diagnosing retinopathy of prematurity has shown significant variability among experts, as some may prioritize certain features more than others [234]. Given that the issue underlying this inconsistency in diagnosis stems from the high-dimensional nature of the data, the application of algorithms for identifying the most relevant features, a process known as FS, offers a viable solution.

Feature Selection is a field that combines statistical multivariate and ML methods to reduce the number of input variables or features [229]. The main goal of FS is to identify an optimal subset of input features, specifically the most relevant ones. Unlike dimensionality reduction methods like PCA, which derive a smaller set of features through linear combinations of the original ones, FS focuses on selecting a subset of the original features. This is particularly important for making high-dimensional data more interpretable [229], as it retains the original meaning of the features rather than transforming them into abstract components.

In ML, FS also improves model interpretability by offering clearer insights into the importance of individual features in decision-making, which is the fundamental task in 'explainable ML' (see Section 2.3.3). Additionally, by removing redundant or irrelevant input data, FS enhances classifier performance. This leads to more efficient classifiers with better cost-performance ratios, as reducing the number of features decreases both computational and memory requirements [235].

As mentioned earlier, DL models are highly effective at achieving compressed representations of the input data. In particular, AEs were originally designed for this purpose, dimensionality reduction to obtain an optimal representation in a lower-dimensional space. However, they operate at a higher level of abstraction, which complicates the interpretability issues seen in classical methods such as PCA, as the resulting features are too abstract and impossible to understand. Since DNNs are inherently capable of learning compressed representations of data, they should, in theory, be well-suited to performing FS. However, DL models typically do not perform feature selection; they do not discard irrelevant or redundant features from the original input space. Consequently, a typical architecture of a DNN is not well-suited for FS.

In this chapter, an adaptation of the DL model will be presented to achieve a compressed representation of the input based on FS, and how this approach can be useful in obtaining a more comprehensive interpretation of data in a medical case study. The model will be analyzed using the IT approach, identifying key characteristics required to achieve an optimal solution, while highlighting the limitations of the proposed method. This chapter is structured as follows: first, the state-of-the-art in FS will be reviewed. Then, the proposed FS method, based on DL and the IT perspective, will be introduced to provide a more in-depth understanding of the problem-solving mechanism. A case study focused on band selection in HSI will be used. Finally, validation of the IT perspective will be achieved through IP estimation.

5.1 Taxonomy of Feature Selection methods

Feature Selection methods can be traditionally categorized into the following classes: filter, wrapper, and embedded methods [229]. This section describes various state-of-the-art methods, along with relevant examples.

5.1.1 Filter Methods

Filter-based methods for feature selection are characterized by their reliance on a scoring criterion or statistical property to identify and filter out less important features or variables [229]. Unlike some other techniques, these methods do not require a predictive model for scoring variables, making them suitable for use as a preprocessing step. However, the filter-based feature selection is a computationally efficient and straightforward method, but it has some limitations. It does not consider feature interactions or the context of the model, so it may not always select the best features for a particular model or dataset.

One example of a filter-based method is Relief-based feature selection, which evaluates the relevance of each feature by measuring its ability to distinguish between instances that are near to each other. Relief-based algorithms are sensitive to feature interactions and can handle various data characteristics, such as classification, regression, missing values, and noise. Relief-based feature selection has been applied to various biomedical problems, such as gene expression analysis, biomarker discovery, and disease diagnosis [236].

Another popular approach involves a filtered method based on MI. As it has been described in Section 3.1.3, This metric serves to evaluate the relevance of a subset of features in predicting the target variable, as well as assess redundancy in relation to other variables [237]. In addition to MI, other filtered-based methods may utilize statistical criteria such as correlation or chi-squared to evaluate the relationship between features.

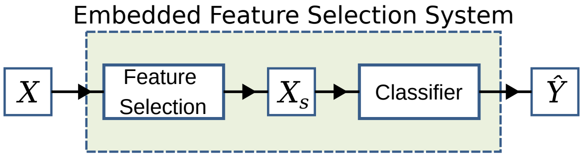


Figure 5.1: Block diagram of an embedded method for FS. The variable X_s represents the subset of the input variable with the selected features S.

5.1.2 Wrappers and Embedded Methods

Wrapper methods involve defining the search space, encompassing all possible variable subsets, and treating a model as a 'black-box' [229]. Within this framework, a search, and evaluation strategy is implemented to determine the optimal selection of variables or features. While this approach can lead to substantial advantages in terms of generalization, it comes with the drawback of being computationally intensive and might result in subsets that are overly specific to the particular classifier used.

The most common example of wrapper methods is the Recursive Feature Elimination (RFE) [238] and the Genetic Algorithm (GA) [239]. The RFE method was originally proposed to use with a SVM classifier and this method, as the name suggests, consists on removing iteratively the least relevant features based on the SVM weights until a desired number of features is reached. However, it can be applied to other classification models. On the other hand, GA is a wrapper method that is based on biological evolution. It is inspired by the process of natural selection and relies on biologically inspired operators such as mutation, crossover, and selection. Another wrapper FS method widely used for different applications is the Ant Colony Optimization (ACO) [240].

The embedded methods incorporate the variable selection in the training process, for instance using a regularization for reducing the number of variables used during classification. For this reason, this methodology is more efficient than the wrappers methods. The more illustrative example of these methods is the Least Absolute Shrinkage and Selection Operator (LASSO) [241], originally introduced in geophysics [242]. LASSO, or L_1-regularized regression, assigns zero weights to the least relevant or redundant features, effectively leading to feature selection within the model.

By using the learning machine as a 'black box', wrappers are remarkably universal and simple. But embedded methods that incorporate variable selection as part of the training process may be more efficient in several respects: they make better use of the available data by not needing to split the training data into a training and validation set; they reach a solution faster by avoiding retraining a predictor from scratch for every variable subset investigated [229].

5.1.3 Nested Subset Methods

Nested subset methods represent a category of feature selection techniques that combine both wrapper and filter methods. In these approaches, a wrapper method is initially employed to choose a subset of features, and subsequently, a filter is applied to this subset to further refine the selection process [229].

A notable example of this approach is presented in the work of Peng et al. [243], the minimum-Redundancy Maximum Relevance (mRMR). They introduced a filter-based feature selection method that relies on MI criteria to identify the most relevant features. The selected features were then integrated into a two-step algorithm, resembling a wrapper method. The mRMR is designed to find the smallest relevant subset of features for a given ML task.

Additionally, these methods can unify filter methods for applying a stop criterion in wrapper methods. In this context, the use of MI is particularly relevant, as conditional mutual information (the expected value of mutual information) naturally emerges when bounding the ideal regression/classification errors achieved by different subsets of features [237].

5.2 Deep Learning approach for Feature Selection

Utilizing wrapper methods for FS with DL models is a viable approach, albeit demanding in terms of computational resources. The employment of a DL model for FS represents an embedded method, wherein the selection of the feature subset is optimized throughout the training process.

If the DNN is considered as an embedded FS method, it becomes necessary to encourage the DNN to generate a sparse representation in its initial layer. In other words, the DNN should be designed to select a subset of features in the input layer. However, it is important to acknowledge that feature selection in certain problems is inherently an ill-posed problem (see Section 2.4.1), implying that there may not be a unique subset that represents the optimal solution. This characteristic is especially relevant in DNN due to their stochastic training processes, which can lead to variations in feature selection.

Indeed, an initial approach to achieve feature selection within a DNN would involve setting the weights in the input layer to zero for those features that are considered irrelevant, similar to the approach used in LASSO. However, the L_1 regularized approach, LASSO, is primarily designed for linear models. Nevertheless, there are ways to deal with non-linear features by using LASSO such as, for instance, Hilbert Schmidt Independence Criterion Lasso (HSIC Lasso) [244].

The HSIC Lasso method consists on applying the LASSO in the RKHS, a high-dimensional space where lineal solutions are viable. It offers the advantage of providing a clear statistical interpretation, as it can identify non-redundant features with strong statistical dependence on output values using kernel-based independence measures. The HSIC is a non-negative value that attains zero if and only if two random variables are statistically independent [244]. However, a significant drawback of this method lies in its computational complexity, which scales quadratically with the number of observations.

Nevertheless, it is crucial to acknowledge that in DNN, the non-linear space generated does not align with an RKHS. Despite this, certain approaches have implemented L_1 regularization in DNN to achieve a sparse representation. Notably, Feng and Simnon [245] introduced the concept of an input-sparse DNN, where the input weights are penalized using the group L_1 penalty. In this framework, the weights of the upper layers are subjected to a L_2 penalty. The authors demonstrated that the weights of irrelevant features converge to zero. However, it is crucial to emphasize the significance of attaining a well-fitted configuration for the Lagrangian multipliers associated with the diverse regularization penalties, as these represent hyperparameters requiring meticulous tuning.

According to the authors, LassoNet [246] stands as an extension and generalization of the previously mentioned sparse-input DNN approach. This method distinguishes itself by incorporating an input-to-output residual connection, a skip-connection representing a linear link between the input and output layers. This connection allows a feature to possess a non-zero weight in a hidden unit

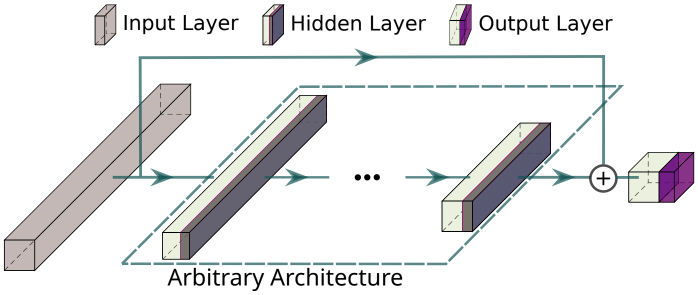


Figure 5.2: LassoNet architecture [246]. The architecture of LassoNet consists of a single residual connection and an arbitrary feedforward neural network. The residual layer and the first hidden layer are jointly passed through a hierarchical soft-thresholding optimizer.

only if its linear connection is active (refer to Fig. 5.2). The objective function is defined as follows:

$$\begin{aligned} &\min_{\theta} \ \mathcal{L}(Y, X; \theta) + \lambda ||\omega^r||_1, \\ &\textbf{subject to } ||w_j^{(1)}||_\infty \leq M|\omega_j^r|, \ \ j = 1, ..., d, \end{aligned} \tag{5.1}$$

where $\mathcal{L}(.)$ represents the objective function. The parameters of the DNN is composed by $\theta \in \{w, \omega^r\}$, where ω^r are the parameters of the residual network, the skip-connection, and w denotes the parameters of the different hidden layers. The notation $w^{(1)}$ corresponds to the parameters of the first hidden layer. The $||.||_\infty$ denotes the max norm, which signifies the maximum absolute value among the components of the vector [120].

Observing Eq. (5.1), it is clear that it is necessary to tune two important hyperparameters: the Lagrangian multiplier, λ, which controls the effect of the L_1 penalty, and the M that controls the trade-off between the linear and non-linear components. In addition, it is specially relevant because it defines the soft-thresholding for obtaining the sparse representation. Regarding the hyperparameter M, in the extreme where $M = 0$, the formulation recovers exactly the LASSO; in the other extreme (by letting $M \to \infty$), one recovers a standard feed-forward neural network with L_1 penalty on the first layer [246].

As it was previously mentioned, the concept behind FS is to acquire a sparse representation of the input vector. Typically, the sparsity of a vector is quantified using the pseudo-norm L_0, as illustrated in Eq. (2.26). However, LASSO offers a relaxation approach that establishes a convex problem in which sparsity is achieved by constraining and shrinking the weights.

In the context of DNN, the operation of the L_1 penalty becomes more intricate due to the network's deep architecture. This complexity is why LassoNet relies on the implementation of a linear residual connection where input weights are penalized to encourage convergence towards zero by this approach, Fig. 5.2. In addition, in practice, introducing a L_1 penalty into gradient descent does not sparsify the weights and requires post-training thresholding, as LassoNet proposed.

In summary, L_1 penalization offers a relaxed approach to the problem that is suitable for linear models. However, when applied to DL models, it necessitates precise tuning to achieve a satisfactory approximation, and it does not drive the weights to zero at any point. Hence, in the subsequent section, an alternative probabilistic relaxation of the L_0 norm is proposed, specifically designed to work more effectively with DL models.

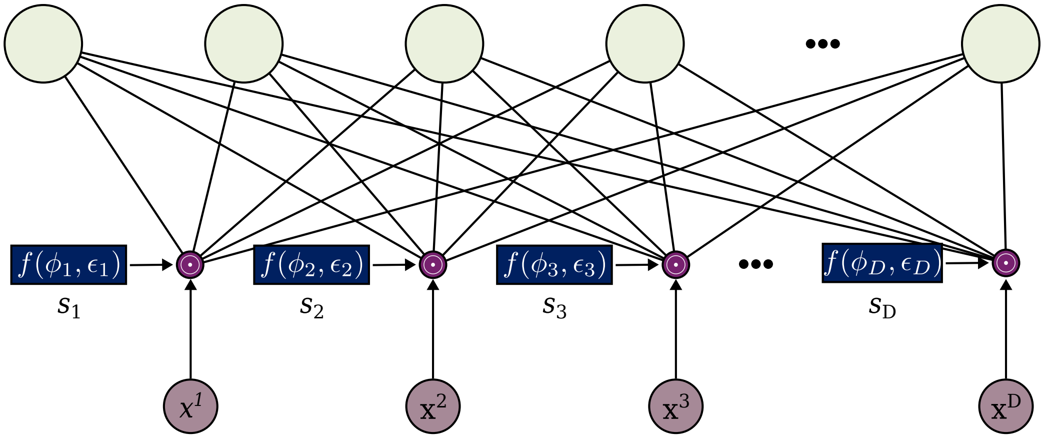


Figure 5.3: The proposed architecture uses the dropout technique for feature selection. Each feature is controlled by a 'gate,' with its probability of being opened or closed determined by the variational parameter ϕ_i, which represents the *dropout rate*. Features with a gate where $\phi_i \simeq 1$ are discarded, thereby reducing the number of features selected by the DL model.

5.2.1 Feature Selection by Dropout

The L_0 regularization is non-convex and computationally expensive and intractable for high-dimensional data [120] and it has been demonstrated to be a NP-Hard problem [247]. For this reason, it is necessary to generate a relaxation of this L_0 for obtaining an optimal solution to this ill-posed problem.

The Dropout technique [73] plays a fundamental role in proposing a probabilistic interpretation of the L_0 regularization for both linear and non-linear models. Its application for DNN has been applied previously [248, 249]. Intuitively, Dropout can be likened to a binary 'gate', determining whether a variable is active or 'dropped out' (see Section 2.2.4). Examining Eq. (2.19), the contribution of a variable is determined by a Bernoulli distribution with ρ_d as the dropout probability, known as the dropout rate. Each variable, as represented by a 'gate', follows a Binomial distribution, i.e., n variables following a Bernoulli distribution. However, when the probability for each variable is a constant ρ_d, it implies that each variable has an equal chance of activation, and the 'gates' are randomly opened. In other words, there is no a preference among the different variables.

In the context of FS, it is crucial to ensure that the 'gates' controlling the most pertinent features remain consistently open, while the less relevant ones are more likely to be closed. Therefore, it becomes imperative to assign a probability per 'gate', denoted as $p(s_i = 0) = \phi_i$ (a custom dropout rate per variable).

In this way, considering a Dropout technique as a feature selector method, the subset $X_s \subset X$ – with its indices defined by a vector $S \in \{0, 1\}^d$ where d represents the dimension of the input vector X – is defined by the individual probabilities ϕ_i:

$$\begin{aligned} S &\sim \prod_{i=1}^{d} Bern(\phi_i), \\ X_s &= X \odot S, \end{aligned} \tag{5.2}$$

where $\odot$ is the Hadamard element-wise product.

Loss function

In the context of applying the Dropout technique for FS, the process becomes a statistical model, necessitating a well-defined approach for drawing inferences from the model during optimal training. Just like any DL model, it is posed as a problem of minimizing a loss function, but with an added regularization-based constraint. As any DL model, the problem rephrased as a problem of loss function minimization but, in this case, there is a regularization-based constraint. Given a DNN $f(.;\theta)$, with θ as the model parameters, and considering the subset indices S, the problem can be defined as:

$$\begin{aligned} &\min_{\theta,\phi} \mathcal{L}(f(X \odot S; \theta), Y) \\ &\textbf{subject to} \quad J(S) \leq k. \end{aligned} \tag{5.3}$$

As it can be observed, the regularization factor affects exclusively to S. This $J(.)$ regularization aims to find a unique solution for the ill-posed problem (see Section 2.4.2). This regularization. Specifically, this regularization aims to derive a sparse representation of X to create the subset X_s while ensuring the number of elements remains lower than k. Essentially, the regularization factor works toward reducing the input vector X to contain k elements or fewer.

Based on the L_0 pseudo-norm being optimal for assessing sparseness in the vector X_s, the goal is to introduce a regularization factor that represents $J(S) \approx ||S||_0$. This factor relies on a Bernoulli distribution to select different variables, following the Dropout technique as seen in Eq. (5.2). Intuitively, the objective is to introduce more noise to those features that are considered less relevant. For instance, if the i^{th} element $x_i \in X$, is deemed irrelevant, the corresponding ϕ_i dropout rate should tend towards 1. A $\phi_i \simeq 1$ implies that this feature x_i is unlikely to be considered for problem-solving.

Following the aforementioned deduction, the regularization factor can be described by the Cumulative Distribution Function (CDF) [248]. A higher CDF, means that the dropout rate is low. For this reason, the regularization factor can be considered the expected value of the different Bernoulli distribution, defined by the dropout rate ϕ_i.

As a result, considering a supervised method, given an input X and output Y, the problem can be described as:

$$\min_{\theta,\phi} \mathbb{E}_{X,Y,S} \left[\mathcal{L}(f(X \odot S, Y; \theta)) + \lambda \Phi(S) \right], \tag{5.4}$$

where $\Phi(.)$ denotes the Bernoulli's CDF. The Lagrangian multiplier λ shares a relationship with the constraint k in Eq. (5.3). Increasing λ effectively decreases k. Hence, the regularization factor aims to boost the dropout rate, ϕ_i. The expectation $E_S[\Phi(S)]$ penalizes activated 'gates', s_i, as it increases them:

$$E_S[\Phi(S)] = \sum_{s \in S} s \cdot \Phi(s) = \sum_{i}^{d} s_i \cdot (1 - \phi_i). \tag{5.5}$$

The loss function detailed in Eq. (5.4) aims to resolve the problem by integrating a regularization factor that distinguishes the most significant features. Consequently, the dropout rate, ϕ_i, for the various features ϕ_i tends toward 0 for the most significant features and to 1 for the least relevant ones. Additionally, this constraint is essential for attaining a sparse solution, reducing the number of features in the subset X_s significantly compared to the original input features.

5.2.2 Bernoulli Continuous Relaxation

As it was previously described, the DNN are optimized via gradients descent for determining the optimal model parameters θ (see Section 2.2.2). For this reason, it is natural to think that the parameters for the Dropout-based FS, the different Bernoulli's dropout rate ϕ_i, have to be optimized by this method too. However, the discrete behavior of the Bernoulli distribution, will suffer high-variance, so it is not optimal to optimize by a gradient descent method.

To address this limitation, Maddison et al. [250] introduced a continuous relaxation of discrete distributions, a concept later applied in Concrete Dropout [251]. This technique provides a variational approach to optimize the dropout rate. For a given feature i of sample j, denoted as $x_{i,j} \in X$, the binary 'gate' defined by a Concrete binary distribution is formulated as follows:

$$\begin{aligned} s_{i,j} &= f_C(\phi_i, \epsilon_{i,j}) = \sigma\left(\frac{1}{\tau}(\log\phi_i - \log(1-\phi_i) + \log\epsilon_{i,j} - \log(1-\epsilon_{i,j}))\right), \\ \epsilon_{i,j} &\sim \mathcal{U}(0,1). \end{aligned} \tag{5.6}$$

In this context, $\sigma(.)$ represents a sigmoid function, and $\mathcal{U}(0,1)$ denotes a uniform distribution. The parameter $\tau \in (0,1]$ serves as a temperature, controlling the "noise level" in the generated samples [250]. Notably, as $\tau \to 0$, the approximation to the discrete Bernoulli distribution improves.

Leveraging this relaxation, Chang et al. [249] introduced a FS approach that involves penalizing the number of features not dropped out. Essentially, the authors suggested constraining the solution using a regularization factor derived from the sum of the elements of S. However, it is important to note that this approach lacks stability, and its results are highly dependent on the choice of the regularization factor.

In the continuous relaxation of the distribution mentioned earlier, the 'gate' is not strictly constrained to values of 0 or 1. Within a narrow range, defined by t and centered around ϕ_i, the Concrete distribution lets s to take on values between 0 and 1, rather than being strictly 0 or 1. It may be considered as spike and slab distribution [252].

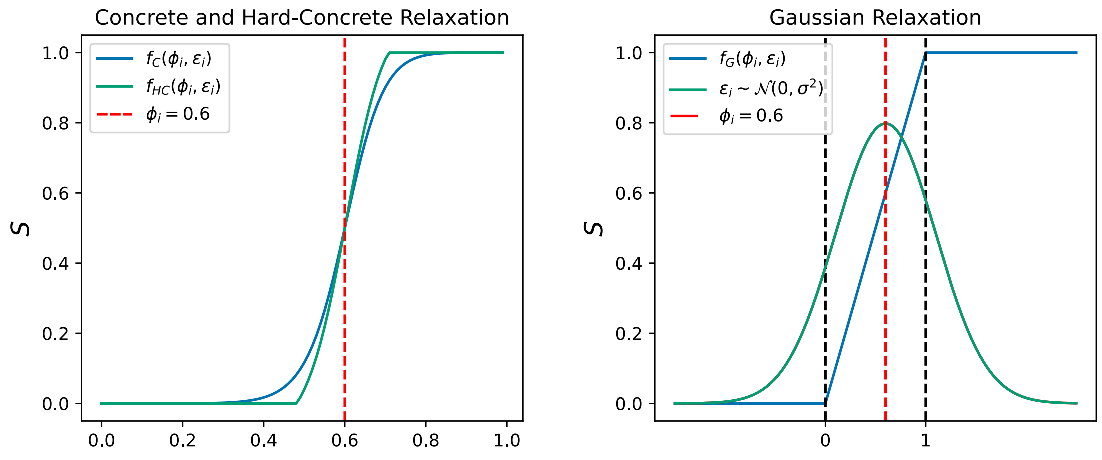
In order to mitigate the aforementioned limitation, Louizos et al. [248] proposed the hard-concrete distribution. This distribution is a modification of the Concrete distribution that forces the values to tend to 0 or 1.

$$\begin{aligned} s_{i,j} &= f_{HC}(\phi_i, \epsilon_{i,j}) = min(1, max(0, \overline{s}_{i,j})), \\ \overline{s}_{i,j} &= f_C(\phi_i, \epsilon_{i,j})(\zeta - \gamma) + \gamma \end{aligned} \tag{5.7}$$

with $\gamma < 0$ and $\zeta > 1$. As it is expressed, the Hard-Concrete distribution 'stretches' the Concrete distribution to the (γ, ζ) interval and the apply a hard-sigmoid.

An example of the proposed Hard-Concrete distribution is illustrated in Fig. 5.4a. As observed in this figure, the Hard-Concrete, $f_{HC}(.)$, has a steeper slope than the Concrete approach, $f_C(.)$. It is intuitive to understand that this function forces the value to 0 or 1.

Additionally, Yamada et al. [253] proposed a Gaussian-based continuous relaxation of the Bernoulli distribution. In this approach, the idea is to estimate the relaxation of the Bernoulli distribution by a normal distribution $\mathcal{N}(\phi_i, \sigma^2)$. To compute the gradients effectively, it is applied the Reparameterization Trick [254] in this normal distribution. As a result, this Gaussian-based



(a)

(b)

Figure 5.4: Different approaches for Bernoulli continuous relaxation. Both figures illustrate an example where the dropout rate ϕ_i is 0.6. In (a), the temperature t has been set to 0.2 and, for the Hard-Concrete distribution, $f_{HC}(.)$, the γ, and ζ has been set to -0.1 and 1.1 respectively. In (b), the Gaussian relaxation, $f_G(.)$, the σ^2 is set to 0.5.

relaxation is defined as follows:

$$\begin{aligned} s_{i,j} &= f_G(\phi_i, \epsilon_{i,j}) = min(1, max(0, \phi_i, +\epsilon_{i,j})), \\ \epsilon_{i,j} &\sim \mathcal{N}(0, \sigma^2) \end{aligned} \tag{5.8}$$

where the relaxation depends on the σ^2 instead of a temperature t as in the previous approaches.

In the context of the Gaussian-based relaxation, it is important to note that the estimation of the CDF for solving Eq. (5.4) has to consider σ. In this case, the estimation is obtained by $\Phi_G(\phi_i/\sigma)$, where $\Phi_G(.)$ is the standard Gaussian CDF [253].

As a conclusion, in this section, various relaxations have been introduced to obtain a continuous relaxation of the discrete Bernoulli distribution. The alternatives to the binary Concrete distributions [250] have been proposed to force the values to 0 and 1. However, as described in the following sections, this thesis proposes a training procedure to achieve this sparse representation in the input X.

5.3 Information Theoretical Perspective for Feature Selection

Leveraging IT concepts for FS is a logical choice, given that the goal is to identify a subset that provides the best compression of information. It is specially relevant since MI is measure of redundancy satisfies the following properties: it is symmetric, nonlinear, non-negative, and does not diminish when adding new features [255]. A popular algorithm is the previously mentioned mRMR [243]. There are more alternatives, as the Conditional Infomax Feature Extraction (CIFE) [256] or the Conditional Mutual Information Maximization [257, 255], which are a wrapper and filter-based method respectively.

In the previous section, it has been introduced a statistical approach for introducing the FS in DL models based on dropout technique. This section demonstrates the IT perspective of the proposed approach for FS. At it will be shown, this perspective has similarity with the mRMR and CIFE and how this approach based on dropout technique may be considered for wrapper methods, although it is an embedded method because the feature selector is followed by a DNN.

In addition, aligning with the objectives outlined in Chapter 4, this thesis aims to gain a more in-depth understanding of the problem-solving mechanisms of DNN through the lens of information-theoretical concepts. By adopting this approach, it becomes more feasible to enhance the interpretability of the model. The information-theoretical perspective allows for the description of DL models based on their information flow, elucidating the compression achieved, and the information lost across different layers. Both concepts, information and compression, are intuitive and contribute to obtaining a clearer perspective on these complex models.

In the previous chapter, it was established that DNN excel at data compression. Moreover, in the context of supervised tasks, the inherent trade-off between compression and predictability compels the model to achieve a compression that retains features relevant for solving the supervised task. This section seeks to extend the information-theoretical perspective to the realm of FS tasks. Finally, the insights gained from the IT perspective will be applied to ensure the method is configured for obtaining the optimal solution.

5.3.1 Feature Selection by Dropout: Information-Theoretic Learning

As it was previously mentioned, the FS may be considered a lossy compression, it consists of obtaining a subset X_s (a compression of X) in which it may be considered that there is a loss of information. In IT, it may be measured by MI. As a result, the compressed representation X_s has to maximize the MI regarding a task defined by a random variable Y. In other words, X_s has to be compressed with that features that contain most information of Y, any other information is irrelevant. Based on a IT-learning approach, the FS optimization may be formulated as [253]:

$$\max_{S} \mathcal{I}(X_s; Y) \quad s.t. \quad |S| = k. \tag{5.9}$$

In this way, the FS is a task as selecting S, which is hypothesized that contains k relevant features represented by the cardinality of the set ($|S|$), such that the MI between X_s and Y is maximized.

In the proposed FS method based on the dropout technique (Section 5.2.1), deterministic optimization is not applied. The method is inherently stochastic, employing a binomial distribution as described in Eq. (5.2). Given an input vector X with d components, the stochastic regularization factor $\mathcal{R}(S)$ might be defined as $\mathcal{R}(S) = E[S] \leq k/d$, but it is not relevant because it is out of the scope of the present study. Additionally, the $\mathcal{I}(X_s; Y)$, as developed in Appendix A.2.1, is expressed as follows:

$$\mathcal{I}(X \odot S; Y) = E_S \left[\mathcal{I}(X, Y) \right]. \tag{5.10}$$

Therefore, the defined FS approach based on dropout, from an IT-learning perspective, is as follows:

$$\operatorname*{argmax}_{\phi} E_S \left[\mathcal{I}(X, Y) \right] \quad s.t. \quad \mathcal{R}(S). \tag{5.11}$$

In conclusion, the goal is to obtain the parameters, i.e., the different dropout rates $\phi \in \{\phi_1, ..., \phi_d\}$ that define S as depicted in Eq. (5.2), maximizing this expression. As deduced, this approach follows the InfoMax principle described in Eq. (3.23) (See Section 3.5).

Given that FS involves a lossy compression, it is guaranteed that $\mathcal{I}(X;Y) \geq \mathcal{I}(X_s;Y)$. The optimal scenario is to attain the Markov Blanket [255], which is the subset containing the full information necessary for inferring Y. However, in practical terms, even if there is minimal loss of information, X_s is still considered a Markov Blanket. This perspective arises from the understanding that the original structure might not be apparent for establishing a relationship with Y. Consequently, a transformation of X_s defined by a function $f(.)$ can achieve the maximum MI, denoted as $\mathcal{I}(f(X_s);Y) = \mathcal{I}(X;Y)$.

Reformulating as a Minimization Problem

In the preceding section, it has been outlined the approach of FS based on Dropout, emphasizing the maximization of the expected MI between X and Y across the 'gates'. However, in this section, it will approach the same problem from a minimization perspective, offering, in my opinion, a more intuitive understanding of the optimization process.

Following the definition of MI in Eq. (3.10), it becomes apparent that maximizing MI over S can be reframed as minimizing the conditional entropy $\mathcal{H}(Y|X_s)$, given that $\mathcal{H}(Y)$ is constant. In essence, the optimal solution aims to establish a 'noise-free' channel (Section 3.2.1), requiring the noise, quantified by the conditional entropy, to approach zero ($\mathcal{H}(Y|X_s) = 0$):

$$\max_S \mathcal{I}(X_s;Y) \Leftrightarrow \min_S \mathcal{H}(Y|X_s).$$

In Appendix A.2.2, it has been developed how the conditional entropy for FS based on dropout technique should be minimized. In this way, the conditional entropy, $\mathcal{H}(X_s|Y) = \mathcal{H}(X, S|Y)$ corresponds to a minimization of the joint entropy $\mathcal{H}(X, Y, S)$.

In this way, the FS based on dropout technique is defined as follows:

$$\underset{\phi}{\operatorname{argmin}}\ \mathcal{H}(X,Y) + \mathcal{H}_\phi(\mathcal{S}) \quad s.t. \quad \mathcal{R}(S). \tag{5.12}$$

As it is indicated, the solution proposed consists in reducing the entropy $\mathcal{H}(S)$, the entropy of the 'gates'. The joint entropy $\mathcal{H}(X,Y)$ is not optimized, but it is indicated for the future mentions in the next sections.

Keys insights from the IT Perspective

In the context of this information-theoretic learning approach, several key insights can be derived. As per Eq. (5.12), the FS method based on Dropout operates by minimizing the uncertainty introduced by the 'gates' responsible for dropping out features from X. This provides valuable information about the optimal initialization of the dropout rates, ϕ_i.

Since the 'gates' are defined by the indices of the subset S where $s_i \sim \text{Bern}(\phi_i)$ (see Eq. (5.2)), the initial configuration must be determined by the distribution of S that maximizes entropy. The optimal initial value is found when $\phi_i = 0.5$. This is illustrated in Fig. 3.2, where it is evident that the entropy is 1 bit, the maximum entropy value, when $p = 0.5$.

Considering this problem as a minimization of $\mathcal{H}(S)$, as it is depicted in Eq. (5.12), implies the objective of reducing the uncertainty introduced by S, akin to mitigating the 'white noise' generated by uncertainty. It is imperative to note that the minimization of $\mathcal{H}(S)$ occurs in two scenarios: when $\phi_i \to 0$ and when $\phi_i \to 1$, as illustrated in Fig. 3.2. Therefore, the incorporation of the regularization factor is pivotal for ensuring a stable solution and establishing a well-posed problem (see Section 2.4.2).

It is intuitively evident that a classifier would seek to utilize all features provided by X, so implementing the optimization of Eq. (5.4) without accounting for the regularization factor would lead to dropout rates converging to 0. The regularization factor guarantees that $\mathcal{H}(S)$ is reduced, causing dropout rates to converge to 1 for most features, and to 0 for a select few — the relevant ones.

5.3.2 Deep Learning Perspective: Information Bottleneck

The FS based on dropout technique has been defined by an IT perspective in the previous section. This section extends this perspective to incorporate the DNN, as this process is carried out within the DNN. This extension aims to enhance the interpretability for balancing the interpretability-accuracy trade-off for the selection of the proposed model.

The standard analysis of lossy compression, such as the FS process, often involves rate distortion theory, as described in Eq. (3.20). A primary challenge in this context is the need to properly specify the distortion function, which determines the relevant features [160]. Rate distortion theory, however, does not provide a complete solution, as the choice of the distortion function is not inherently addressed within the theory. One approach to address this limitation is to consider the generalization of this theory, the IB [160].

As it has been validated in the Chapter 4, a supervised DL algorithm, as the proposed in this chapter, can be formulated as an information theoretic trade-off between compression and prediction. This property is formalized by the IB framework [97], see Section 4.1. In this way, the DNN has to detect the most relevant features that guarantee the predictability of Y. Additionally, due to DL architectures, typically forming a Markov chain, it has to considered that the information flow in the DNN follows the DPI, as it described in Eq. (4.1) and Eq. (4.2).

Following the IB, the FS based on dropout technique, described in Section 5.2.1, forces to identify the 'bottleneck' in the 'gates'. In other words, the bottleneck Z in Eq. (3.22) is the subset X_s defined as Eq. (5.2). In this way, the optimal solution is given by the following equation:

$$\min_{p(x_s|x)} \mathcal{I}(X;X_s) + \beta\mathcal{I}(X_s;Y)\,, \tag{5.13}$$

which involves compressing the information on X while retaining the least necessary information about Y.

As you may have noticed, it may seem contradictory to what was presented in Eq. (5.9) since minimizing the second term, controlled by the Lagrangian multiplier β, appears to involve reducing the MI with respect to Y. However, there is an explanation for this, considering the fundamental role of data structure in information, as mentioned in Section 3.1.

The estimation of $\mathcal{I}(X_s;Y)$ quantifies the information shared between both variables. Since $\mathcal{I}(X_s;Y)$ is not equivalent to $\mathcal{H}(Y)$, it implies that $\mathcal{H}(Y|X_s) \neq 0$, as shown in Eq. (3.10). In this context, it can be inferred that X_s represents a compressed form of X that has introduced additional uncertainty in identifying Y, as indicated by the conditional entropy $\mathcal{H}(Y|X_s)$. However, if X_s contains sufficient information ($\mathcal{H}(X_s) > \mathcal{H}(Y)$), the features in this subset can be transformed into a new latent space where their relationship with Y becomes more evident, leading to a reduction in $\mathcal{H}(Y|X_s)$.

In general, any DL model with various hidden layers, defined by the model parameters θ, aims to reorganize and transform the information within X to map to $p_\theta(Y|X)$. The multiple layers work to transform the information, identifying features that are more salient for predicting Y, as illustrated in Eq. (4.2). From an IT perspective, this process can be considered a reduction in the entropy of the model, denoted by $\mathcal{H}_\theta(X, Y)$

In the previous section, it was introduced how the FS method by dropout technique may be formulated as a minimization of the joint entropy $\mathcal{H}(X, Y, S)$, where S is the set of indices which identifies X_s. Following the Eq. (5.12), the $\mathcal{H}(X, Y)$ can be replaced by $\mathcal{H}_\theta(X_s, Y)$ since the DNN has to minimize this expression. The use of the X_s in the expression $\mathcal{H}_\theta(X_s, Y)$ is justified by the architecture, see Fig. 5.3, since the hidden layers directly operate on the subset X_s defined by the indices S. As a result, the FS method by dropout technique is defined as:

$$\underset{\theta,\phi}{\operatorname{argmin}}\ \mathcal{H}_\theta(X_s, Y) + \mathcal{H}_\phi(\mathcal{S}) \quad s.t. \quad \mathcal{R}(S). \tag{5.14}$$

In conclusion, the bottleneck defined by Eq. (5.13) represents the minimal information that has to be propagated to the DNN for solving the problem defined by $p(X, Y)$. In an ideal case, the compressed representation X_s would have $\mathcal{I}(X_s; Y) = \mathcal{H}(Y) = \mathcal{H}(X_s)$. It means that X_s contains the minimal information possible and is sufficient for predicting Y. In this way, the information propagation in the DNN would be considered as a noise-free channel (see Section 3.2), and the DPI would be satisfied by the special case represented by the equality in Eq. (4.1). In other words, there would be no loss of information in the information flow through the different layers.

Key Algorithmic Details

This section introduces key insights that can be derived from the IT perspective. These insights are crucial for establishing the optimal configuration of the DNN with the FS method.

As it is observed in Eq. (5.14), the model has to solve two problems: $\mathcal{H}_\phi(S)$ that consists in solving the FS process and $\mathcal{H}_\theta(X_s, Y)$ that is the typical solving problem of a supervised problem. Each problem is well-defined by fitting the parameters θ (parameters of the hidden layers in the DNN model) and the parameters ϕ, which consists on the FS method depicted in Eq. (5.2). However, following the expression of the joint entropy, Eq. (3.6), it depends on $\mathcal{H}(X_s)$ that depends on S so it is deduced that $\mathcal{H}(X_s) \sim \mathcal{H}(S)$. It may represent a potential problem to converge the DNN to an optimal solution.

Considering that $\mathcal{H}(S)$ is constantly changing, specially in the initial epochs of training where the variational parameters ϕ_i (dropout rates) are initialized to have the highest entropy value, as proposed in Section 5.3.1, the first epochs can be challenging for the model. To address this, the use of the annealing trick [258], originally proposed to improve the training of the Variational AutoEncoder (VAE) [254], is proposed. This involves applying a scheduler that gradually increases the Lagrangian multiplier of the regularization factor, λ in Eq. (5.4), during the training process.

By doing so, the model in the initial epochs focuses on resolving $\mathcal{H}_\theta(X_s, Y)$, and after a fitting stage, it gradually shifts towards minimizing $\mathcal{H}_\phi(S)$. More details about the scheduler proposed in Section 5.4.1.

Following the IB, another key factor to consider is that the architecture directly influences the number of features selected. As observed in Eq. (5.13), the bottleneck lies in the FS method, and, as justified, it defines the minimal information that needs to be propagated to the hidden layers of the DNN, which aims to reorganize and transform the information within X to solve the problem. Consequently, a deeper DNN might require less common information for predicting Y from X_s, i.e., a lower $\mathcal{I}(X_s; Y)$.

From this observation, it is deduced that using a deeper DNN may pose a challenge in terms of interpretation with respect to the original variables. While a deeper DNN results in a more sparse representation of the input X, the relationship between X_s and Y becomes more difficult to interpret. Therefore, the network architecture must be carefully considered when applying this method.

5.4 HyperSpectral Band Selection Case

Nowadays, characterized by an ever-growing volume of data, the application of FS methods becomes crucial to tackle the challenges associated with data analysis. This issue is particularly evident in HSI, where the continuous increase in data volume poses new challenges, leading to a significant burden on manual labor and material resources. Addressing these challenges requires effective strategies to reduce this burden, enhance efficiency, and extract meaningful insights from the data [259].

In the medical domain, the use of HSI is a commonly explored field due to its substantial diagnostic and surgical guidance potential (see Section 2.6). Nevertheless, the extensive information furnished by HSI often comprises redundancy and non-essential data. It is vital to discern the most pertinent wavelengths for a specific application to enhance prediction accuracy and reduce classification algorithm execution time. Moreover, eliminating noisy wavelength bands can further enhance the classification process [260].

In this section, it is assessed the effectiveness of FS by DL in the context of HSI, considering its challenging high-dimensional feature space. The primary objective is to identify the subset of wavelengths that provide crucial information, identifying the Markov Blanket, essential for resolving the supervised problem. Our experiment focuses on the application of HSI for the identification of biological tissues. To ensure robust validation, the obtained results will be compared with information available in the state-of-the-art.

5.4.1 Experimental Setup

In HSI, the information is captured and represented by a data structure consisting of a series of spectral bands or channels, with each band or channel capturing information about the light reflected, emitted, or transmitted by objects in the environment over a narrow range of wavelengths. Let the hyperspectral image of spatial dimensions $W \times H$ with L spectral bands be denoted by $X \in \mathbb{R}^{W \times H \times L}$. However, in this experiment, it is proposed to find the spectral band more relevant for identifying the variable Y. For this reason, the hyperspectral image is reshaped to produce the matrix $X \in \mathbb{R}^{N \times L}$ where $N = WH$ is the number of pixels in the image.

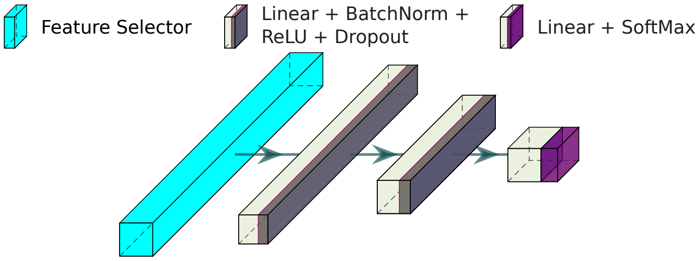


Figure 5.5: DNN architecture proposed for experimental results. The feature selector is based on the FS method based on dropout technique, which has the purpose of identifying the subset X_s.

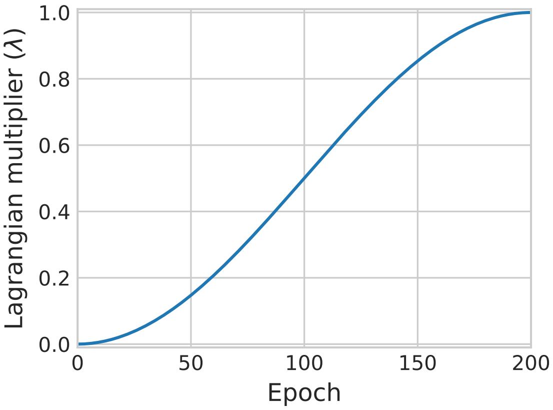


Figure 5.6: Proposed scheduler for applying the annealing trick during the training process of 200 epochs.

For the purpose of this study, a DNN is employed, featuring an architecture consisting of various fully-connected layers, as illustrated in Fig. 5.5. Notably, the proposed architecture for the experimental results is intentionally not excessively deep. This decision, as discussed in Section 5.3.2, is made to preserve the interpretability of the data, preventing the generation of a subset X_s where the relationship with Y becomes less evident. The number of neurons in the hidden layers is adjusted for each dataset. Furthermore, a dropout rate of 0.5 has been applied to the hidden layers to mitigate overfitting and encourage the model to achieve a higher compression in the latent space, as demonstrated in Section 4.3.2. It is important to note that the variational parameters of the feature selector are set to $\phi_i = 0.5$, as this has been identified as the optimal initialization value, as demonstrated in Section 5.3.1.

Annealing Trick

As it was discussed in Section 5.3.2, we recommend applying this technique to mitigate the problems associated with the entropy of the subset X_s. This technique is previously proposed for training VAEs [258] by increasing linearly the regularization factor. However, in this thesis it is proposed a scheduler based on cosine function. This approach of scheduler are previously proposed for applying the annealing trick to the learning rate parameter [261] and DPM [262].

The cosine scheduler is defined as follows:

$$s(t;T) = 1 - cos\left(\frac{t}{T}\cdot\frac{\pi}{2}\right)^2, \tag{5.15}$$

where T is the number of total epochs of the training and t is the current epoch. A visualization of this scheduler is illustrated in Fig. 5.6.

Examining Fig. 5.6, where the scheduler is defined for training a DNN for 200 epochs, the regularization factor, intended to emphasize sparseness or reduce $\mathcal{H}_\phi(S)$, is minimal in the initial epochs. Around epoch 50, the regularization factor gradually increases, following an approximately linear trend until around epoch 150. In the final 50 epochs, it slowly converges to the maximum value.

This scheduler is particularly relevant in the initial steps, as the model initially focuses on solving the classification problem. The FS process is considered in the subsequent epochs, thereby increasing the importance of this function. As the DNN tackles the problem in an environment with maximum uncertainty, defined by the different ϕ_i values and the dropout rates of the hidden layers, once the model generalizes effectively, the reduction of uncertainty in S would not negatively affect the model.

Training process

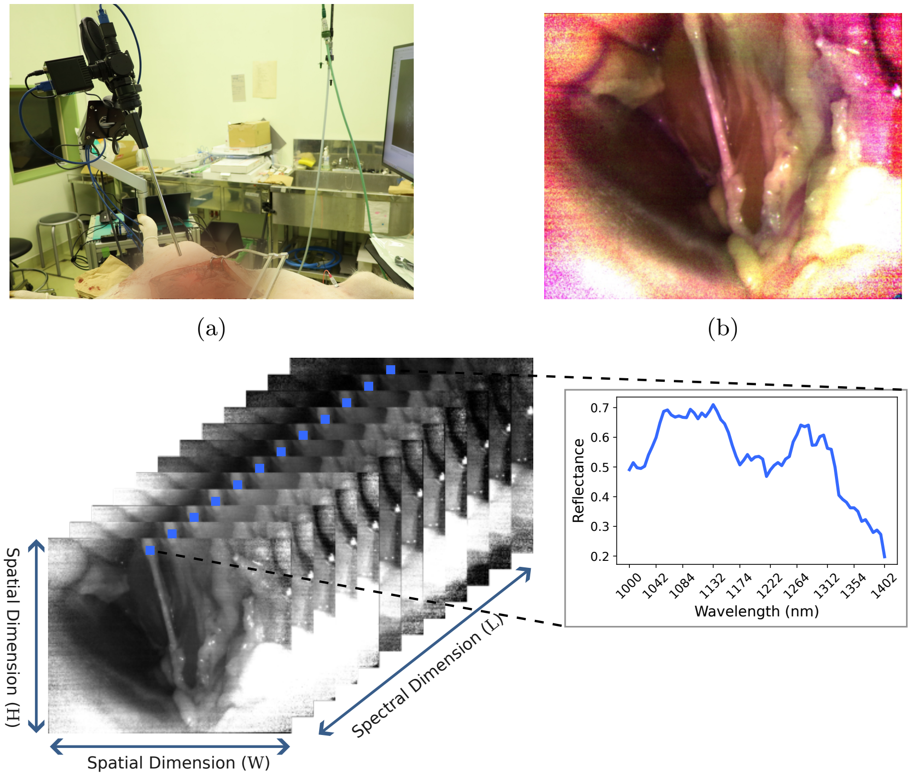
To ensure consistency across various experiments, common practices have been applied. The experiments are trained for a total of 200 epochs. The batch size is adjusted for each dataset to ensure 100 iterations per epoch, resulting in a total of 20000 iterations throughout the training process. The chosen optimizer is ADAM [65] with a learning rate set to 10^{-3}.

Additionally, the dataset is divided into a training set and a test set to validate the training process and ensure that overfitting is avoided. The test set comprises a randomly selected subset, corresponding to 20% of the dataset. For the sake of reproducibility, a fixed seed has been applied across different experiments, ensuring that the subset of samples for a specific dataset remains the same.

5.4.2 HyperSpectral Laparascopy Dataset

HyperSpectral Imaging plays a crucial role in providing non-destructive analysis of object components that are challenging to distinguish with visible light (see Section 2.6). Examples include identifying fat in meat, distinguishing between similarly colored resins, and detecting foreign matter contamination [263], especially in the NIR region (850-2500 nm). At the Tokyo University of Science, a HSI rigid scope system has been developed for medical applications [263]. Indeed, this device functions as a HSI laparoscopy device, making it suitable for minimally invasive surgery as it can be introduced through a small incision to operate on internal organs in the abdomen or pelvis. Such a device holds the potential to enhance current trends in this technology, including robot-assisted surgery powered by AI or augmented reality [264].

However, the current acquisition process for the system is time-consuming, requiring modifications to the light source and an acousto-optic tunable filter [263]. Consequently, the acquisition system takes minutes to obtain an image, as camera settings, optical design, and the number of wavelengths have not been optimized to prioritize imaging speed.



(a)

(b)

(c)

Figure 5.7: Data acquisition with the HSI rigid scope system. In (a), the scenario for data acquisition is illustrated. In (b) displays a synthetic RGB image generated from the NIR, and (c) illustrates the data structure of hyperspectral imaging. Here, W and H correspond to the spatial resolution of 1280×1024, and L represents the spectral dimension with 68 bands.

The proposed FS method aims to optimize the number of wavelengths required, reducing this number, and thereby improving acquisition time. Additionally, this approach is relevant for enhancing data interpretability by reducing the number of wavelengths for analysis.

In this section, a hyperspectral image obtained during a research study is utilized to identify relevant features for distinguishing different tissues, including fat, muscle, nerve, and the hypervascularized zone. Figure 5.7 illustrates the data acquired in-vivo from a porcine. In Fig. 5.7a, the scenario where the image was acquired is depicted, showing the introduction of the HSI rigid scope into the subject. An RGB representation derived from the hyperspectral image is observed in Fig. 5.7b.

Concerning the data structure employed in this experiment, as depicted in Fig. 5.7c, the hyperspectral image has spatial dimensions $W \times H$ (corresponding to 1280×1026) with L spectral bands (68 bands) denoted by $X \in \mathbb{R}^{W \times H \times L}$. However, as the objective of this work is to identify the most relevant spectral bands, the input X has been reshaped to consider each pixel independently, i.e., $X \in \mathbb{R}^{N \times L}$ where $N = W \cdot H$. For this reason, only the spectral information is considered for classifying each pixel, which goes from 1000 nm to 1402 nm.

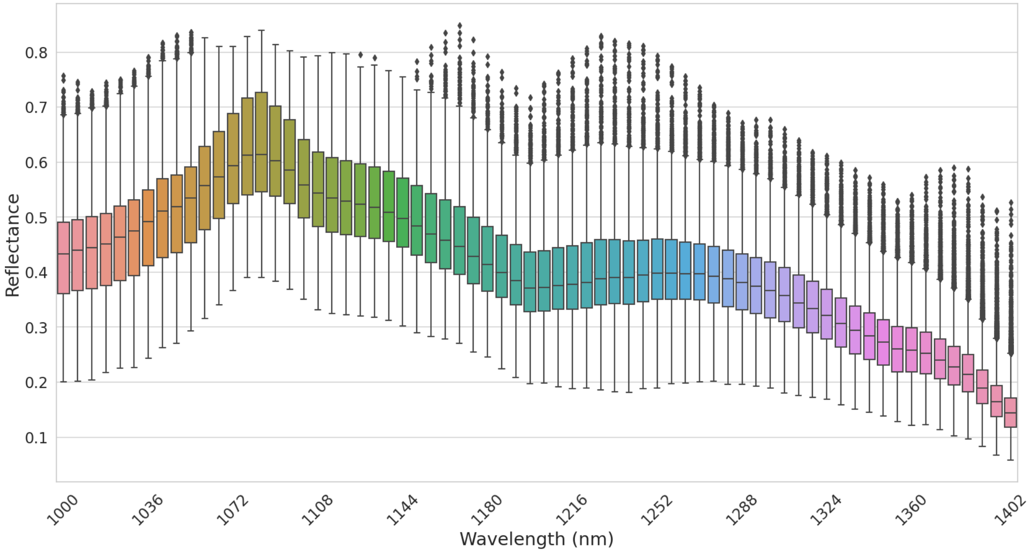


Figure 5.8: Spectral data exploration of the dataset using boxplots. The small dots outside the minimum/maximum values represent the outliers. The box boundaries represent the IQR.

Table 5.1: Laparoscopy Dataset Summary. The label 'vessel' represents the hypervascularized zone.

Dataset	Fat	Muscle	Nerve	Vessel
Original	199674	230971	27991	170854
Undersampled	10000	10000	10000	10000

Preprocessing

In the acquisition process, four hyperspectral images were obtained from three different angles. The different images were unified, and the hyperspectral signatures have been reduced by identifying outliers using the well-known Interquartile Range (IQR) methodology [142]. In this way, those pixels that contain an outlier in a one or more bands are removed from the dataset. Figure 5.8 shows the result of this methodology.

The dataset comprises more than $600,000$ samples, which correspond to the pixels of the different images. Table 5.1 depicts the number of samples per class that you can find in the dataset. To mitigate the high imbalance among samples for different classes, particularly focusing on the identification of nerves, an undersampling technique was employed to balance the dataset.

Three different methods were evaluated for balancing the dataset, the NearMiss method [265], a second method based on K-Means [260] and a method based on Orthogonal Complements Subspace Projection (OCSP) [266]. The downsampling method has the purpose of reducing the number of samples to 10000 per class, as it is shown in Table 5.1, having a dataset composed of the most relevant 40000 samples. For the evaluation of the different methods, it was considered a quantitative evaluation based on cross-validation using a SVM for classification and a qualitative result using Potential of Heat-diffusion for Affinity-based Trajectory Embedding (PHATE) [267] for dimensional reduction. The use of PHATE allows to obtain a representation that preserves local structure from the high-dimensional data without losing the global structure.

As a result, the OCSP method was used for carrying out the downsampling. The reason why the OCSP was selected is due to that this method allows obtaining more variability in the subset of samples extracted. More details can be found in Appendix A.2.3.

5.4.3 Feature Selector Configuration: Experimental Evaluation

The FS method based on dropout technique has some hyperparameters that have to be considered for obtaining an optimal solution. The first of them, is the Lagrangian multiplier (λ) of the loss function defined by the Eq. (5.4). As it was mentioned in Section 5.2.1, λ shares a direct relationship with the sparseness of the X. For this reason, it has to be defined as a hyperparameter of the algorithm that will indicate the maximum value of the penalization factor, which is controlled by the cosine scheduler defined in Eq. (5.15). In this way, the regularization factor in an epoch t is defined as follows:

$$\lambda_t = \lambda \cdot s(t;T). \tag{5.16}$$

Regarding the configuration of the FS method, it is necessary to take into consideration different hyperparameters depending on the Bernoulli relaxation (see Section 5.2.2). In this section, the Concrete approach, see Eq. (5.6), and the Gaussian approach defined by Eq. (5.8) have been evaluated.

The feature selector using the Gaussian approach has a 'locality' feature in the estimation because of the use of the Reparameterization Trick, as it was indicated in Section 5.2.2. As illustrated in Fig. 5.4b, the ϵ is centered on the dropout rate ϕ, and for this reason, the σ of the Gaussian distribution, which is a hyperparameter, is important. A small σ means that the dropout technique is not well implemented, and an extremely high σ would reduce the option to a binary case, so the continuous relaxation is not reflected. As the range is from 0 to 1, the optimal value is using a σ of 0.5, as it is illustrated in Fig. 5.4b. Yamada et al. [253] provide a more in-depth analysis of the σ hyperparameter tuning, and in conclusion, it was indicated that the optimal value of σ is 0.5.

The Concrete approach for FS is not so intuitive, and it may depend on the dataset. Contrary to the previous case, the ϵ is defined by a uniform distribution (see Eq. (5.6)) so the effect of the σ in the previous Gaussian case can be associated to the τ for the Concrete approach. However, it is not intuitive and, for this reason, different configuration would be considered. In the experimental results, it will be evaluated different values of τ as 0.2, 0.3 and 0.5.

Experimental Result

As detailed in Section 5.4.1, the DNN architecture proposed is based on Fig. 5.5. In this case, it is used fully-connected layers of 68 (the number of bands in the hyperspectral image as it is indicated in Section 5.4.2), 100, 32, and 4 neurons. Remember, considering the output layer's role in tissue identification, it is essential to tailor it to the specific classes of interest: fat, muscle, nerve, and the hypervascularized zone labeled as 'vessel'. The role of the feature selector is to eliminate irrelevant features through the dropout technique.

Figure 5.9 illustrates bands selected by different feature selector configurations. When varying the regularization factor (λ in Eq. (5.4)) from left to right, a notable increase in sparseness is observed in Fig. 5.9 from left to right. The sparsity converges to approximately 81% with various implementations for $\lambda = 2$. This observation underscores the relationship between λ and the constraint k in Eq. (5.3). An additional test was carried out using a regularization factor of 3, observing

Table 5.2: Normalized sparsity of the input achieved through various FS methods under different regularization factors.

FS Approach		**Sparsity** λ **=1**	**1.5**	**2**	**3**
Concrete	$\tau = 0.2$	0.66	0.71	0.81	0.84
	$\tau = 0.3$	0.66	0.74	0.81	0.85
	$\tau = 0.5$	0.68	0.75	**0.82**	**0.87**
Gaussian	$\sigma = 0.5$	**0.72**	**0.78**	0.81	0.81

the consistent pattern discarding the Gaussian approach, which maintaining sparsity at 81% (see Table 5.2).

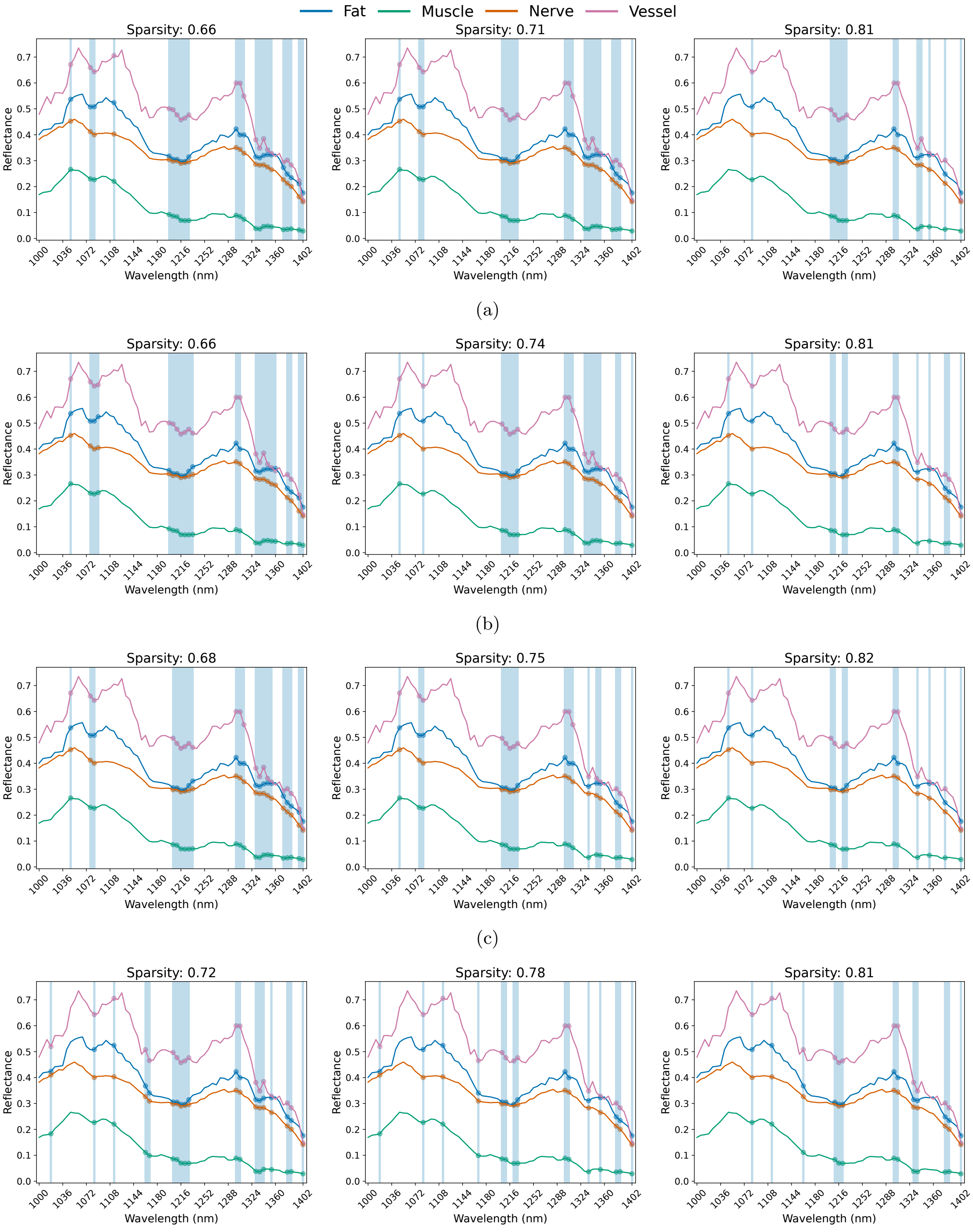
Concerning the different implementations, the Concrete approach exhibits consistent behavior across different experiments, with modifications to the hyperparameter τ, as illustrated in Fig. 5.9a, Fig. 5.9b, and Fig. 5.9c. In general, the Concrete approach for FS identifies distinct regions in the wavelength, grouping various bands of the signal. As λ increases, these clusters consistently reduce, resulting in a subset of bands in each region. For instance, observing Fig. 5.9b, in the region approximately between 1213 and 1220, using $\lambda = 1$ selects 6 bands, and increasing to 1.5 and 2 the regularization, this number is reduced to 5 and 4, respectively.

Certainly, as it has been discussed in this chapter, the application of FS methods is instrumental in improving the interpretability of high-dimensional data, as the HSI. In the context of tissue identification, state-of-the-art literature indicates that the distinctive absorption characteristics of fat tissue can be observed in the spectral range around 1200 and 1400 nm, marked by a peak in lipid absorption [268, 269]. This knowledge is particularly relevant for muscle identification, given the absence of lipids in muscle tissue [270], especially around 1200 nm, where the absorption coefficients of water (abundant in muscle) and lipids are comparable [270]. In addition, around 1300 nm there is clear evidence for discriminating between water and lipids [270, 268], which a range of the wavelength that it is considered by the FS method.

The identification of nerve tissue is closely tied to the characteristics of the myelin sheath surrounding the nerve axon, which, in turn, is enveloped by collagen. Collagens, crucial components of Schwann cells, play a vital role in peripheral nerves [271], such as those captured in this dataset. In existing literature, a soft peak of absorbance around 1100 nm has been identified [268], potentially justifying the consistent selection of the band corresponding to 1084 nm by the FS method. Additionally, there is another peak around 1500 nm [268], possibly explaining the selection of the band corresponding to 1402 nm in various implementations. However, it is worth noting that the 1200 nm range is also significant for collagen.

The identification of the hypervascularized zone, labeled as a ‘vessel’ in the dataset, poses a challenge in justifying the selection in the range of 1000 to 1400 nm. Upon reviewing the existing literature, it appears that there is no specific band identified as relevant in this range. However, a potential justification might be linked to the identification of cholesterol in the blood, leveraging the bands crucial for identifying lipids, and considering the high composition of water.

Finally, Fig. 5.10 illustrates the classification result of an image acquired from a different porcine, distinct from the one shown in Fig. 5.7. This image lacks labeled samples and has been utilized to qualitatively assess the model’s performance using the bands selected by the proposed approach.



(a)

(b)

(c)

(d)

Figure 5.9: HSI band selection using the FS based on dropout technique with different configurations. The Concrete approach is evaluated with different τ, 0.2 (a), 0.3 (b) and 0.5 (c). The Gaussian approach (d) is evaluated using the optimal hyperparameter setting $\sigma = 0.5$. From left to right, there are applied three different regularization factors, 1, 1.5, and 2, respectively.

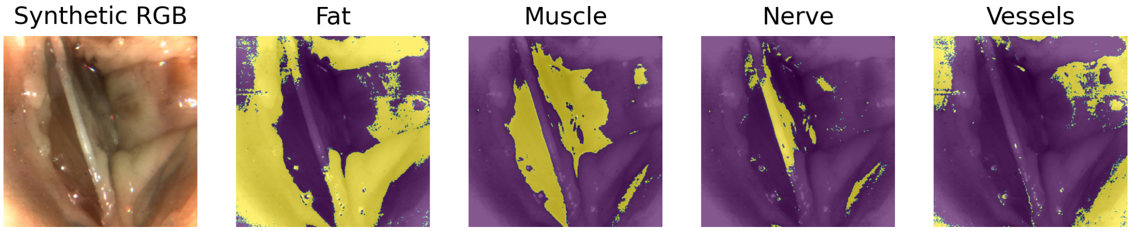
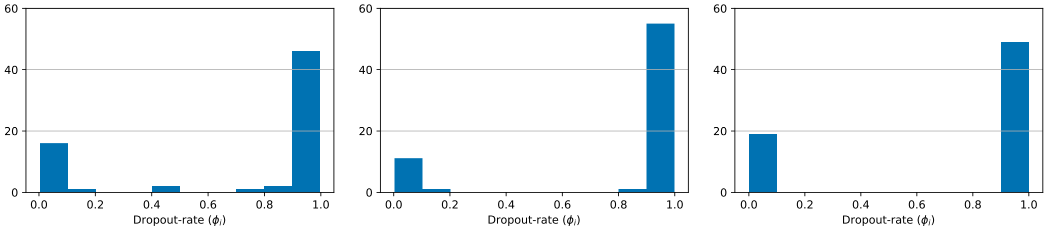


Figure 5.10: Classification of a hyperspectral image obtained from a second porcine that does not contain labeled samples.



(a)

(b)

(c)

Figure 5.11: Histogram depicting the dropout rate values in various configurations of the FS method based on dropout technique. In (a) and (b), the histogram illustrates the Concrete approach with $\tau = 0.3$ and λ set to 1 and 2, respectively. In (c), the Gaussian approach is employed with λ set to 1.

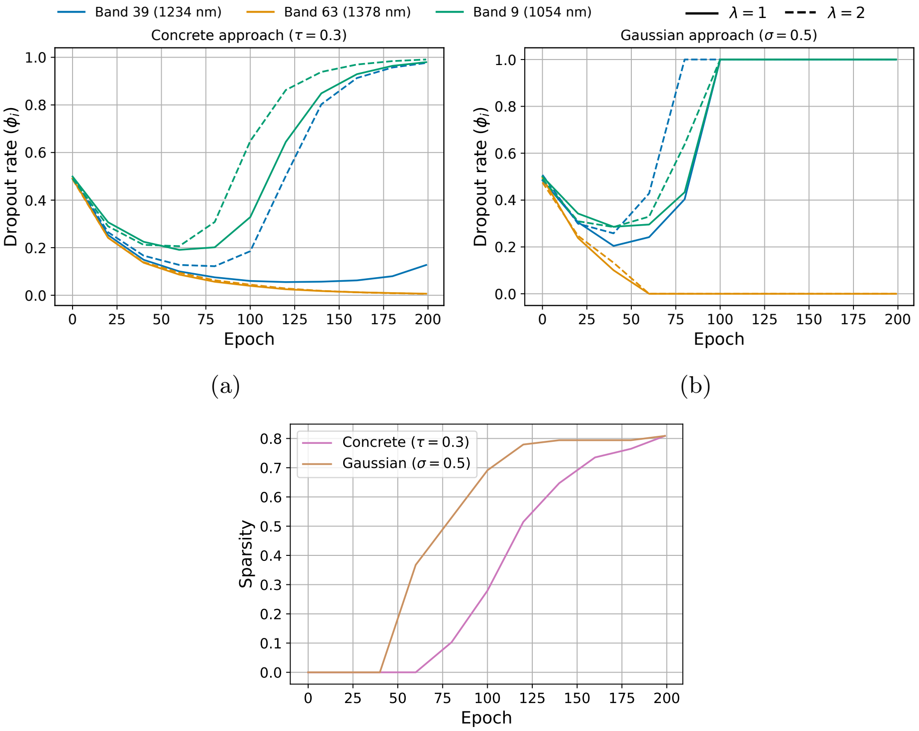
Specifically, the bands selected by the Concrete dropout with a regularization factor of 2 were employed (see Fig. 5.9b). Despite the challenge of identifying different tissues in the synthetic RGB representation, the model successfully differentiated between various tissues, as evidenced by the classification result for the 'Nerve' tissue shown in Fig. 5.10.

Dropout rate (ϕ_i) Progression

Another crucial factor to consider is the value of ϕ in different implementations, which it will be discussed in this section. As explained in Section 5.2.2, the Gaussian approach was introduced to address the problem associated with the high variance in the Concrete and Hard-Concrete approaches due to their heavy-tailed nature [253]. In other words, the Concrete approach faces challenges in obtaining a precise $\phi_i = 1$ and may converge to a value close to $\phi_i \simeq 1$, but achieving an exact 1 is much more complex. The Hard-Concrete approach attempts to mitigate this problem [248], but it is still present. This problem in the Concrete approach is evident in Fig. 5.11.

Examining Fig. 5.11a, it is evident that there are ϕ_i values that are not located near 0 or 1 when the Concrete approach is applied with τ set to 0.3 and λ set to 1. This issue is less noticeable when using λ set to 2, as shown in Fig. 5.11b. However, there are still some features that cannot be considered 0 or 1. In contrast, the Gaussian approach with λ set to 1 ensures that all ϕ_i values are located at 0 or 1, as depicted in Fig. 5.11c.

Figure 5.12 illustrates the dropout rate progression during the training process. Examining Fig. 5.12a, it is evident that ϕ_i values converge to 1 for some features but still cannot reach exactly one. Therefore, it is proposed to apply a threshold in this approach to establish the selected features.



(a)

(b)

(c)

Figure 5.12: Progression of the dropout rate in the FS method based on the dropout technique. In (a), the method is illustrated using the Concrete approach, and in (b), it is illustrated using the Gaussian approach. Two cases are presented based on the regularization factor, with λ set to 1 represented by the solid line and set to 2 depicted by the dashed line. Finally, in (c) is illustrated the 'sparsity' during the training in both approaches using a $\lambda = 2$.

The results obtained from the Concrete approach in Fig. 5.9 were obtained using a threshold of 0.9, considering features where $\phi_i < 0.9$ as relevant. Observing Fig. 5.11a, the features that are located in the last bin are discarded. Intuitively, it can be understood that the 'gates' with a 90% of probability of dropout are irrelevant for the model because it is likely deactivated. If it is necessary, this threshold might be more aggressive, such as 0.5. The Gaussian approach does not require this threshold because values converge rapidly to 0 or 1, as illustrated in Fig. 5.12b.

As a result, the Gaussian approach achieves sparsity in fewer iterations compared to the Concrete approach, as illustrated in Fig. 5.12c. The figure indicates that the Gaussian approach begins to attain a sparse representation around epoch 40, while the Concrete approach initiates this process only around epoch 60. In other words, there are no dropout rates in the Concrete approach to surpass 0.9 (the threshold used) until this epoch. Furthermore, the Gaussian approach reaches its highest sparsity around epoch 130, whereas the Concrete approach attains its maximum value in the later epochs. Therefore, the Gaussian approach converges to a sparse representation more efficiently in terms of number of iterations.

Finally, it is noteworthy that in the initial steps, when λ_t is small, the different ϕ_i tend to approach 0. For instance, Fig. 5.12b shows that the different features have a decreasing ϕ_i, but after 40 epochs, the regularization factor λ_t starts to become relevant, penalizing the number of activated 'gates.' This behavior was expected, as described in Section 5.3.1 because the model aims

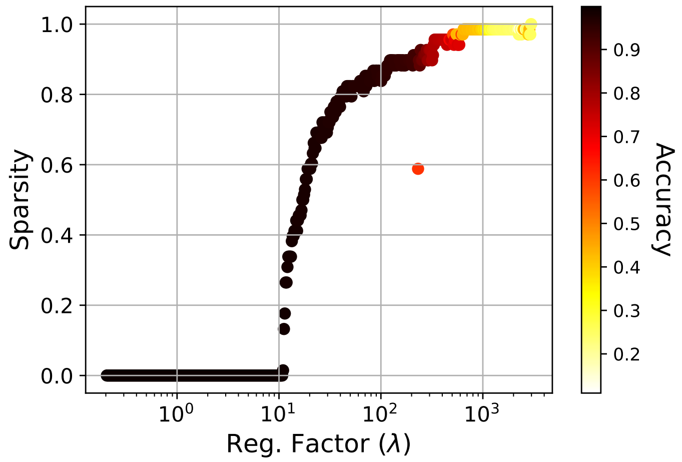


Figure 5.13: LassoNet progression by increasing the L_1 penalty indicated by λ.

to minimize $\mathcal{H}(S)$ in the case where all features are considered. In other words, it attempts to minimize the uncertainty produced by the dropout rates while propagating the most information contained in the input X. When λ_t becomes relevant in penalizing the activated 'gates,' the irrelevant features start to be discarded, reducing $\mathcal{H}(S)$ by $\phi_i \to 1$.

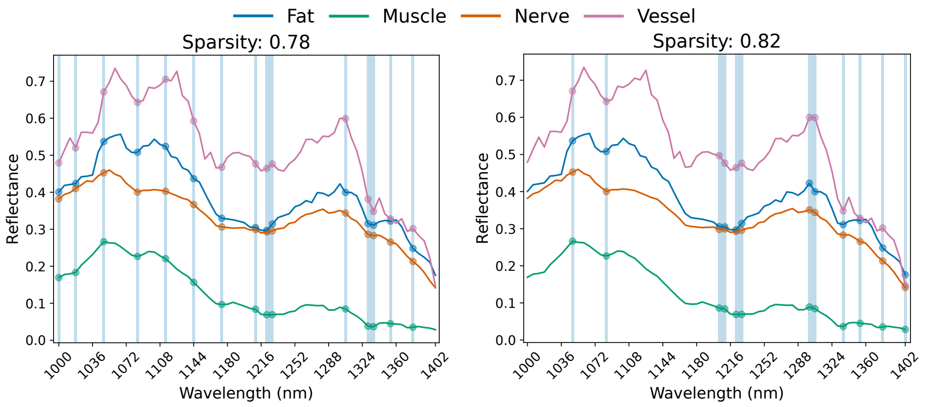
5.4.4 Comparative Evaluation

In this section, the FS method based on dropout will be compared with the other two relaxations of the L_0 methods mentioned in this chapter, namely, LASSO and LassoNet. As mentioned earlier, these methods address the non-convex L_0 problem by incorporating L_1 regularization, transforming it into a convex problem suitable for linear models. LassoNet [246], another DL approach for FS (see Fig. 5.2), requires post-thresholding due to the optimization of the L_1 norm using gradient descent. This section will provide a comparative analysis of both methods with the proposed approach in this chapter.

Firstly, it is important to note that LassoNet has two key hyperparameters for optimal performance: M, which controls the thresholding, and λ, the regularization factor (see Eq. (5.1)). Setting $M = 0$ corresponds to exactly LASSO, so it is crucial to choose a value that considers the non-linear components introduced by the feed-forward DNN. Various values of M have been evaluated, and for this experiment, it has been determined that $M = 2$. Additionally, it is essential to set λ to control the L_1 penalty. Figure 5.13 illustrates the progression of the DNN as λ increases, resulting in higher sparsity of the subset X_s. Here, it can be observed that the Lagrangian used for the regularization factor has to scale highly for obtaining a sparse representation, which complicates the selection of the optimal λ value. Notably, around 85%, the accuracy of the model begins to be penalized. As a result, a value of $\lambda = 100$ is chosen, corresponding to a point of sparsity around 80%.

Regarding the LASSO method, the L_1 penalty needs to be set. For comparison purposes, this value is set to $\lambda = 100$ to demonstrate the differences with LassoNet.

The wavelengths selected for the Laparoscopy HSI dataset by LASSO and LassoNet are illustrated in Fig. 5.14. It is observed that the sparsity obtained with $\lambda = 100$ is similar to that obtained in the FS based on dropout using $\lambda = 2$, as shown in Fig. 5.9 on the right.



(a) (b)

Figure 5.14: HSI band selection using (a) LASSO and (b) LassoNet. Both methods have the regularization factor of the L_1 penalty (λ) set to 100.

Table 5.3: Classification metrics results for each FS method. The specific hyperparameters for each method are indicated.

Method	Accuracy	Precision	Recall	F1-Score
Concrete ($\tau = 0.3,\ \lambda = 2$)	0.9934	0.9933	0.9933	0.9933
Gaussian ($\sigma = 0.5,\ \lambda = 2$)	**0.9959**	**0.9958**	**0.9959**	**0.9959**
LassoNet ($M = 2,\ \lambda = 100$)	0.9666	0.9672	0.9664	0.9665
Lasso ($\lambda = 100$)	0.9834	0.9833	0.9833	0.9833

In the case of LassoNet, as shown in Fig. 5.14b, the band selection corresponds to the results obtained by the method proposed in this chapter. Specifically, the features selected are the same as those observed in Fig. 5.9c. This result is understandable since both DL methods use a relaxation of the L_0 penalty. The difference lies in the way of obtaining that relaxation, being the proposed method, based on the dropout technique (see Section 5.2.1), a statistical approach.

Concerning the LASSO method, utilizing a linear model, it demonstrates a distinct pattern of selecting equidistant bands, as illustrated in Fig. 5.14a. Furthermore, it exhibits a less sparse representation compared to LassoNet, despite sharing the same value for λ, and the FS method based on dropout. This distinction emphasizes that the non-linear components controlled by M in LassoNet contribute to obtaining a more compressed representation of the input X.

Table 5.3 presents the different classification metric values obtained by the various models. As observed using the test set, the results demonstrate that the different methods perform well in solving the problem. The primary difference among the methods is the sparsity obtained, with LASSO, the non-DL method, exhibiting lower sparsity.

It is essential to highlight that FS methods aim to enhance the interpretability of the data, as emphasized in the introduction section. In Section 5.4.3, the justification for the selected features in Fig. 5.9 was provided by referring to relevant studies in the literature. This rationale was not solely derived from the higher sparsity of the FS method based on dropout; rather, conclusions were drawn based on the specific pattern observed in different clusters of bands. Examining the result obtained by LASSO (Fig. 5.14a), it becomes evident that such a pattern is not manifested with this

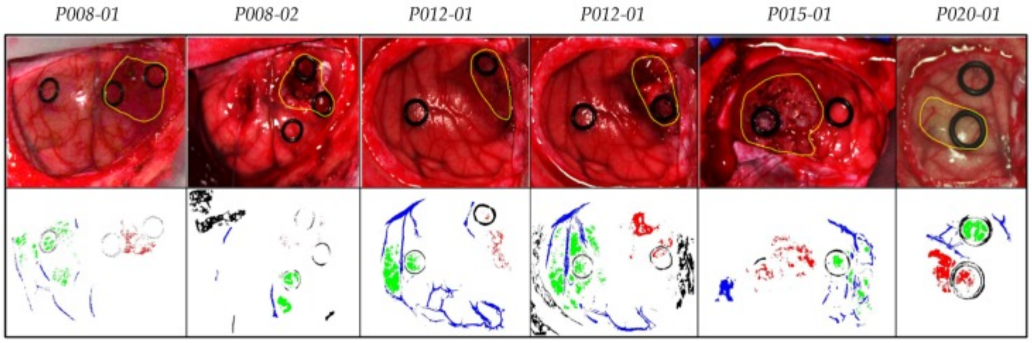


Figure 5.15: The hyperspectral images from 4 different patients, where the tumor area is surrounded in yellow (first row). The ground-truth, labeled samples from a semi-automatic labeling tool, are illustrated in the second row. Normal, tumor, hypervascularized and background classes are represented in green, red, blue, and black, respectively. Image obtained from [260].

method. However, it is expected that LassoNet could capture this pattern, although the scale of the regularization factor (see Fig. 5.13) complicates this observation.

5.4.5 Additional Experiment: Band Selection for Brain Cancer Detection

The high-dimensional structure of HSI is expected to contain essential information for tissue identification, making it potentially valuable for cancer detection. Martinez et al. [260] identified the most relevant wavelengths for brain cancer detection by reducing the HSI data from 128 bands, spanning the spectral range from 400 to 1000 nm, to 48 bands.

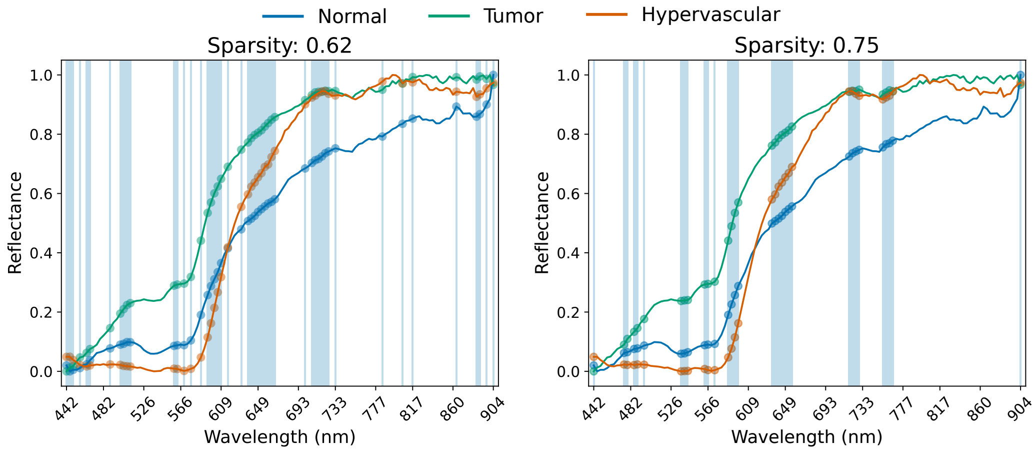
In their study [260], the authors proposed the use of different wrapper FS methods, such as ACO and GA. As mentioned in Section 5.1.2, these methods are computationally intensive and may yield subsets that are overly specific to the particular classifier used.

This section aims to provide a qualitative evaluation of the proposed FS method, comparing the results obtained with those achieved by the computationally intensive methods proposed in the original work. The goal is to assess if the proposed method's results are comparable to those obtained by the more resource-demanding approaches.

Dataset and Preprocessing

For this study, six hyperspectral images from four different patients were employed. These images encompass pixels labeled into four distinct classes: normal tissue, tumoral tissue, hypervascularized, and background. Figure 5.15 illustrates the hyperspectral image, represented by a synthetic RGB image, along with the labeled pixels. The label 'background' is ignored, as it represents elements outside the brain. Image labeling [260] was conducted using a combination of pathology assessment and neurosurgical criteria through a semi-automatic tool based on the Spectral Angle Mapper algorithm [272].

As evident from Fig. 5.15, the dataset is imbalanced. The tumoral tissue, as a reference, contains 11,054 labeled pixels, while both normal tissue and hypervascularized contain over 25,000 pixels each. To address this imbalance across the six images, the proposed K-Means-based method was employed for dataset balancing [260]. The authors suggested undersampling the dataset to 1000 samples per class, justifying this reduction due to the quadratic computational complexity of the SVM, the 'black



(a)

(b)

Figure 5.16: Most relevant spectral bands for brain cancer detection. In (a), the bands proposed in [260]. In (b), the relevant features are illustrated based on the FS method employing dropout technique.

box' used in their wrapper methods. This limitation is not necessary in the proposed method but to ensure unbiased conclusions, the same undersampling method is applied in the results presented in this section.

Ultimately, the images undergo preprocessing, adhering to the guidelines outlined in the original work [260]. Moreover, any outliers have been removed using the IQR methodology, as it was done in Section 5.4.2.

Band Selection Result

The experimental setup, which is described in Section 5.4.1, has been applied using a DNN architecture identical to the evaluated using the Laparoscopy HSI dataset (see Section 5.4.3). The DNN architecture based on Fig. 5.5, it is composed by fully-connected layers of 128, 100, 32, and 3 neurons. The FS configuration used corresponds to the Concrete approach, Eq. (5.6), setting the hyperparameter τ to 0.3 and the regularization factor $\lambda = 2$.

Figure 5.16 presents a comparison between the spectral bands selected in [260] and those selected by the proposed approach. Firstly, it is evident that the proposed approach achieves higher compression of subset X_s, resulting in increased feature sparsity. Consequently, the 48 bands proposed in [260], illustrated in Fig. 5.16a have been reduced to 32 spectral bands, depicted in Fig. 5.16b.

From a biological analysis perspective, certain wavelength ranges have been associated with specific optical properties of cancer tissues [144]. Hemoglobin is particularly relevant for identifying tumoral tissue, and absorbance peaks in the spectral range from 450 to 600 nm have been observed [273, 268]. Specifically, oxygenated hemoglobin exhibits two absorbance peaks around 550 and 600 nm [268].

The results in Fig. 5.16 indicate the presence of the aforementioned wavelength ranges in both methods. Specifically, the wavelength range around 600 and 560 nm is selected in both cases, with the band around 600 nm appearing more relevant. Around 650 nm, both methods identify a group of bands that appear relevant.

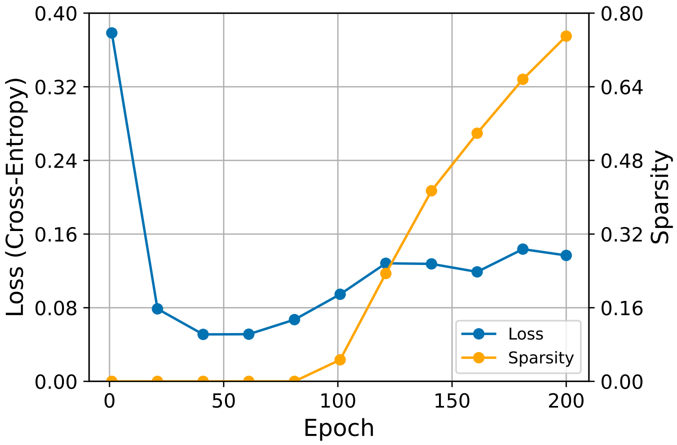


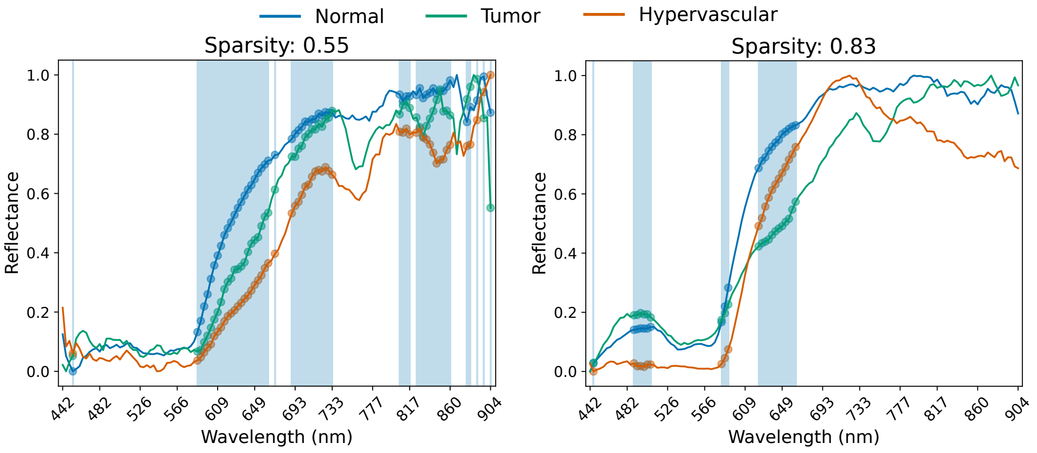
Figure 5.17: Model performance progression with increasing sparsity. The graph illustrates a decline in performance as sparsity is increased.

In the NIR spectrum region, spanning from 700 to 900 nm, there is a region selected by both methods, particularly around 720 nm. However, there is more discrepancy in this region. Figure 5.16a shows a sparse selection of bands in that region, choosing 8 different bands from 777 to 900 nm. On the other hand, Fig. 5.16b places specific emphasis around 760 nm. However, the specific reason for the relevance of this wavelength remains unclear, suggesting that the proposed methods may have identified a pattern influenced by the acquisition system or environment.

Finally, it is worth noting that the feature sparsity in the DL model leads to a loss of performance in the model, as it is depicted in Fig. 5.17. As it is observed, the model starts to obtain the sparse representation in the epoch 100 and from that moment the model's performance, represented by the CE, is slightly worse, converging to ~ 0.15. It is not specially relevant, but it provides a perspective of the problem associated with the wrappers methods, as it was indicated in Section 5.1.2, these methods might result in subsets that are overfitting the model.

Considering that the dataset consists of different images, it can be assumed that some hyperspectral images may pose challenges in identifying distinct patterns. For example, Fig. 5.18 presents two examples using the images P008-01 (Fig. 5.18a) and P015-01 (Fig. 5.18b). In these experiments, it is evident that P008-01 requires the inclusion of bands in the NIR range, specifically from 800 to 900 nm. On the other hand, P015-01 utilizes bands reflected in Fig. 5.16b. Consequently, it can be inferred that the subset X_s observed in Fig. 5.16b impacts the performance in accurately estimating the signals extracted from P008-01. However, the selected bands proposed in [260] may better handle these signals from P008-01, as they include more bands in the NIR range.

Putting aside details, the observations in Fig. 5.18a might be a consequence of the noise present in the hyperspectral image due to the acquisition process, but this cannot be guaranteed. It is essential to emphasize that the results presented in Fig. 5.18 are obtained using a heavily unbalanced dataset. Despite applying techniques to address this issue, the results are presented for comparison purposes and should not be considered definitive.



(a) (b)

Figure 5.18: Most relevant spectral bands for brain cancer detection in (a) P008-01 and (b) P015-01.

5.5 Feature Selection by Dropout: Information-Theoretical Validation

This section aims to validate the IT perspective described in Section 5.3. By employing the aforementioned IT perspective, it seeks to establish a theoretical understanding of DL models, with the goal of bridging the gap between interpretability and accuracy trade-offs in these models. Continuing from the work presented in Chapter 4, this validation will be conducted through IP analysis.

Figure 5.19 illustrates the IP estimation obtained by the kernel-based entropy estimator using the RBF kernel and applying the kernel-aligment for the kernel-width estimation (see Section 3.6.3). However, for obtaining a better estimation in the first iterations, the kernel-width has been initialized by the normalized Silverman rule, Eq. 3.31. Regarding the DL model used, it has been obtained by replicating the experiments indicated in Section 5.4.3, specifically using the Concrete approach by setting the hyperparameter τ to 0.3 and the Lagrangian multiplier in the objective function λ to 2, see Eq. (5.4). This model obtained the result illustrated in Fig. 5.9b on the right.

As mentioned in Section 5.4.3, the DNN consists of four layers: the feature selector, responsible for generating the subset X_s, and three fully-connected layers with 100, 32, and 4 neurons, represented in Fig. 5.19 as L_1, L_2, and L_3, respectively.

Firstly, considering the Markov property of the DNN architecture, it is crucial to validate the DPI. In Section 4.1, it was indicated that a supervised DL model employs mechanisms to solve the problem defined by $p(x, y)$. In this process, the transmission of information is governed by two DPI: the forward DPI defined by Eq. (4.1) and the backward DPI depicted in Eq. (4.2). This pattern is well identified in Fig. 5.19, decreasing the $\mathcal{I}(X;T)$ in the deeper layers and increasing $\mathcal{I}(T;Y)$ in the deeper layers.

As documented in Section 5.3.2, within the IB framework, the bottleneck in this architecture is located in the subset X_s obtained by the FS method using the dropout technique, as shown in Eq. (5.13). Therefore, subset X_s defines the trade-off between compression and prediction, indicating the minimal information on X required for predicting Y. This compression of information is illustrated in Fig. 5.19, continually compressing the input X by selecting the relevant features defined by the

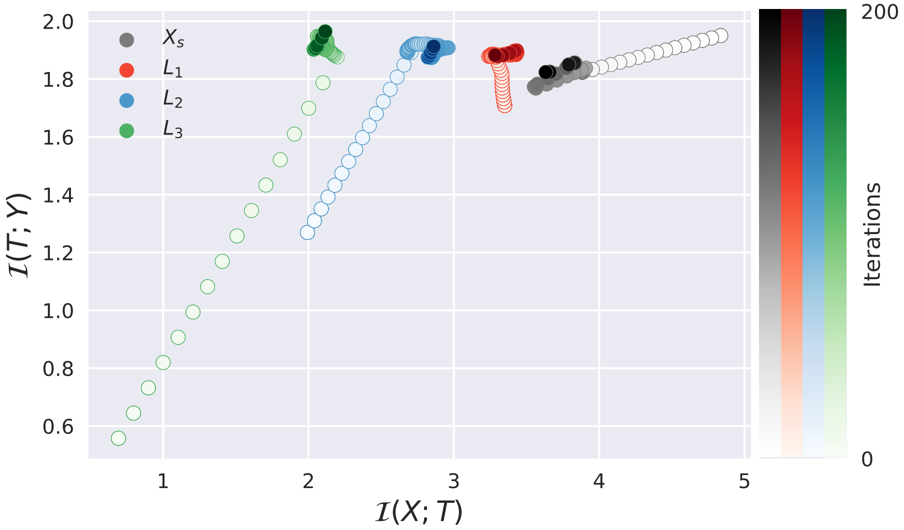


Figure 5.19: IP estimation using the FS method based on dropout technique. As reference, the estimation of $\mathcal{H}(X)$ is around 5 bits.

subset X_s.

As observed in Fig. 5.19, the loss of information in X results in a smooth reduction in the predictability of Y. The subset X_s compresses X by more than 1 bit, and this reduction has an impact on predictability of around 0.1 bit. Based on this observation, it is concluded that in the original space the compressed information is not fully representative for obtaining Y and the hidden layers of the model have to 'transform' the information contained in the subset for getting a more evident representation for predicting Y.

Regarding the hidden layers of the DNN, the different layers show a clearly fitting phase during the training of the model. It can be deduced a light compression phase in the layer L_1 but it is not specially relevant. Additionally, the output layer, L_3, converges to a point where $\mathcal{I}(X;L_1) = \mathcal{I}(L_1;Y) = 2$. Given that the test dataset is balanced and there are 4 classes to identify, the theoretical maximum value corresponds to $log_2(4) = 2$. Given this, it can be considered that the model is perfectly identifying the pattern of both variables, X and Y. This might be a consequence of defining the bottleneck by X_s, since the DNN has less liberty to extract features from the original space to extract features that he considered relevant, as it is commonly done in the input layers of a DNN.

In conclusion, the IT perspective of the proposed method described in Section 5.3 has been validated through the IP analysis. This analysis, based on IP, is particularly relevant for illustrating that X_s is a compressed version of the input X and how this compression involves a loss of information for identifying Y. In other words, the compression of information leads to a loss of predictability, as expected, and it is far from the theoretically ideal compression mentioned earlier, where $\mathcal{I}(X_s;Y) = \mathcal{H}(Y) = \mathcal{H}(X_s)$. For this reason, it is essential to note that, in general, higher compression would imply that the selected features become less representative. However, as it has been previously indicated, a subset X_s with a minimal loss information for predicting Y, in practical case, can still be considered as a Markov Blanket. Furthermore, if the output layer is considered, it is observed that it meets the requirements to consider said space as a Markov Blanket, but in a space obtained from different non-linear iterations that pose a challenge to interpret in terms of the original variable.

5.6 Discussion

As introduced at the beginning of this chapter, in medicine, the FS is not focus exclusively on reducing the dimensionality of the problems but also on reducing costs involved, as it is crucial to understand reasons behind disagreements regarding disease diagnosis among image-analysis experts [232]. Therefore, it is necessary to use a FS method that, in this chapter, is proposed a supervised DL model.

In this chapter, it has been proposed to use a FS method based on the dropout technique, essentially solving the problem of the sparse representation defined by L_0 regularization, which is non-convex and computationally expensive and intractable for high-dimensional data [120]. On the other hand, LASSO is a linear approach that transforms the problem into a convex problem by using the L_1 as a regularization factor, instead of the L_0. This approach has been adapter to DL model previously [245], with the current reference being LassoNet [246]. As it has been demonstrated in Section 5.4.4, the solution obtained by LassoNet and the FS method described in Section 5.2.1 yield similar results. This is because both methods are addressing the same problem, defined by the L_0, but from different perspectives, with the FS method based on dropout employing a statistical approach.

The primary goal of this thesis is to enhance the interpretability of DL models to improve their reliability through the use of IT. This task is specially relevant in different fields, such as medicine, where the trade-off between accuracy and interpretability of the model is relevant [91]. For this reason, an IT perspective has been to describe the solving-mechanism of this FS method by dropout, as discussed in Section 5.3.

Given this IT perspective, it has been identified a specific configuration for improving the performance and dealing with problems of the proposed FS method. Firstly, it has been demonstrated which is the optimal initial value for the configuration of the different dropout rates, ϕ_i, for the mechanism of this approach described in Eq. (5.2). In this case, it has been established that the dropout rates should be initialized as $\phi_i = 0.5$, starting with the maximum entropy scenario. This is because the algorithm aims to solve the problem by minimizing entropy, as indicated in Eq. (5.12). Secondly, the problem has been shown to be well-defined in two additive terms, as seen in Eq. (5.14), in which the second term has a restriction. To address the challenge of solving both problems simultaneously, the annealing trick has been proposed using a cosine scheduler, as illustrated in Eq. (5.15). Consequently, the model initially focuses on fitting the parameters of the DL model, and gradually, the sparse representation in the subset X_s, the restriction $\mathcal{R}(S)$ in Eq. (5.14), becomes more important as λ increases in Eq. (5.4).

Initially, the model focuses on reducing the uncertainty produced by the dropout used in the feature selector by minimizing the entropy, $\mathcal{H}_\phi(S)$ in Eq. (5.14) without considering the restriction in the set of indices S. This involves propagating the maximum possible information from the input X, represented by $\phi_i \to 0$ as it means that the different features will not be dropped out. This progression is illustrated in Fig. 5.12, where it is observed that in the first epochs, the different ϕ_i tend to 0 and later increase in those features that are considered irrelevant. Additionally, Section

5.5 validates the IT perspective by using the IP. As a conclusion, the subset X_s, based on the IP analysis, can still be considered as a Markov Blanket, the subset containing the full information necessary for inferring Y.

Finally, the results obtained for band selection in HSI data have been validated by the spectral analysis reported in the state-of-the-art. In this case, it has been used a hyperspectral image obtained from a HSI rigid scope system that has been applied as a laparoscopy device (see Section 5.4.2) and a HSI dataset, reported in the literature, for brain cancer detection (Section 5.4.5). Using both dataset, it has been identified that the FS method has selected a subset of wavelengths that are specially relevant for solving the problem. For instance, the image obtained by the HSI laparoscopy device has different tissues and the FS method has special emphasis in the range around 1200 nm, which is specially relevant for identifying muscle and fat tissue (see Section 5.4.3). In the case of the HSI brain cancer dataset, the subset of wavelengths obtained by the FS method has much similarity with the band selected in [260], see Section 5.4.5. Additionally, the FS method has selected multiples bands that are located in the spectral range from 450 to 600 nm, which has been reported to be specially relevant for identifying tumoral tissue [144].

Dropout Approach: Overfitting Concern

One of the most common concerns associated with DL models is overfitting. In the case of the study described in Section 5.4, the samples consist of pixels obtained from a limited number of hyperspectral images. Therefore, it is essential to ascertain whether overfitting is present when using the DL approach for FS described in Section 5.2. It is noteworthy that the IP analysis observed in Section 5.5 provides clear evidence that the model is not overfitting, following the work documented in Chapter 4. It is especially relevant that the output layer converges to $\mathcal{I}(X;T) = \mathcal{I}(T;Y)$, representing the theoretical Markov Blanket.

The aforementioned problem has been reported by reviewers from different works that we have published where it has been used the proposed method for DFU detection, using the dataset described in Section 4.2. In these works [226, 227], this approach was employed by using the scarce dataset for DFU. The FS by dropout technique was employed as a feature ranking instead of FS by applying different techniques, such as cross-validation, for establishing a ranking based on the parameters ϕ_i of the aforementioned FS method. However, for non-specialists, this process might be questionable because the features selected might be the consequence of a overfitting of the model. For this reason, we provided a intuitive explanation to mitigate the concerns.

Given that the dropout technique was initially designed to address overfitting in DL models, it is important that this technique be not only used in the first layer but included in the different layers of the model. In Section 4.3.2 has been exposed different benefits of using this technique. However, in the first layer, specifically, the feature selector for obtaining the subset X_s, the dropout rate acts as a variational parameter that the model can adjust.

Using the dropout technique to obtain the subset X_s imposes a significant restriction, compelling the model to place particular emphasis on working within the original space to achieve optimal compression of this space defined by X. This limitation is advantageous for the DNN, enabling it to identify distinctive ‘patterns’ present in the original space. For example, consider a subset of pixels in a hyperspectral image that exhibits a specific pattern due to noise injected by the sensor of the acquisition system, evident in a few specific bands. When employing a FS method, such a pattern may not be deemed significant, especially if it appears in only a subset of samples. Consequently, the 'information' (is just noise) conveyed by this pattern should not be transmitted, and it will not be considered for obtaining the new 'representation' of the input X.

In the FS method using the dropout technique, the input layer is characterized by variational parameters that define a Binomial distribution consisting of d independent Bernoulli 'relaxed' distributions. This configuration serves as a 'gate' to identify irrelevant features by introducing noise through an increase in ϕ_i. In an ideal scenario, relevant features tend to have a dropout rate close to 0, while irrelevant features tend towards a dropout rate of 1. Essentially, this imposed restriction in the input layer helps mitigate concerns related to overfitting, as irrelevant information is not propagated to the subset X_s. Consequently, the model might experience a loss in performance, as observed in Fig. 5.17. As indicated in Section 5.4.5, this could be a consequence of utilizing images from different patients and conditions, where specific patterns may result in a loss of precision for detecting certain samples.

5.7 Conclusions

Feature Selection methods play a fundamental role in those sceneries where the data interpretation is fundamental. For instance, in the context of healthcare, the comprehensive data interpretation is essential since a bad interpretation of the data can result in inconsistencies among experts when diagnosing a disease, leading to increased variability in clinical decision-making. Additionally, those techniques permit tackling the multitude of issues triggered by the so-called 'curse of dimensionality'.

This chapter aims to introduce a FS embedded method utilizing DNN. The current state-of-the-art includes various FS methods based on DL, and it has been argued that the most compelling approach for this task is the one presented here.

As this thesis endeavors to enhance the interpretability of the solving mechanisms employed in DNN models, aiming for a more profound understanding of these algorithms, an explanation has been offered based on an IT perspective, extending the discussed in Chapter 4. By delving into the internal mechanisms utilized by DNN to solve problems, this exploration provides insights that can contribute to improving results and ensuring accurate identification of relevant features.

This study has been evaluated through the hyperspectral band selection problem, with the aim of identifying the most crucial wavelengths for distinguishing various tissues. The proposed FS has been validated for identifying relevant wavelengths for the identification of different tissues, aligning with findings documented in the literature. Furthermore, one case of study has been compared with the estimation of wavelength selection obtained by wrapper methods, known for their computational demands. Results indicate similarity between the proposed method and wrapper methods, with the proposed approach yielding a more compressed subset. This compression discrepancy may stem from wrapper methods being overly specific to the classifier, selecting wavelengths pertinent to only a limited number of samples.

As a result, this chapter outlines a DL model that delivers a solution and enhances the interpretability of the provided data. This aspect can be particularly significant for the integration of DL-powered AI in healthcare scenarios. The FS approach using DNN introduces a more interpretable trade-off between compression and prediction, as discussed in Chapter 4. With the emphasis on obtaining the best compression in the original space defined by the input, the model's decision-making process becomes more intuitively understandable. This improvement can contribute to enhancing efficiency in medical applications' workflow and instill greater reliability and trustworthiness.

Chapter 6

Signal Decomposition: HyperSpectral Unmixing

> The purpose of computing is insight, not numbers.
>
> *Richard Hamming*

In Chapter 5, FS was introduced as a method for improving data interpretability, with HSI employed as the case study for experimental validation. As previously discussed, HSI is a rich image modality that provides a substantial amount of information. However, the inherent high-dimensionality of this data presents challenges for analysis and estimation, as well as other issues related to the 'curse of dimensionality' [228]. Chapter 5 addresses these concerns by extracting a subset of relevant features, thereby identifying the most informative wavelengths for understanding the spectral content of the hyperspectral images (see Section 5.4).

Nonetheless, the information contained in HSI data can be further leveraged to obtain a more in-depth understanding of the captured scenes. Specifically, identifying the composition of HSI data is a crucial task that enables a more comprehensive interpretation of the scene. HyperSpectral Unmixing (HSU) plays a fundamental role in this process, providing a new data representation based on the extraction of endmembers —pure signatures— and their abundance estimations. In this chapter, a DL model for blind HSU will be proposed, drawing inspiration from contrastive learning approaches and the IT perspective developed in previous chapters, to obtain the endmembers and their abundance estimation.

HyperSpectral Unmixing is formulated as an inverse problem (see Section 2.4) which aims to decompose the pixel of the hyperspectral image into pure spectral signatures —endmembers— and their corresponding proportions, represented as the abundance map. In blind unmixing, both endmembers and abundances are estimated simultaneously, making it an ill-posed problem [110]. The number of endmembers and their spectral signatures are typically unknown, and the observed spectral measurements are often affected by factors such as atmospheric interference, sensor noise, and spectral variability of the endmembers [131, 274]. It is also important to note that the notion of a "pure material" is subjective and problem-dependent [275].

While HSU is comparable to dimensional reduction algorithms such as PCA and ICA, it introduces a critical factor absent from traditional methods: the emphasis on interpretability. Specifically, HSU decomposes signals within the original input space, guided by certain constraints. This approach ensures that the resulting formulations enhance the physical understanding of the data, so the solutions are constrained to physically meaningful ranges and regularized solutions.

PCA and ICA have been explored in the context of blind HSU [276] but, as discussed in Chapter 5, they do not necessarily improve interpretability. While HSU is not primarily a dimensionality reduction problem, it involves projecting the data into a vector space where the basis vectors are defined by the endmembers. This results in a more structured representation of the data, where each pixel is expressed as a combination of a smaller set of physically meaningful components, which in turn improves interpretability by using the original spectral bands. This approach contrasts with methods like PCA, where the generated dimensions are often linear combinations of the original features, which can complicate interpretability.

The core of HSU lies in defining the spectral mixture model that relates the abundance map and the endmembers. The Linear Mixture Model (LMM) is the most widely used [277], particularly in remote sensing, although Non-linear Mixture Model (NMM) are recognized for their applicability in more complex scenarios [278].

HyperSpectral Unmixing has been extensively applied to material identification in hyperspectral images, providing not only spectral signatures but also spatial locations within the image. This is particularly prevalent in remote sensing [275], though its applications have expanded to new fields, such as disease diagnosis and surgical guidance [135], forensic science [136], food quality evaluation [137], plastic identification for recycling [138], and art conservation [139].

In healthcare, where data interpretability is paramount for reducing inconsistencies in disease diagnosis [232], HSU methods offer a promising avenue for the integration of HSI in medical applications. For example, HSU has been applied to distinguish between healthy and tumorous brain tissue using hyperspectral imaging [276], as described in Section 5.4.5. By estimating the abundance maps of these endmembers, the study enabled classification based on similarity metrics like the Spectral Angle Mapper (SAM) [279] or Spectral Angle Distance (SAD), methodologies commonly used in remote sensing [280]. Additionally, Fabelo et al. [281] used HSU for quantifying tissue composition, validating the use of HSI as a complementary diagnostic tool.

The application of DL-powered AI in HSU tasks is a practical and expected approach. As discussed in previous chapters, using DL to obtain more interpretable data is a logical-desired step, given its high performance in obtaining compressed representations of input data. Illustrative examples of DL's effectiveness in data compression include AEs, as discussed in Section 4.3, which demonstrate high-performance compression, and the DL-based FS method in Section 5.2 that achieves a more interpretable form of compression. In the context of HSU, it can be viewed as another form of dimensional reduction, compressing information into a new representation using original variables. However, aligning with the hypothesis of this thesis, the IT-perspective is provided to gain more in-depth understanding of the problem-solving mechanism of the DNN architecture.

In this chapter, we introduce a novel DL framework for blind HSU, referred to as the Contrastive Learning for blind Hyperspectral Unmixing (CLHU). This framework is strongly inspired by contrastive learning techniques in DL, see Section 2.2.5. In blind HSU, the goal is to estimate both endmembers and their abundances, with contrastive learning playing a crucial role in endmember estimation, controlled by a regularization factor that helps reduce model uncertainty — an essential factor in enhancing interpretability.

While the proposed DL approach is designed specifically for blind HSU, it holds the potential to extend to other high-dimensional data modalities. In other words, although we can consider that the definition of an endmember is problem-dependent, this chapter focuses on blind HSU, the primary domain for which this framework was developed.

The chapter is structured as follows: it begins by presenting state-of-the-art techniques and foundational concepts for understanding hyperspectral unmixing. Next, the proposed DL method for HSU, Contrastive Learning for blind Hyperspectral Unmixing (CLHU), is introduced. Following this, the IT perspective is applied to offer deeper insight into the problem-solving mechanisms. Finally, experimental results are provided using a state-of-the-art dataset and a hyperspectral image for biological tissue identification, followed by a discussion and conclusions

6.1 Foundational Concepts in Hyperspectral Unmixing

In this section, different concepts will be introduced to help to obtain a more in-depth understanding of the HSU process.

6.1.1 Spectral Mixture Model

The spectral mixture model is an essential part of the HSU and the LMM is widely used. The rationale behind this is that, despite its simplicity, it serves as a commendable representation of the light scattering mechanisms in numerous real-world scenarios [275].

From a physical perspective, the LMM is predicated on the assumption that incident light interacts with only one material on the surface before being captured by sensors [277]. This premise underpins the LMM's validity, particularly when the mixing scale is macroscopic [275].

According to the LMM, a mixed pixel $x_i \in X \in \mathbb{R}^{N \times L}$ is a convex combination of endmembers, with its abundances as coefficients as follows:

$$\begin{gathered} \hat{x}_i = a_i M + \epsilon_i \\ s.t.\ 1^\top a_i = 1 \ and \ a_i \succeq 0 \,, \end{gathered} \tag{6.1}$$

where i denotes the pixel index, a_i is a row vector in $A \in \mathbb{R}^{N \times \mathcal{P}}$, with A representing the abundance matrix, $M \in \mathbb{R}^{\mathcal{P} \times L}$ is the endmembers matrix and ϵ_i is an additive noise term. Here, The $\mathcal{P}$ denotes the number of endmembers, N the number of pixels and L the number of bands in the HSI data. As it can be deduced, the matrix $M = \{m_1, \ldots, m_{\mathcal{P}}\}$ defines the different endmembers that represent the basis of the convex combination. The symbol $\succeq$ denotes componentwise inequality. As depicted in Eq. (6.1), the abundance matrix A has two constraints: the Abundance Sum-to-one Constraint (ASC) and the Abundance Non-negativity Constraint (ANC).

However, the LMM does not consider the spectral variability that is an effect commonly observed in many scenes. Spectral variability refers to the differences in the spectral signatures of a pure material produced by external factors, such as illumination, or can also be intrinsic to the pure material [282, 283]. Spectral variability was addressed by the Extended Linear Mixture Model (ELMM) [284]. It allows varying the endmembers within a hyperspectral image by using scaling factors, resulting in the following representation:

$$M_i = M_0 \mathbb{I}(\varphi_i)\,, \tag{6.2}$$

where M_0 represents the original endmember library and $\mathbb{I}(\varphi_i)$ is the diagonal matrix with the vector $\varphi_i \in \mathbb{R}^{\mathcal{P}}$, which contains the scaling factor of the $\mathcal{P}$ materials. Despite its ability to capture the spectral variability caused by changes in illumination and topography [285], the ELMM model's capacity to represent more intricate variability is limited. This led to the development of a Generalized Linear Mixture Model (GLMM), which incorporates a scaling factor for each band, as proposed in [286].

Although the LMM is commonly used in HSU, several NMM have been proposed in the literature [287]. The NMM is usually due to physical interactions between the light scattered by multiple materials in the scene. The scattering interaction can be modeled by different levels, a classical, also known as multilayered level, or microscopic level [287, 275]. As it is expected, the microscopic level is unapproachable since it is an extremely complex ill-posed problem. For this reason, the multilayered based models are commonly used.

In general, the first-order multilayered models are sufficient to handle the scattering effect, and this leads to the bilinear model [287, 275]. There are several works where the bilinear model has been proposed for HSU [288, 289, 290], incorporating the cross products across different endmembers to account for the non-linear mixing effects. An extended version of the aforementioned mixture model is the Generalized Bilinear Model (GBM), which introduces a scalar γ value that denotes the degree of interaction among the endmembers, m_j and m_k:

$$\hat{x}_i = MA_i + \sum_{j=1}^{\mathcal{P}-1} \sum_{k=j+1}^{\mathcal{P}} \gamma_{i,j,k} A_{i,j} A_{i,j} m_j \odot m_k + \epsilon_i\,, \tag{6.3}$$

where $\odot$ is the element-wise Hadamard product. In this way, the strength of the non-linear components is controlled by γ regarding the overall mixture.

6.1.2 Endmember Extraction

One of the most relevant limitations inherent in the HSU algorithms is the necessity to have prior information about the endmembers, the pure material signature. In those cases where such information is not available, these signatures, the endmembers, have to be estimated from the data [275]. This estimation is carried out by the endmember extraction algorithms.

The most popular endmember extraction algorithms are geometrical based approaches [275], although there are methods where the spatial contextual information in the analysis is applied for the endmember extraction process [291]. There are different examples of non-geometrical endmember extraction algorithms [292, 293], but in this section it will be briefly introduction specific geometrical endmembers extraction algorithms.

Due to spectral redundancy, a hyperspectral image often lies in a low-dimensional subspace [294] and, according to convex geometry theory, any data point within a hyperspectral image that satisfies the LMM assumption must reside within a simplex that is defined by the endmembers [295]. Therefore, a noise-free hyperspectral pixel has to be contained in the simplex defined in the

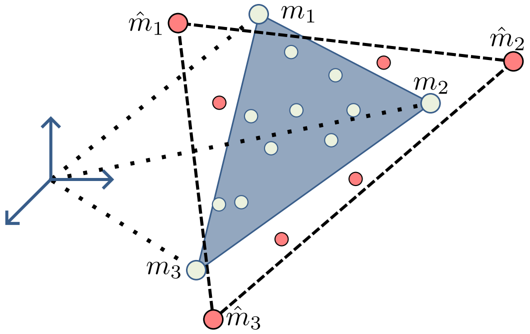


Figure 6.1: Illustration of simplex set to $\mathcal{P} = 3$, $M = \{m_1, m_2, m_3\}$. In this case, it is observed the simplex with the highest volume, represented by dashed lines, is obtained by noisy points, indicated by red dots.

$(\mathcal{P} - 1)$-dimensional subspace. Thus, any data point that contains artifacts produced by noise will be located outside the simplex, see Fig. 6.1. In blind HSU, it is a common practice to extract the endmembers from the hyperspectral image. Most popular endmember extraction methods rely on the geometry of the data for the estimation of the pure material signature, such as Pixel Purity Index (PPI) [296], Vertex Component Analysis (VCA) [297] or N-FINDR [298], which aim to identify the simplex in the low-dimensional subspace whose vertices represent the endmembers, denoted by $M = \{m_1, \ldots, m_{\mathcal{P}}\}$.

In the case of the N-FINDR algorithm, it operates by identifying the simplex with the maximal volume, premised on the principle that the volume delineated by a simplex constituted by pure signatures, the endmembers, surpasses that of any other pixel combination [298, 275]. A notable challenge with this technique is its susceptibility to noise in the data, potentially mistaking noisy pixels for pure ones. A example of this is illustrated in Fig. 6.1. The VCA addresses this issue by iteratively projecting data onto a direction orthogonal to the subspace defined by the already identified endmembers [297, 275]. Consequently, each new endmember signature aligns with the extremity of this projection, effectively mitigating the influence of noise. In the case of the PPI, it introduces a preprocessing that reduces the dimensionality and to improve Signal-to-Noise ratio (SNR), [296, 275].

6.1.3 Unmixing Algorithms

The HSU algorithms aim to optimally estimate the abundance map based on endmembers, which, in blind unmixing, need to be derived from the data, as discussed in Section 6.1.2. Given the constraints of the LMM (Eq. (6.1)), one prevalent method is the Non-negative Matrix Factorization (NMF) algorithm, which decomposes the hyperspectral data matrix into non-negative matrices of endmembers and abundances [299].

Additionally, the Fully Constrained Least Squares (FCLS) [300] algorithm, another widely employed approach, treats unmixing as a constrained optimization problem. The FCLS aims to minimize the sum of squared differences between observed and modeled spectra, where the model is a linear combination of endmember spectra weighted by abundances, subject to constraints on the abundances such as ANC and ASC. As it has been previously mentioned, both constraints are crucial for the LMM.

As it happens with the endmember extraction algorithms, the geometrical approach is popular in unmixing algorithms. These methods are commonly known as minimum volume-based algorithms [275]. As it has been described in the Section 6.1.2, those algorithms discussed are associated with the identification of a pure-pixel endmember extraction, and it is carried out by finding maximum volume simplex estimation. However, as it has been previously discussed, those algorithms are data-dependent and the introduction of noise can generate a simplex in the low-dimensional subspace whose vertices, the estimated endmembers, do not correspond to the real pure material. To counter this, these HSU algorithms introduce a new constraint, the Minimum Volume Constraint (MVC) [301].

In these algorithms, initial endmember estimates, M_0, obtained through endmember extraction methods like N-FINDR, are refined during HSU solution. Taking as reference, Miao and Qi [301] proposed to use the NMF by including this new constraint, MVC. In this way, the algorithm is solving both problems, the abundance and endmember estimation. This constraint is also applied in methods such as the simplex identification via variable splitting and augmented Lagrangian (SISAL) [302] and the minimum volume simplex analysis (MVSA) [303].

These algorithms are not only pivotal in remote sensing but also serve as valuable preprocessing tools in various domains, including the analysis of biological tissues [304].

Deep Learning Algorithms

As it has been discussed in previous chapters, DL models provide an easily adaptable output with high accuracy, and they have been recently applied to carry out HSU problems. These models have been previously proved to be effective for pixel unmixing in multispectral images. Baraldi *et al.* [305] compared the performance of a MLP algorithm with two neuro-fuzzy classification schemes. Nevertheless, Licciardi *et al.* [306] proposed for the first time the use of an AE architecture [67] for abundance estimation in HSU, providing the endmembers as hyperparameters of the model. Rasti *et al.* proposed UnDIP [307] for abundances estimation based on DIP [117], which is an approach for solving different inverse problems such as super-resolution, inpainting, and denoising. In [308], the authors proposed a non-negative sparse AE for endmember extraction. Finally, Palsson *et al.* [309] presented a method for blind HSU based on AEs, where the weights of the decoder are the endmember spectra, see Fig. 6.2, and evaluated the performance using different objective functions. In addition to the previous works, Hong *et al.* proposed the EGU-Net [310] for HSU using a two-stream end-to-end network, which is and AE where the encoder is used for the abundance map estimation. In this case, one stream is employed to estimate the abundance map of pseudo-endmembers, obtained by another HSU method. The decoder is not used in this stream. The other stream estimates the abundance map of the mixed dataset, utilizing the decoder to penalize the model through reconstruction. Finally, Transformer-based architectures [54] have emerged as a dominant paradigm lately, and the use of these architectures has been applied for HSU. Ghosh *et al.* [311] proposed the use of transformers architecture for abundance and endmember estimation, providing the spectral and spatial information, obtaining a consistent performance across different datasets.

Most of these methods leverage DL models' reconstruction capabilities for abundance estimation, often following an AE-like architecture. In this context, the models operate as regressors, mapping input spectral signatures to abundance values. While these models can reconstruct data, they tend

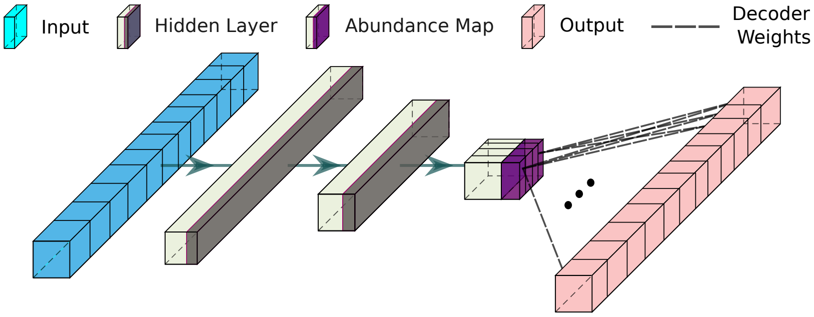


Figure 6.2: AE proposed by Palsson *et al.* [309] for blind HSU. The weights of the decoder are considered the endmember spectra.

to produce physically meaningless endmembers in blind HSU due to the lack of effective guidance for actual endmembers. For example, endmembers derived from the weights of the final layer in the study by Palsson *et al.* [309] often do not match the spectral signals present in the dataset. In some cases, like EGU-Net [310], there is even an absence of evidence for the endmembers underpinning the abundance map estimation.

The use of DNN for HSU, particularly in enhancing the physical comprehension of the data, is pivotal for addressing the challenges posed by spectral variability. As explored in Section 6.1.1, spectral variability can arise from external factors like illumination or the inherent properties of the pure materials. Traditional methods tackle this issue using mixture models like the ELMM or the GLMM. However, as it has been explored through this thesis, the DNN has to reduce the uncertainty of the data, which can be a consequence of the variability in the data, by obtaining an optimal compressed representation in the latent space, see Section 4.3.2.

6.2 Contrastive Learning for Hyperspectral blind Unmixing

In the previous section, it has been described the use of DL models for HSU. However, as it has been discussed, most model rely on the reconstruction capabilities for abundance estimation and the endmembers for obtaining a physical meaning have to be provided as hyperparameters.

As it has been discussed in the introduction of this chapter, HSU algorithms aim to obtain a more in-depth understanding of the scene captured in HSI data. Similar to FS methods, the idea is to obtain a better interpretability of the data and, for this reason, in HSU is specially relevant to obtain a signal decomposition using the original variables, in the high-dimensional space, providing the physically meaningful. For this reason, in Section 6.1 has been focused on geometrical approach since those methods are the most interpretable.

The challenge of spectral variability for external environmental factors, particularly issues like illumination, poses significant difficulties for classical methods in HSI. This is an area where DL models demonstrate a distinct advantage. As discussed throughout this thesis, DL models excel in managing noise within the compressed latent space. From an IT perspective, which has been a focus of this thesis, the primary objective of the model is to minimize the uncertainty in this latent space Z, represented as $\mathcal{H}(Z|X)$. The spectral variability induced by an external factor, such as illumination, can be conceptualized as a form of external noise with uncertainty independent of the

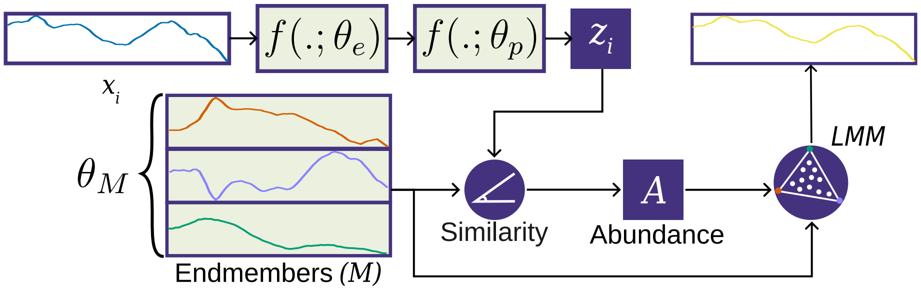


Figure 6.3: The proposed architecture for blind hyperspectral unmixing based on contrastive learning, where θ_i represents the parameters of the deep learning model.

data X, it is a white noise in the inverse problem (see Section 2.4). In other words, this external factor has the same behavior that applying a dropout technique, it is independent of the data and the uncertainty is the sum of both variables $\mathcal{H}(X) + \mathcal{H}(\psi)$, where ψ would represent the 'illumination'. In this way, DNN are tasked with identifying a representation of Z where this issue is substantially alleviated.

Nonetheless, akin to the scenario with PCA, tackling the latent space Z would pose a formidable challenge in obtaining physical significance in HSI. As was the case with FS methods, true comprehensibility requires reverting to the original variable space, denoted by X. Consequently, this section introduces a DL framework specifically designed to achieve interpretable unmixing in blind HSU. This framework focuses on providing effective guidance for identifying real endmembers, a critical aspect often overlooked in standard approaches, effectively bridging the gap between advanced DL techniques and the physical realities of HSI.

The proposed DL framework aims to concurrently estimate endmembers and accurately determine the abundance map. As a result, it is proposed a DL framework for addressing the problem of blind unmixing by jointly estimating the endmembers and fractional abundances map. Unlike most existing DL algorithms, this framework integrates the endmembers estimation and accurately estimates the abundance map by establishing a relationship between the input signal and the estimated endmembers.

6.2.1 Network Architecture

The CLHU framework introduced in this thesis draws inspiration from contrastive learning methods, incorporating several key features typical of these approaches. The architecture of CLHU, as illustrated in Fig. 6.3, follows a structure akin to contrastive learning models. It begins with an encoder, represented by $f(.;\theta_e)$ where θ_e denotes the parameters of the encoder. This encoder is responsible for generating the underlying representation of the input X. Subsequently, this representation is further processed by a projection head, $f(.;\theta_p)$, characterized by parameters θ_p. Unlike SimCLR [80] where the projection head is a small model with no-linear activation in the different layers, CLHU's projection head comprises a single linear hidden layer, designed to project the underlying representation back to the original dimension. In the final stage, CLHU employs the LMM approach to reconstruct the original spectral signature. This reconstruction is achieved through a linear combination of M, which are model parameters denoted as θ_M.

The use of an encoder and a projection head in the design is influenced by contrastive learning techniques. Additionally, the *Similarity* module, another component inspired by contrastive learning methods, is introduced to describe the relationship between the input X and the endmembers M for the purpose of abundance map estimation. As discussed in Section 6.1.3, the majority of DL-based unmixing algorithms employ an AE architecture where the decoder is tasked with estimating the abundance map A [306, 67, 308], aiming to reconstruct the original signal $\hat{X}$. Conversely, in the proposed CLHU framework, A is estimated by the similarity or approximation to M of the latent space Z, positing that the contribution of a endmember is directly proportional to its similarity.

In conclusion, this approach differs from the aforementioned DL-methods by leveraging the relationship between the latent space and the endmembers to estimate the abundance map, rather than relying exclusively on the reconstruction capabilities of the AE's decoder. As mentioned before, in this work the LMM, (6.1), is considered for the reconstruction. Nevertheless, atmospheric effects and instrumental noise [131, 274] usually degrade the original hyperspectral image, so the input X does not interact exclusively with one material, which is the main assumption of LMM. The encoder in the CLHU framework has the purpose to obtain the underlying representation of the input X, which has the best representation in the convex geometric space constrained by M. It can be considered a denoising effect, reducing the spectral variability by external factors, modifying the signal to satisfy the optimal conditions of the LMM, ensuring that all points are contained in the simplex defined in the low-dimensional subspace (see Section 6.1.1). At the same time, the encoder can deal with the spectral variability intrinsic to the pure material.

Similarity Module

As previously mentioned, the CLHU framework enforces a latent space that is constrained by the endmembers M through the *Similarity* module, as shown in Fig. 6.3. This module ensures that all points in the latent space are distributed within a simplex whose vertices correspond to the endmembers. Drawing inspiration from contrastive learning approaches, the *Similarity* module estimates the abundance map $a_i \in A$ based on the distances between the point $z_i \in Z$ and the endmembers M:

$$\begin{gathered} a_i = \sigma\left(\log\left(\frac{\nu}{1-\nu}\right)\right) , \\ \nu = (\xi \cdot 0.5) + 0.5 \, , \\ \xi = cos(\langle 1_{\mathcal{P}}^T, z_i \rangle, M^T) \, , \end{gathered} \tag{6.4}$$

where $1_{\mathcal{P}}$ is the row vector of dimension $\mathcal{P}$ with the scalar value 1, and $\langle .,. \rangle$ is the inner product.Finally, a Softmax function, $\sigma(.)$, is applied to A to force the value to sum-to-one, guarantying the (6.4) satisfied the ANC and ASC. Consequently, the fractional abundance of each pixel, denoted as $a_i \in \mathbb{R}^{1\times\mathcal{P}}$, is obtained. This similarity-based approach allows the framework to capture the relationship between the latent space Z representations and the endmembers M following a logit function before applying the Softmax, noted as $\tilde{a} \in \tilde{A}$, see Fig. 6.4.

The use of the $\sigma(.)$ function, specifically the Softmax function commonly used in DL-based unmixing models [307, 309, 306], imposes a notable constraint on the estimation process. Softmax ensures that the abundance map represents a mixture of various endmembers, thereby complicating the identification of instances where a hyperspectral pixel is represented by a singular endmember

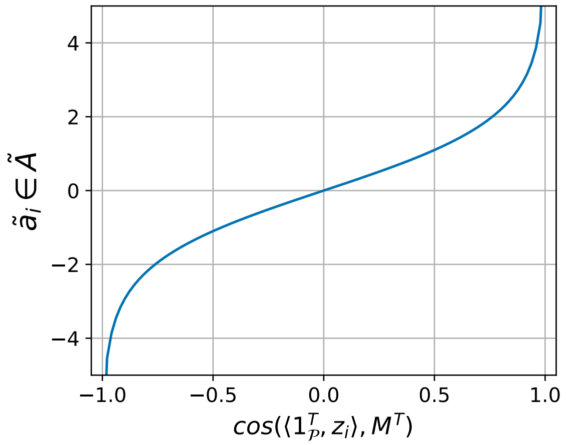


Figure 6.4: Before applying a Softmax for obtaining the abundance map, $\tilde{A}$, the similarity relationship between Z and M follows a logit function.

contribution in A.

Endmember Estimation and Gradient Update

In the proposed CLHU framework, which is designed for blind HSU, the estimation of endmembers is an essential task. In this framework, the endmembers M are treated as model parameters and are initialized using a predefined set of endmembers denoted as M_0. The results obtained are dependent on this initialization and, for this reason, it is recommended to use an endmember extraction algorithms such as the introduced in Section 6.1.2, such as N-FINDR or VCA. The aforementioned methods are based on simplex volume maximization as a search criterion, providing a suitable starting point for the CLHU framework, and are recommended for obtaining interpretable results as it is expected in blind HSU.

The model parameters θ_M, that illustrate the endmembers M, have two interactions in the workflow of the CLHU, particularly in the *Similarity* module and the reconstruction process based on LMM as it is illustrated in Fig. 6.3. However, the gradient estimation for updating the parameters that represent the M are exclusively derived from the LMM, see Eq. (6.1), and the abundance matrix A. Specifically, the gradient with respect to M can be computed as follows:

$$\frac{\partial \mathcal{L}}{\partial M} = \frac{\partial \mathcal{L}}{\partial \hat{X}} \frac{\partial \hat{X}}{\partial M} = \frac{\partial \mathcal{L}}{\partial \hat{X}} A \,,$$

where $\hat{X}$ was estimated by LMM as illustrated in Fig. 6.3, see Eq. (6.1). The objective function $\mathcal{L}$ measures the quality of the reconstruction and its details can be found in Section 6.2.2. This gradient is then used to update the endmember matrix M during the optimization process within the CLHU model. The *Similarity* module indirectly contributes to the estimation of endmembers by influencing the optimization of the abundance matrix A.

6.2.2 Objective function

The HSU task can be interpreted as a reconstruction problem, where the objective is to obtain the best reconstruction, denoted as $\hat{X}$, of the input X. Given the interest in the interpretability of data in HSU, the use of regularization factors is necessary for enhancing the injectivity and stability of this inverse problem, as it has been described in previous sections. Considering the purpose, the loss function defined for CLHU is given by:

$$\mathcal{L} = \frac{1}{2}\|X - \hat{X}\|_F^2 + \mathcal{L}_{Con}(M) + \lambda\frac{\mathcal{L}_{Vol}(M)}{\phi}\,. \tag{6.5}$$

Here, the first term represents the widely used reconstruction error between the input X and the reconstruction based on LMM, denoted as $\hat{X} = MA$. This reconstruction approach, widely used in HSU algorithms [307], is commonly employed for estimating the abundance matrix A.

As it is observed, two regularization terms are depicted in Eq. (6.5), although the second one is considering optional, and it is inspired by geometrical approaches in HSU. The first regularization term, $\mathcal{L}_{Con}(M)$, promotes the contrastiveness of the endmembers in the endmember matrix M, in other words, increasing the independence among the different endmembers. The second regularization term, $\lambda\mathcal{L}_{Vol}(M)$, encourages the minimum simplex volume constraint in M, which is usually considered for endmember estimation in unmixing algorithms [312, 301]. The normalization factor ϕ is introduced to control the scaling of the volume to maintain a desired range or magnitude.

The contrastive regularization factor, $\mathcal{L}_{Con}(M)$, as it was proposed in the NT-Xent contrastive loss function applied in SimCLR [80] and proposed in [85], focuses on minimizing the CE among endmembers. Using this regularization factor, the CLHU framework encourages the endmembers to be distinct and dissimilar from each other. This regularization term promotes the diversification of the endmember estimates by penalizing similarity and encouraging large dissimilarities among the endmembers. Consequently, it helps to prevent redundancy in the endmember set, M. The underlying representation of X in the latent space Z is impacted by this contrastive approach, as it was mentioned before because of the *Similarity* module, and, consequently, this contrastive approach also influences the estimation of abundances A. This regularization factor is defined as follows:

$$\mathcal{L}_{Con}(M) = -\frac{1}{\mathcal{P}}\sum_{\hat{e}\in S}\mathbb{1}(\hat{e} = c)\log\hat{e}\,, \tag{6.6}$$
$$S = \frac{1}{\tau_c}cos(M, M^T)\,,$$

where S is the similarity matrix of the endmembers M and the contrastive regularization factor is defined by the negative log-likelihood where $\mathbb{1}(.)$ is the indicator function. The temperature parameter, $\tau_c \in (0, 1]$, is for scaling the logits in S which it is mapped as a probability value using a Softmax function.

The second regularization factor, $\mathcal{L}_{Vol}(M)$, imposes a constraint on the volume of the simplex formed by the endmembers in M. This regularization is employed to reduce the search space for the endmembers and ensure that they lie on the vertices of a simplex with the smallest possible volume, while maintaining a significant separation from each other [301]. For this reason, in the previous section, it was recommended the endmember initialization by endmember extraction algorithms that are based on simplex volume maximization as a search criterion. This volume constraint regularization factor is given by:

$$\mathcal{L}_{Vol}(M) = \frac{1}{2(\mathcal{P}-1)!} det^2 \begin{bmatrix} 1_{\mathcal{P}} \\ \hat{M} \end{bmatrix} , \tag{6.7}$$
$$\hat{M} = U^T(M - \mu 1_{\mathcal{P}}^T) ,$$
$$U \in \mathbb{R}^{L\times(\mathcal{P}-1)} ,$$

where $1_{\mathcal{P}}$ is the row vector of dimension $\mathcal{P}$ with the scalar value 1 and $\hat{M}$ is the endmember set projected in the low-dimensional space. The orthogonal matrix U is formed by the $\mathcal{P}-1$ most significant principal components of the dataset X, which are obtained through PCA. To address the variability of the simplex area across different datasets, it is suggested to normalize the regularization factor by the initial state, denoted as $\phi = \mathcal{L}_{Vol}(M_0)$ in Eq. (6.5). This normalization ensures that the regularization term is relative to the initial volume of the simplex formed by the initial endmember set M_0.

It is noteworthy that the $\mathcal{L}_{Vol}(.)$, which constraint can also be addressed using distance metrics [313], aligns with HSU's geometrical strategies for enhanced interpretability. However, the relevance of this constraint can be considered optional, and it is particularly relevant in situations where the initial endmember initialization results in a small-area simplex, which can occur when the number of endmembers is relatively high. In such cases, the $\mathcal{L}_{Con}(.)$ constraint would significantly expand the volume. Enforcing this constraint helps in obtaining a set of endmembers that does not change significantly from the signal spectrum present in the hyperspectral image. In addition, given the recommendation of using endmember extraction based on simplex volume maximization, this regularization factor can help to deal with the problems associated with a noisy M_0.

6.3 Information Theoretical Perspective on CLHU

Aligning with the objective of this thesis, improve the interpretability of the DNN giving a explanation more than a 'black-box', in this section it will be discussed the IT perspective of the CLHU framework. As it has been indicated in previous chapters, IT [10] provides a powerful framework for studying the theoretical foundations underlying in DL models.

Understanding the CLHU model as a variant of the AE, a self-supervised method, it naturally aligns with the InfoMax principle [166, 167], Eq. (3.23), as discussed in Section 4.1. Considering the InfoMax principle, the model aims to maximize the MI between an input X and its latent representation Z, obtaining the compression with the highest information rate. In the absence of explicit constraints in the latent space Z, the AE can be expressed as in Eq. (4.6), aligning with the IT perspective of AEs outlined by Yu and Principe [190].

Yu and Principe conceptualize the decoder's role as an 'undo' function, reversing the encoding performed at various stages of the AE encoder [190]. However, as demonstrated in Section 4.3.2, this 'undo' function does not fully apply when the dropout technique is utilized in the encoder. The application of dropout results in potential information loss, as highlighted in Section 3.2.1, even through the decoding process.

Unique to the CLHU framework, and not previously discussed in this thesis dissertation, is the introduction of a secondary information 'channel' within the model's 'bottleneck', depicted in Fig.

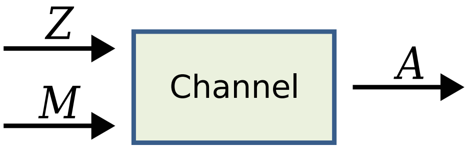


Figure 6.5: Channel of two inputs. Using the reference of the CLHU, the 'bottleneck' information transmission is carried out by two channels AZ and AM.

6.3. Understanding the theoretical implications of this dual-channel approach within the CLHU model is crucial for offering an IT interpretation of its problem-solving mechanism.

6.3.1 InfoMax Constraint in CLHU

Considering the information flow in the CLHU, and understanding that the decoder task consists on 'decoding' the encoded message, it is intuitive to understand that A, discarding the Softmax function at the end, must contain the information on the latent space Z. However, if the dropout technique is applied and following the information rate theory, this property will not be guaranteed. The way to deal with this problem is simple, exploiting redundancy among different channels for minimizing the information loss. This is a feature presented in CLHU as it is illustrated in Fig. 6.5, where there are two channels that transmit information to A, the channel AZ that corresponds to the contribution of Z in A and AM that provide information regarding M.

The MI in A regarding M and Z, $\mathcal{I}(A;M,Z)$, can be reformulated to considering the contribution of information per channel. As a result, see Appendix A.3.1, it is described as:

$$\begin{aligned}\mathcal{I}(A;M,Z) &= \mathcal{I}(A;M) + \mathcal{I}(A;Z|M) = \\ &= \mathcal{I}(A;Z) + \mathcal{I}(A;M|Z),\end{aligned} \tag{6.8}$$

where the first term represents the amount of information contributed by a channel and the second term is the additional information provided by the second channel.

Equation (6.8) outlines the objective of the *Similarity* module, as detailed in Section 6.2.1. The goal here is to enhance the additivity of information across both channels, ensuring that each channel contributes additional, distinct information relative to the other. As a result, $\mathcal{I}(A;M,Z) \geq \mathcal{I}(A;Z)$ with equality if $\mathcal{H}(A|Z) = \mathcal{H}(A|Z,M)$, meaning that M does not provide additional information.

Given the detailed explanation, it is evident that the constraint in the InfoMax Principle, denoted as $\mathcal{R}(Z)$ in Eq. (3.23), is not solely dependent on Z in this case; the inclusion of M is also imperative. Consequently, the constraint applicable to the latent representation should be appropriately denoted as $\mathcal{R}(Z,M)$. Regarding A, since the *Similarity* module (Section 6.2.1) operates as a non-trainable function that delineates the relationship between Z and M, it does not necessitate inclusion in the constraint formulation.

Formulation of the constraint $\mathcal{R}(Z,M)$

Following the InfoMax principle, in contrast to the AEs where the $\mathcal{R}(Z)$ is not always explicitly defined, it has been identified that CLHU has a specific constraint defined. This constraint definition can be originally deduced because of the use of regularization to address the ill-posed problem [119] in the loss function, Eq. 6.5. In the context of HSU, where the primary goal is oriented to data interpretation, it becomes crucial to explicitly incorporate $\mathcal{R}(.)$ to guide the learning process.

Given the description in the previous section, it is intuitive to understand that the constraint can be defined by $\mathcal{I}(A; M, Z)$. In essence, the objective is to maximize the information propagated to the 'decoder' of the CLHU, which is used for the abundance estimation. Given the expression of MI, $\mathcal{I}(A; M, Z) = \mathcal{H}(A) - \mathcal{H}(A|M, Z)$, and knowing that $\mathcal{H}(A)$ is constant since *Similarity* module does not contain trainable parameters, then maximizing the MI can be reformulated as minimizing the conditional entropy:

$$\underset{\theta}{\text{argmax}}\ \mathcal{I}(A; M, Z) \Leftrightarrow \underset{\theta}{\text{argmin}}\ \mathcal{H}(A|M, Z)\,,$$

where θ are the parameters of the DL model.

Following the description of Eq. (3.6), it can be concluded that reducing the $\mathcal{H}(A|M, Z)$ corresponds to a reduction of the $\mathcal{H}(M, Z) - \mathcal{H}(A, M, Z)$. Given that, as it has been previously mentioned, A is fixed by the *Similarity* module, it can be concluded that $\mathcal{R}(Z, M) = \mathcal{H}(Z, M)$. As a result, CLHU is described by the following InfoMax principle:

$$\underset{\theta}{\text{argmax}}\ \mathcal{I}(X; Z) + \beta\mathcal{H}(Z, M). \tag{6.9}$$

In summary, Eq. (6.9) aims to enhance the information within the combined set of Z and M. As indicated by Eq. (6.8), with Z and M serving as compressed representations of the input X, it is essential that both variables contribute unique information not present in the other.

Reformulation as Information Bottleneck principle

Equation (6.9) provides an interpretation of the CLHU framework aligned with the InfoMax principle, detailed in Section 3.5. Nonetheless, this equation can also be adapted to align with the IB principle.

As it is observed in Eq. (6.9), the CLHU has a constraint that consists of increasing the $\mathcal{H}(Z, M)$, increasing the information contained by the compressed latent space Z and the endmember set M. However, it might be expressed as a minimization in the MI between both variables, $\mathcal{I}(Z; M)$, as inferred from Eq. (3.6) and Eq. (3.10). Consequently, Eq. (6.9) can be reformulated to reflect the IB principle as:

$$\underset{\theta}{\text{argmax}}\ \mathcal{I}(X; Z) - \beta\mathcal{I}(Z; M). \tag{6.10}$$

This reformulation directly correlates with the IB principle depicted in Eq. (3.22). In this way, the proposed framework achieves statistical properties desired for the unmixing problem on the representation Z.

Keys insights from the IT Perspective

Given the problem-solving mechanism of the CLHU by IT perspective, both expression, Eq. (6.9) and Eq. (6.10), allow obtaining a better interpretation about how CLHU is working. Although both expression has the same meaning, the different descriptions allow obtaining a better interpretation of the different variables Z and M.

Considering Eq. (6.9), offers insights into the functionality of the latent space Z and the *Similarity* module within the CLHU framework. As it is observed in this equation, the constraint is to maximize the $\mathcal{H}(Z;M)$ and this task has to be carried out by the *Similarity* function. For that reason, using the proposed *Similarity* and observing Fig. 6.4, it is intuitive to understand that, given the logit function, the entropy maximization of this relationship is found where ξ in Eq. (6.4) is around 0. This condition suggests that the lowest $\mathcal{H}(Z,M)$ - indicative of maximal information sharing between Z and M - occurs in situations where certain endmembers have minimal or overly significant contributions to Z.

Given the relationship between Z and M, Eq. (6.9) formulates the CLHU constraint as a minimization in $\mathcal{I}(Z;M)$. As it has been previously discussed, the optimal solution is obtained by a description of Z where there is not a predominant endmember that contributes to the signal representation. In this way, reducing $\mathcal{I}(Z;M)$ depends on M. In this way, this task is constrained by the $\mathcal{L}_{Con}(.)$ in Eq. (6.5), aiming to minimize the MI among the elements of M, ideally achieving $\mathcal{I}(m_1;\ldots;m_{\mathcal{P}}) = 0$ in an optimal scenario where the different endmembers are independent.

Both description fits with the understanding of input X as a convex combination of M based on LMM. If it is considering that $m_i \in M$ are independent, they hold no information about each other, minimizing MI among them, thereby the information you can obtain about X from knowing a single component m_i is limited. In that case, Z would be a compressed representation of X different endmembers contributing equality, increasing the entropy $\mathcal{H}(M;Z)$.

Finally, following the IB principle formulation in Eq. (6.10), the latent space Z in CLHU can be interpreted as a trade-off between a compression of X preserving as much as possible the information on M. For that reason, this problem is heavily dependent on M initialization, M_0. This description fits with the aforementioned explanation, where the idea is to deal with spectral variability induced by external factors that does not belong to the endmembers interaction.

6.3.2 Data Processing Inequality

As it has been pointed out in previous chapters, due to the learning mechanism of DL models, where the input signals are propagated from one layer to another until they reach the output layer and then errors are back-propagated through all layers in reverse order, this can be interpretable as a Markov process. However, both propagation are unidirectional and depend exclusively on the previous variable, such as a Markov chain [191, 190]. In other words, the output of the i^{th}-layer in the DL model, T_i, depends on the output of the previous layer, T_{i-1}.

As a consequence, DNN are characterized as cascaded channel, see Section 3.2.1, where there is loss of information as it is depicted in the the DPI, Eq. (3.17). As a result of a ideal AE would be satisfied by Eq. 4.3 in the encoder and Eq. 4.4 in the decode path. However, as it has been previously reported, the use AEs with common techniques for avoiding the overfitting force the Eq. (4.7) that is a restriction in the decode path. In other words, the decode cannot recover more information than the contained in Z.

As it has been described in Section 6.3.1, the use of two variables, Z and M, has a special purpose from a IT-perspective, recovering information that it is not contained in Z. As a consequence, A estimation by the *Similarity* module in CLHU - that can be considered as part of the decoder - has to contain more information than Z exclusively. In other words, the introduction of the random variable defined by M serves the purpose of incorporating information that aids in the reconstruction of the original signal.

In the optimal solution, where the endmember are well estimated by CLHU and the reconstruction model, the LMM in this case, fits with the data, the $\mathcal{I}(X;A) \simeq \mathcal{H}(X)$. The information lost in the sequence of layers of the CLHU model has to be recovered by A, which contains information about how the different endmembers interact. As a result, CLHU should satisfy the following DPI:

$$\mathcal{H}(X) \geq \mathcal{I}(X;A) \geq I(X;Z). \tag{6.11}$$

6.3.3 Approximation to Independent Components Analysis

It is possible to approximate the CLHU framework to the ICA [167] in which the InfoMax principle, Eq. (3.23), is also applied. ICA is a statistical technique used to extract independent sources from a set of observations that are linear mixtures of those sources, and it has been previously used for HSU [314, 315, 316]. However, ICA has been questioned as an approach to HSU [317], given that the ASC in abundance map implies statistical dependence, but recent works have demonstrated that this method is competitive for this task [318].

The CLHU framework can be approximated as an ICA method, where the goal is to extract the underlying structure or patterns in X by representing it as a linear combination of independent components, which are defined by the variable M in the unmixing task. Furthermore, it is important to note that the quality of the representation in Z and the estimation of A depend on the selected number of independent components.

On the one hand, if the number of independent components is not accurately estimated, the resulting space will not be able to capture all the underlying independent sources of variability in the data, which may result in a loss of information in the reconstructed signals. On the other hand, if the number of independent components is overestimated, the resulting space may include some unnecessary or redundant sources of variability, which may also lead to a less informative representation. Therefore, the accurate estimation of the number of independent components is crucial for obtaining an effective and informative representation of the data.

In ICA, one of the well-known approaches for estimation is to minimize the MI between the independent components. This approach aims to decrease the mutual dependence between the independent components and the input data. As a result, the MI between the independent components and the input data should decrease. In light of this, it can be considered as follows:

$$\mathcal{I}_0(Z;M) \geq \mathcal{I}_T(Z;M)\,, \tag{6.12}$$

where $\mathcal{I}_0(Z;M)$ is the initial state of the latent space and $\mathcal{I}_T(Z;M)$ at the end of the training.

6.4 Experimental Results

In this section, the CLHU framework has been subjected to experimental evaluation using various datasets. Since the state-of-the-art dataset for remote sensing lacks a real ground truth, a synthetic dataset was initially employed to evaluate the performance of CLHU. Subsequently, the proposed framework was tested on different datasets from the state-of-the-art. As a proof of concept, a hyperspectral image comprising similar biological tissues was employed to demonstrate the utility of the proposed framework in medical applications.

The proposed framework is quantitatively evaluated by comparing the estimated results with the ground truth reconstruction, denoted as Y, which is obtained from the true endmembers and abundance maps provided by the dataset. The evaluation metrics used in this study include the Root Mean Squared Error (RMSE) between the estimated reconstruction, $\hat{X}$, and ground truth reconstructions, Y (RMSE$_Y$), the RMSE between the estimated and ground truth abundance fractions (RMSE$_A$), and the SAD in radians, which measures the difference between the estimated and ground truth endmembers (SAD$_M$).

To ensure unbiased conclusions, the same configuration of hyperparameters was used across all experimental results. The details of the hyperparameter configuration can be found in Section 6.4.

Hyperparameters configuration

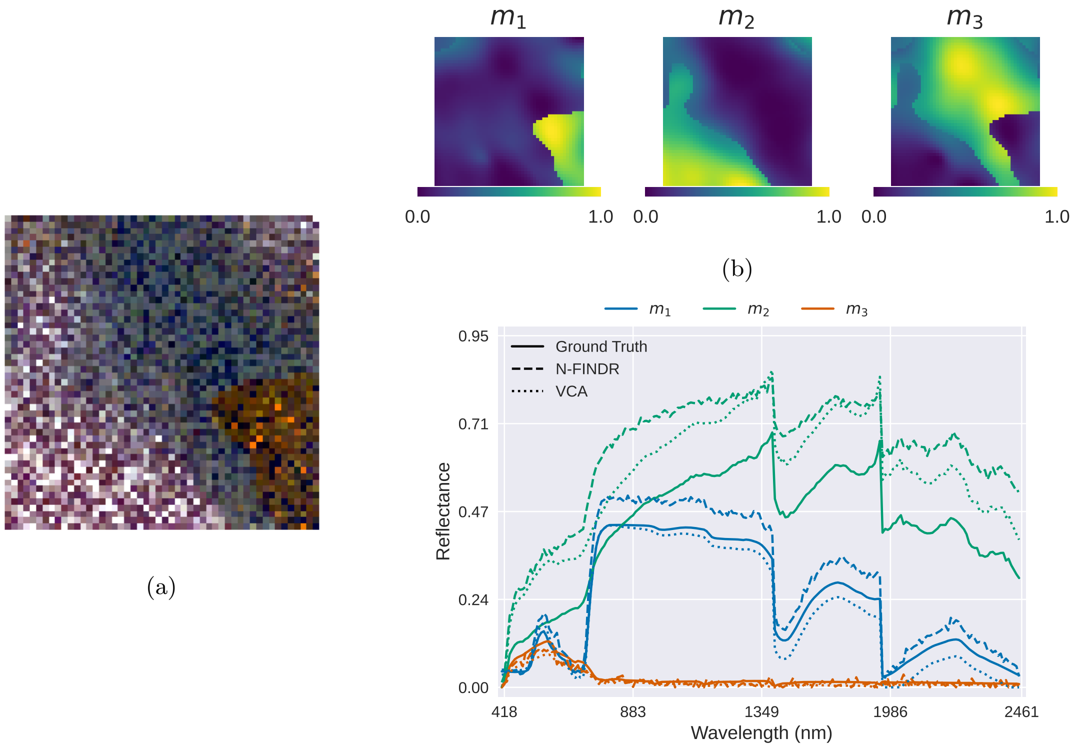
In DNN-based models, the obtained results are typically dependent on the hyperparameter settings. In this work, the proposed encoder of the CLHU-based model, denoted as $f(.;\theta_e)$, is composed by three sequential fully connected layers with 512, 128 and 32 neurons, respectively. Each layer is followed by batch normalization [224] and a dropout layer, which is a regularization technique commonly used in DL models to reduce overfitting [73] and a ReLU activation function. The projection head, denoted as $f(.;\theta_p)$, constitutes a fully connected layer that lacks bias and activation function. It can be represented as a common projection matrix. As for the objective function $\mathcal{L}$, the Lagrange multiplier λ is set to 10^{-2}, as depicted in Eq. (6.5). For the contrastive loss function $\mathcal{L}_{Con}(.)$, as presented in Eq. (6.6), a temperature value of 0.1, which is denoted as τ_C, is employed. Lastly, the ADAM optimizer [65] was utilized to train the CLHU-based model with a learning rate of 10^{-3}. The exponential decay rates for moment estimation, represented by β_1 and β_2, were configured to be 0.9 and 0.999, respectively. Finally, the number of epochs is fixed to 50 and the batch size is modified per dataset to obtain a fixed number of batches, 50.

To ensure unbiased conclusions, the same configuration of hyperparameters was used for the different experimental results presented in this section. This approach was adopted to avoid any potential bias in the conclusions due to well-fitted settings for the indicated datasets.

6.4.1 Synthetic Dataset

The Jasper Ridge, Samson, and Urban datasets are often used to assess the endmember extraction and abundance estimation methods for HSU [319]. However, the ground truth provided in those datasets is estimated applying state-of-the-art algorithms, and a measurement to validate the spectral mixture has not been performed. Therefore, the goal of generating a realistic synthetic dataset has been considered for a first validation of the CLHU framework.

The spectral variability, as it has been previously discussed, is produced by many factors such as temporal condition or illumination and it is one of the main challenges from the spectral signature perspective. Borsoi *et al.* [285] generated a synthetic endmember variability dataset, according to simplifications of some radiative transfer models, for evaluating different state-of-the-art algorithms for HSU. The synthetic dataset is composed of three materials: soil, vegetation and water. In order to generate soil samples, a simplified version of Hapke's model [321] was applied. The vegetation endmembers were generated by the PROSPECT-D model [322], according to different



(a)

(b)

(c)

Figure 6.6: Synthetic dataset and ground truth generated with an SNR of 30 dB. In (a), the HSI cube is shown in RGB based on [320]. The abundance maps and endmembers are depicted in (b) and (c). The endmembers m_1, m_2, and m_3 represent vegetation, soil and water, respectively. The figure includes the endmember estimation based on two methods: NFINDR, represented by dashed lines, and VCA, represented by dotted lines.

biophysical parameters. Borsoi *et al.* also considered the atmospheric compensation to generate water endmembers according to different viewing angles [285]. Finally, each pixel was created based on the spectral variability that accommodates the LMM described in (6.1). Three configurations were tested by modifying the SNR to apply additive random noise, denoted as ϵ_i. As a result, three hyperspectral cubes were generated with a spatial resolution of 50 × 50 pixels and 198 bands, based on the Jasper Ridge dataset, by applying additive random noise with SNR values of 10, 20, and 30 dB. Figure 4.3 displays the resulting synthetic dataset generated with a SNR of 30 dB, along with the corresponding ground truth. The endmembers estimation using N-FINDR and VCA is also illustrated in Fig. 4.3(c).

In order to evaluate the CLHU framework, LMM by FCLS (LMM-FCLS) [300], which uses fixed endmembers, and ELMM [284] were considered as references. From this point forward, we will refer to the method LMM-FCLS as FCLS. Both methods are well-known and commonly used [323, 324, 325, 326], specially ELMM has demonstrated to achieve promising performance in the context of non-linear HSU [285]. Given that the ground truth of the synthetic data is linear, both state-of-the-art models are capable of obtaining the optimal solution. As explained in Section 6.1.1, the computational approach of ELMM yields distinct endmembers for each sample, which are modifications of the initial endmember M_0 given the scaling factor ϕ_i, see Eq. (6.2). In contrast, the

Table 6.1: Quantitative results across different configuration of noise in the synthetic dataset (average of 10 executions and standard deviation for the CLHU proposed method). In the experiment, two methods for endmember initialization were considered: N-FINDR and VCA. Since FCLS does not update the initial endmembers M_0, it can serve as a baseline for evaluating how much the CLHU model improves the endmember estimation.

	M_0	N-FINDR			VCA		
SNR (dB)	Metrics	CLHU	FCLS	ELMM	CLHU	FCLS	ELMM
30	$RMSE_Y$	0.0325 ± 0.0011	0.0400	0.0354	$\mathbf{0.0318 \pm 0.0012}$	0.0356	0.0337
	$RMSE_A$	0.1115 ± 0.0055	0.1284	0.1130	0.0892 ± 0.0035	0.0923	**0.0864**
	SAD_M	0.0779 ± 0.0007	0.1277	0.1097	$\mathbf{0.0721 \pm 0.0008}$	0.0965	0.0804
20	$RMSE_Y$	0.0308 ± 0.0009	0.0420	0.0359	$\mathbf{0.0289 \pm 0.0004}$	0.0329	0.0332
	$RMSE_A$	0.1087 ± 0.0056	0.1280	0.0682	$\mathbf{0.0636 \pm 0.0015}$	0.1005	0.0698
	SAD_M	0.0874 ± 0.0005	0.1979	0.1576	$\mathbf{0.0549 \pm 0.0008}$	0.0801	0.0766
10	$RMSE_Y$	0.0368 ± 0.0019	0.0583	0.0486	$\mathbf{0.0311 \pm 0.0004}$	0.0326	0.0467
	$RMSE_A$	0.1284 ± 0.0043	0.1588	0.1053	$\mathbf{0.0717 \pm 0.0019}$	0.0924	0.1448
	SAD_M	0.2309 ± 0.0012	0.4034	0.3345	0.1585 ± 0.0015	**0.1580**	0.1635

FCLS method assumes a single fixed endmember for all samples. Similar to the FCLS, the CLHU framework has a solution with a set of fixed endmembers, but CLHU includes an M estimation process that FCLS lacks.

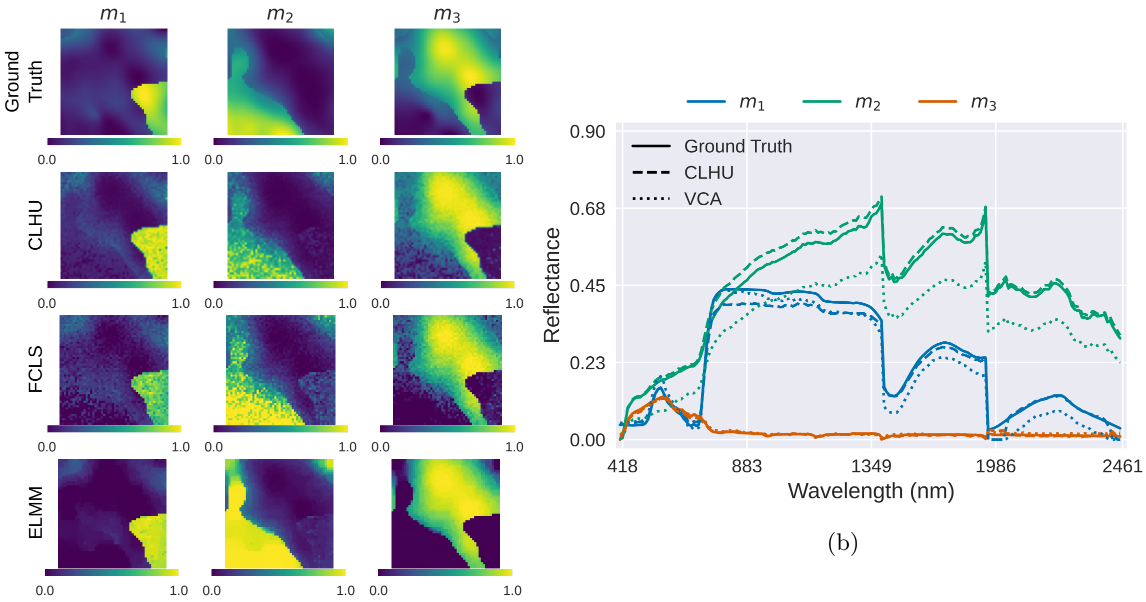
In the different methods, it is necessary to provide an initial set of endmembers, denoted as M_0, which can be estimated using different techniques. In this study, N-FINDR and VCA (see Section 6.1.2) are utilized as endmember extraction techniques. One significant difference between the two methods is that N-FINDR is known to be more sensitive to noise present in the hyperspectral image. As a consequence, it generally exhibits slightly lower performance compared to VCA when dealing with noisy datasets. The outcome of both methods can be observed in Fig. 6.6c. Both state-of-the-art methods, FCLS and ELMM, have been evaluated using both endmember estimation techniques as depicted in Table 6.1.

As mentioned in Section 6.2.1, the endmember initialization M_0 for the CLHU-based model was established through the methods that guarantee the maximum volume of the simplex formed by the endmembers. In addition, it aligns with the concept of the regularization factor $\mathcal{L}_{Vol}(.)$, see Eq. (6.7) for dealing with a noisy M_0, as it was described in Section 6.2.2. In the experimental evaluation, CLHU was tested using both N-FINDR and VCA as endmember initialization methods, see Section 6.1.2.

The quantitative results in the synthetic dataset can be found in Table 6.1. Since the CLHU-approach is a stochastic method, the value presented in Table 6.1 was obtained by conducting the CLHU-approach experiment 10 times and computing the average value and the standard deviation.

As previously mentioned, the FCLS and ELMM approaches are sufficiently effective in obtaining an optimal solution for the synthetic dataset (Table 6.1). The proposed CLHU model is comparable to the state-of-the-art models across different level of noise in the reconstruction and abundance estimation, RMSE$_Y$ and RMSE$_A$. The endmember initialization, M_0, is the differentiating factor of the results, providing VCA better results, as expected.

Observing Table 6.1, it is noticeable that the metrics are quite similar across the different meth-



(a)

(b)

Figure 6.7: Qualitative results for HSU on the synthetic dataset. In (a), it is illustrated A estimation in the synthetic dataset at a SNR of 10 dB. In (b), M estimation in the synthetic dataset at a SNR of 20 dB. In both case, The results have been obtained by using VCA for endmember initialization.

ods, especially when using VCA for M_0 estimation. One notable difference can be observed in the case where the dataset has a SNR of 10 dB and VCA is used as the endmember initialization method. In this particular case, the ELMM exhibits the worst estimation performance based on the RMSE$_A$ metric in Table 6.1. On the other hand, both the CLHU and FCLS methods remain competitive and show comparable performance in estimating the abundance fractions, with the CLHU method slightly outperforming the other. This observation is further supported by the qualitative evaluation shown in Fig. 6.7a, where it can be visually noticed that the CLHU and FCLS methods stand out in estimating the fractional abundances compared to ELMM.

Regarding the endmember estimation, the proposed model outperforms the other methods, surpassing the original M_0 initialization in both cases, whether N-FINDR or VCA is used. This improvement is particularly significant when using N-FINDR for endmember initialization. The SAD$_M$ in the ELMM method is estimated by the average signal obtained from the different samples, M_i, see Eq. (6.2). In the case of the FCLS method, the M is fixed, estimating A by M_0, so it can serve as a baseline for evaluating how much the CLHU model improves the endmember estimation by N-FINDR and VCA. The CLHU model initialized by N-FINDR can obtain a better M estimation than the rest of the models using VCA in the case of SNR 30 dB (Table 6.1). Figure 6.7b provides a qualitative comparison, illustrating how CLHU improves the endmember matrix M estimation compared to the initial estimation obtained from VCA in 20 dB. In this figure, a clear improvement can be observed in the estimation of m_2 compared to the initial estimation based on VCA.

Finally, considering the $RMSE_Y$, the CLHU approach is robust to the noise level, having a good reconstruction in the different noise configurations and with a low standard deviation (Table 6.1). This indicates that the CLHU model is consistent, obtaining a practically identical output in different training sessions.

As a conclusion, the proposed CLHU framework shows a consistent performance across the different noise configurations and the estimated endmembers consistently improve from the initial estimation M_0.

6.4.2 State-of-the-art dataset

In the previous section, the CLHU performance has been evaluated using a synthetic dataset which ground truth is known. However, the HSU algorithms are commonly evaluated with state-of-the-art datasets, which ground truth is not well-defined, and it has been estimated by different HSU methods [319]. For this reason, the results obtained in this section should be interpreted as indicative rather than definitive. Taking this into consideration, in this section, the performance of the CLHU framework will be assessed using three remote sensing hyperspectral images that are widely used in the literature.

The Jasper Ridge dataset, is a hyperspectral image of 224 spectral bands with a spectral resolution of up to 10 nm, ranging from 400 nm to 2500 nm. In HSU, a sub-image of 100 × 100 pixels and 198 spectral bands is used, where the noisy bands contaminated with water vapor density and atmosphere are removed and a labeled ground truth is provided. This hyperspectral image includes four endmembers: *vegetation, water, soil*, and *road* [319].

The Apex dataset, which has been previously used for HSU [311], is a hyperspectral image which contains 110 × 110 pixels and 285 spectral bands that cover the wavelength range from 413 to 2420 nm. In contrast to the Jasper Ridge dataset, Apex has not been preprocessed to reduce the number of bands. Therefore, the bands contaminated with water vapor density and atmosphere have not been removed. This hyperspectral image includes four endmembers: *road, vegetation, roof,* and *water*.

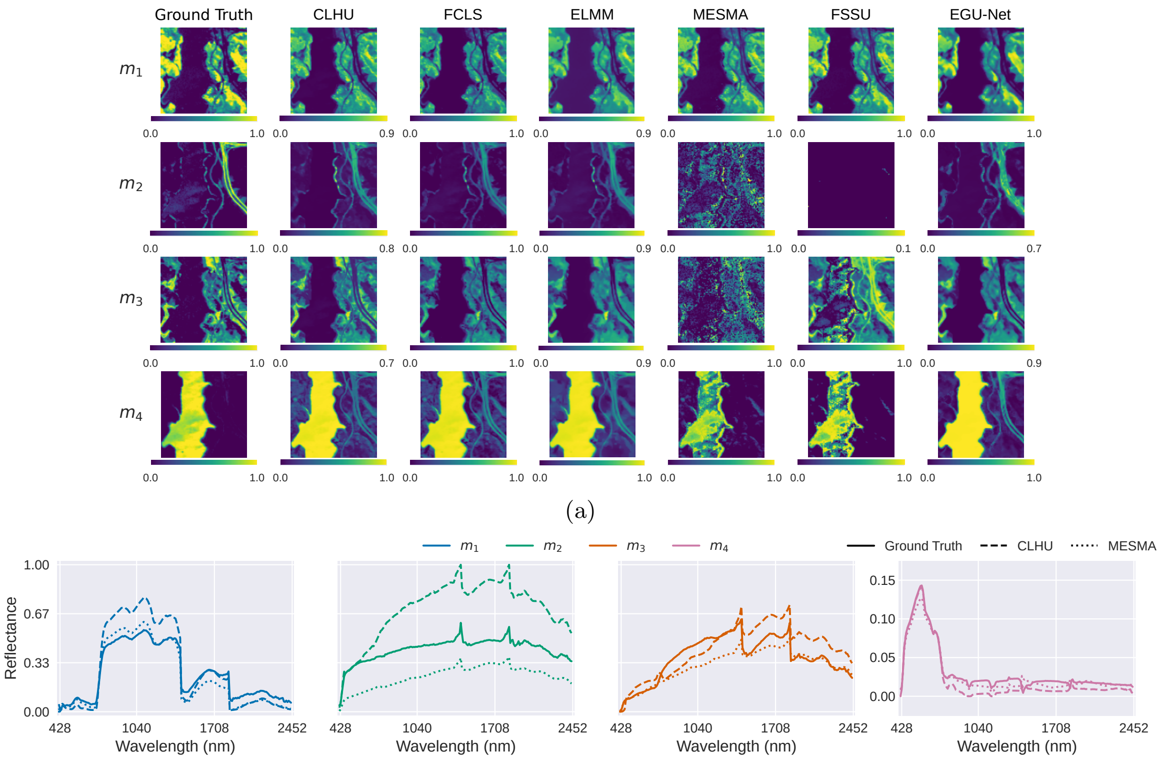
The Samson hyperspectral dataset [327] contains 95 × 95 pixels captured over 156 different spectral bands in the wavelength range from 401 to 889 nm. This image contains three endmembers representing *soil, tree,* and *water*. The endmember spectra were manually selected from the image to provide ground-truth information [327].

Since it cannot be guaranteed that FCLS and ELMM are sufficiently effective in obtaining an optimal solution in these hyperspectral images, two state-of-the-art methods for HSU have been included for comparison: the Multiple-Endmember Spectral Mixture Analaysis (MESMA) [328] and Fractional Sparse Spectral Unmixing (FSSU) [329]. The MESMA algorithm and its variants are computationally demanding optimization methods that achieve good quality in HSU by using a library for the different endmembers, as ELMM. the FSSU extends the MESMA method by incorporating a mixed norm technique.

Furthermore, the EGU-Net [310], an AE model for HSU, was also employed (see Section 6.1.3). Notably, the EGU-Net is a modern DL-model approach that does not explicitly impose spatial information through its methodology, like our proposed CLHU. This distinction sets it apart from other DL-based methods where spatial information is integrated via convolutional layers, such as in the case of UnDIP.[307]. However, it is essential to emphasize that EGU-Net is exclusively designed for abundance map estimation. In contrast to our proposed method, endmember estimation is not a task handled by the DNN in the EGU-Net. In their approach, VCA is used to extract a subset of pseudo-pure endmembers, which are labeled by another HSU method [310]. This labeling is crucial because the encoder, utilized for abundance map estimation, is trained by two streams. One stream

Table 6.2: Quantitative results across different configuration of state-of-the-art datasets. In the experiment, two methods for endmember initialization were considered: N-FINDR and VCA. As EGU-Net methodology depends on a subset of pseudo-pure endmembers extracted by VCA, the N-FINDR has not been applied in this case.

		Jasper Ridge			Apex			Samson		
M_0	Method	$RMSE_Y$	$RMSE_A$	SAD_M	$RMSE_Y$	$RMSE_A$	SAD_M	$RMSE_Y$	$RMSE_A$	SAD_M
N-FINDR	CLHU	0.0460 ±0.0005	0.1776 ±0.0031	0.1555 ±0.0007	0.0685 ±0.0029	0.2691 ±0.0025	0.1856 ±0.0005	0.0331 ±0.0005	0.1890 ±0.0022	0.3799 ±0.0026
	FCLS	0.0471	0.1588	0.1604	0.0672	0.2731	0.1512	0.0209	0.1397	0.0946
	ELMM	0.0461	0.1605	0.1527	0.0677	0.2798	0.1500	0.0225	0.1490	0.0894
	MESMA	0.0437	0.2418	0.1927	0.0493	0.2900	0.3998	0.0177	0.0899	0.0851
	FSSU	0.0440	0.2438	0.1915	**0.0491**	0.4115	0.4010	0.0180	0.0852	0.0841
VCA	CLHU	**0.0424** **±0.0019**	0.1648 ±0.0027	0.1845 ±0.0009	0.0682 ±0.0027	0.2771 ±0.0028	0.1230 ±0.0007	0.0259 ±0.0062	0.1828 ±0.0100	0.2906 ±0.1014
	FCLS	0.0451	0.1559	0.1521	0.0670	0.2801	**0.1089**	0.0206	0.1387	0.0901
	ELMM	0.0451	0.1605	0.1495	0.0675	0.2847	0.1162	0.0223	0.1504	0.0878
	MESMA	0.0435	0.2457	**0.1479**	0.0540	0.1896	0.3422	**0.0151**	0.0624	0.0784
	FSSU	0.0438	0.2354	0.1528	0.0540	0.2585	0.3450	0.0154	**0.0595**	**0.0780**
	EGU-Net	0.0656 ±0.0025	**0.1466** **±0.0033**	0.2315	0.0823 ±0.0004	**0.1692** **±0.0016**	0.3382	0.1025 ±0.0013	0.1969 ±0.0011	0.1125



(a)

(b)

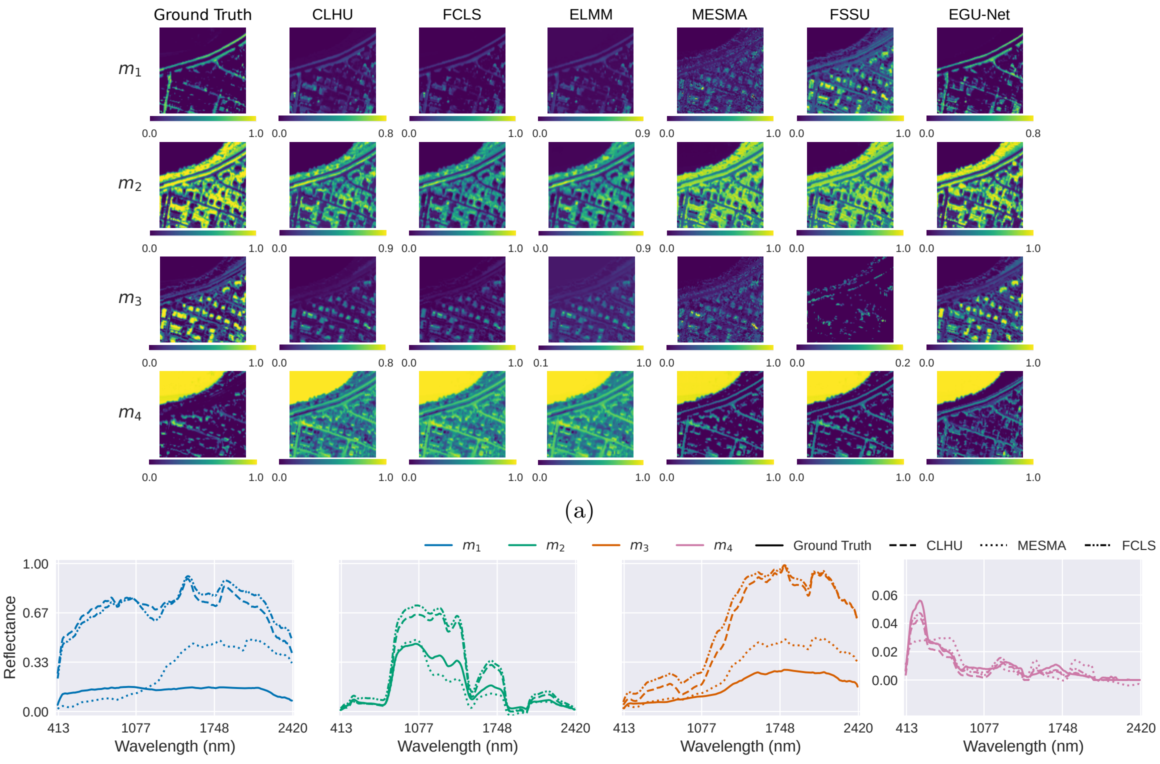
Figure 6.8: Jasper Ridge dataset – Result presented here is obtained by initializing the endmembers by VCA. In (a) is illustrated a visual comparison of A estimation obtained by the different unmixing techniques. In (b) it is shown a comparison of the endmember estimation using CLHU and MESMA. The endmembers m_1, m_2, m_3, and m_4 represent the *vegetation, road, soil,* and *water*, respectively.

optimizes the abundance map of these pseudo-endmembers obtained by another HSU method, while the other stream optimizes the AE as a reconstruction problem using non-labeled pixels.

Given the architectures proposed for EGU-Net, one of them including spatial information [310] by convolutional layers, in this work, the spatial information is not considered. The abundance map of the aforementioned pseudo-pure endmembers has been obtained using Sparse Unmixing by Variable Splitting and Augmented Lagrangian (SUnSAL) [330]. For the quantitative endmember estimation of the EGU-Net, it has been considered as a library method, similar to ELMM, MESMA and FSSU, where the pseudo-pure endmembers serve as the library.

Table 6.2 presents the quantitative results for different datasets. Similar to the synthetic dataset, two methods were used for the endmember initialization, M_0, excluding EGU-Net, in which it is only used VCA for the restriction of the method. Based on the results presented in Table 6.2, it can be observed that the reconstruction, RMSE$_Y$, metric consistently exhibited good performance for all methods and datasets. Therefore, it can be concluded that the reconstructed data is similar to the estimated ground truth data. Upon observing RMSE$_A$ and SAD$_M$ in Table 6.2, it is apparent that the solutions obtained differ from the proposed ground truth. As previously mentioned, HSU is an ill-posed inverse problem, so it is common to obtain multiple alternative solutions for the same input.

The Jasper Ridge dataset yields consistent results across all methods, with the main variation



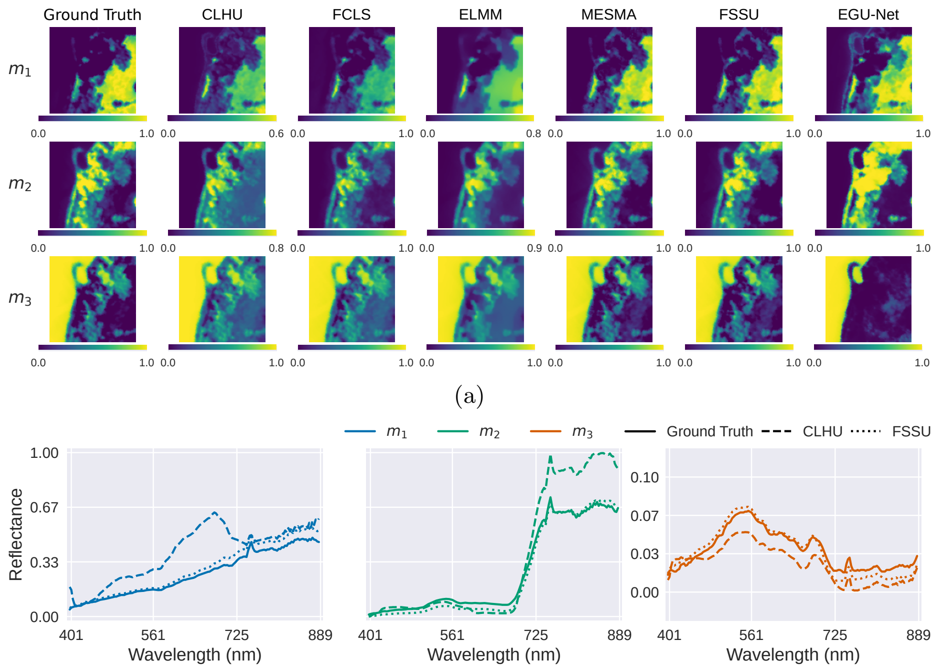
(a)

(b)

Figure 6.9: Apex dataset – Result presented here is obtained by initializing the endmembers by VCA. In (a) is illustrated a visual comparison of A estimation obtained by the different unmixing techniques. In (b) it is shown a comparison of the endmember estimation using CLHU, MESMA and FCLS (VCA). The endmembers m_1, m_2, m_3, and m_4 represent the *road, vegetation, roof* and *water*, respectively.

observed in the RMSE$_A$ metric (Table 6.2). Qualitative results are illustrated in Fig. 6.8b and Fig. 6.8a. Among those methods that rely on endmember libraries, MESMA and FSSU demonstrate a larger deviation in RMSE$_A$ from the estimated ground truth when compared to other methods, and this is confirmed in the qualitative inspection in Fig. 6.8a. At the same time, it can be observed in Fig. 6.8a that the results obtained by CLHU, FCLS, ELMM and EGU-Net are similar, with the latter achieving the best quantitative result. However, in the case of endmember m_3, CLHU shows the best qualitative result. In terms of M estimation, the CLHU model using VCA as the endmember initialization method shows more significant differences to the reference compared to the other methods. This is noticeable in Fig. 6.8b, particularly in the case of endmembers m_2 and m_3, which represent the road and soil, respectively. As expected, since the endmember estimation is not carried out by the EGU-Net, the quantitative result obtained in this method using the pseudo-pure endmembers shows the worst performance in terms of SAD$_M$.

Regarding the Apex dataset, Table 6.2 shows that the abundance estimation by the different methods, excluding the EGU-Net, is notably distant from the proposed ground truth. The MESMA and FSSU methods, both based on endmember libraries, and EGU-Net exhibit significant challenges in accurately estimating the endmembers. It can be observed in Fig. 6.9b and 6.9a that MESMA supplies identical endmembers for the road and roof, namely m_1 and m_3. In both figures, it can



(a)

(b)

Figure 6.10: Samson dataset – Result presented here is obtained by initializing the endmembers by VCA. In (a) is illustrated a visual comparison of A estimation obtained by the different unmixing techniques. In (b) it is shown a comparison of the endmember estimation using CLHU and FSSU, this last one obtained the best SAD_M. The endmembers m_1, m_2 and m_3 represent soil, vegetation and water, respectively.

be observed that MESMA provides a comparable output for the corresponding endmembers. The estimation of A is found to be identical in the qualitative evaluation based on Fig. 6.9a, and similar in Fig. 6.9b, having a poor estimation of both output in MESMA. In addition, Fig. 6.9b also suggests that the identification of m_1 is a challenge for the different methods. The CLHU model appears to struggle with accurately estimating the endmembers corresponding to the road (m_1) and water (m_4), respectively. While the poor estimation of m_1 is understandable due to the difficulties in accurately identifying it, the estimation of m_4 appears promising based on Fig. 6.9b. However, the corresponding A estimation in CLHU is not satisfactory, as happens in FCLS and ELMM. More specifically, the presence of water appears to be observed throughout the entire image (Fig. 6.9a). It is possible that the lack of preprocessing in the Apex image, as it was mentioned before, could generate some issues, particularly with the presence of water vapor, which can cause distortions in the spectral bands. This could result in difficulties for an accurate abundance estimation. However, the methods based on libraries, such as MESMA, look able to deal with this problem (see Fig. 6.9a).

In the specific case of the EGU-Net, it is the only method that obtains an optimal solution regarding the A estimation for the Apex dataset, as depicted in Table 6.2 and Fig. 6.9a. However, the reconstruction and endmember estimation (considering the pseudo-pure endmembers as a library method for comparison reasons) show that this method is only competitive for abundance map estimation.

It is worth noting that the Samson image exhibits a unique behavior in the results obtained by the CLHU model. Specifically, the solution obtained by CLHU differs significantly from the solutions obtained by other methods and also from the estimated ground truth, as it is noted in Table 6.2 and Fig. 6.10b. In the qualitative evaluation, it can be observed that the estimation of A by CLHU, depicted in Fig. 6.10a, does not differ much from the estimation obtained by the other methods. The primary disparity in A estimation can be attributed to the abundance of m_1, which represents the soil. Although the estimated locations of this endmember are consistent across different methods, its composition differs significantly. It appears that in the solution provided by the CLHU method, m_1 is the main contributor for detecting the soil, with the other endmembers also contributing but to a lesser extent. This is easily observable in the colorbar of Fig. 6.10a, where the maximum contribution value is limited to 0.6. As shown in Table 6.2, the SAD_M metric is found to be consistent among the different methods, except for CLHU, whose results differ significantly from the ground truth estimation. The main reason for the considerable difference is found in the m_1 estimation by CLHU, and it can be observed in Fig. 6.10b.

In conclusion, this section presents the results obtained by applying the proposed method to different datasets commonly used in the literature. The performance of the proposed method is compared with other state-of-the-art methods, and the quantitative results are presented in Table 6.2. Overall, the CLHU model shows robust performance and demonstrates that the proposed framework is highly competitive. It is worth noting that all results were obtained using the same configuration of hyperparameters described in Section 6.4 to avoid biased results. Regarding the qualitative evaluation, the CLHU model shows promising results in all experiments, demonstrating a good overall performance. Specifically, in the Samson dataset, the CLHU model shows a solution that differs significantly from the rest of the methods, which is identified in the estimation of the endmember m_1. However, as an inverse problem, this solution cannot be immediately discarded as an adequate solution, especially considering that the ground truth of these datasets is an estimation and there are no actual measurements available for validating the spectral mixture. Insights on endmember estimation will be elaborated in Section 6.4.3.

Comparing with the state-of-the-art DL-based method, EGU-Net, the proposed method provides two main advantages. Firstly, CLHU does not depend on a different method for establishing a reference for the abundance map. Additionally, CLHU includes endmember estimation as a task, something that EGU-Net does not do. These advantages are a consequence of the innovative approach in HSU method based on DL by leveraging the relationship between the input and the endmembers to estimate the abundance map.

6.4.3 Insight in Endmember Estimation by CLHU

As indicated in the previous section, the Samson dataset exhibits a unique behavior in the results obtained by the CLHU model. The A estimation by CLHU does not differ much from the estimation obtained by other methods and the proposed ground truth. However, the primary disparity in A estimation can be attributed to the abundance of m_1, which represents the soil. This is a consequence of a soil endmember estimation that is significantly different from the proposed ground truth.

It is noteworthy that the soil endmember estimation by CLHU exhibits a specific pattern identified in different datasets, particularly in the aforementioned Samson dataset and the Jasper Ridge dataset, where m_3 endmember is labeled as soil (see Fig. 6.8a). Specifically, a peak of absorbance around 700 nm is identified, which is not present in the ground truth, as clearly illustrated in Fig.

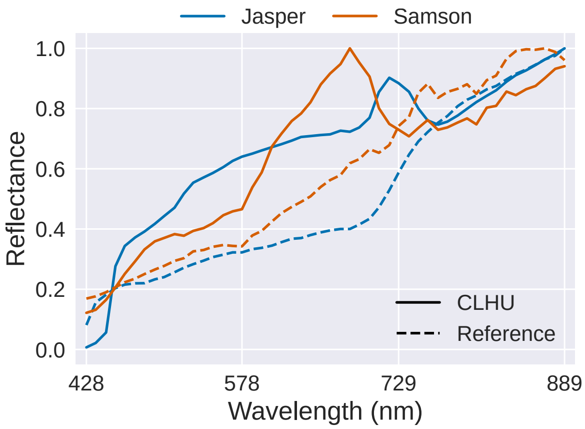


Figure 6.11: Soil endmember estimation by CLHU (solid line) and reference (dashed line) normalized in two independent datasets, Jasper Ridge and Samson.

Table 6.3: Quantitative endmember estimation by SAD_M relative to the Ground-Truth provided by the dataset. In the case of the synthetic dataset, it has been used the dataset with an SNR of 30db. Given $\tilde{M}$ is the endmember estimated by CLHU model $f(.)$, the reconstruction provided for the CLHU is represented by $f(\tilde{M})$.

	SAD_M			
	Jasper	Apex	Samson	Synthetic
$\tilde{M}$	0.1825	0.1100	0.1964	**0.0720**
$f(\tilde{M})$	**0.1055**	**0.1052**	**0.1085**	0.1083

6.10b and less distinctly in Fig. 6.8b. For a direct comparison, Fig. 6.11 illustrates a representation of both endmembers extracted from the different datasets in the same wavelength range. This peak absorbance is observed in Fig. 6.11, where the peak of absorbance is represented by a significant decrease in reflectance.

Observing Fig. 6.11, it is evident that the peak of absorbance is shifted between the two datasets, Jasper Ridge and Samson. This discrepancy might be a consequence of dataset notation or given by the high correlation among adjacent bands, but it cannot be guaranteed. When compared to the reference signal in the illustrated wavelength range (dashed line), the peak of absorbance is not observed. The same discrepancy is observed with other approaches, as illustrated in Fig. 6.8b and Fig. 6.10b, where the endmember estimation is obtained from the library method.

The aforementioned behavior was not observed in the Section 6.4.1, where the experiment is controlled and the endmember are well-known and CLHU has proved his effectiveness in this experiment. As it has been previously reported, the dataset used in this section have not a well-estimated ground truth and the reference has been extracted by different HSU methods [319]. For that reason, the results should be considered as indicative.

As shown in Table 6.2, considering the endmember initialization using VCA, CLHU model appears to yield less favorable results compared to the VCA. For instance, in Jasper Ridge dataset, the original VCA (as seen in the FCLS case) yields a SAD_M of approximately 0.15, whereas in CLHU, this value increases to around 0.18. This trend is consistent across other datasets, with Samson exhibiting the most significant discrepancy.

However, an interesting observation is made when the endmembers estimated by CLHU, denoted as $\tilde{M}$, are 'reconstructed' using the same CLHU algorithm, resulting in $\hat{M} = f(\tilde{M})$, where $f(.)$ represents the CLHU model. This process appears to mitigate the earlier issues. As indicated in Table 6.3, the reconstruction of $\tilde{M}$ by CLHU improves the results and surpasses the better estimations in dataset like Jasper Ridge and Apex. In contrast, for the synthetic dataset with well-known endmembers, the reconstruction quality diminishes, aligning with expectations.

In the absence of more evidence, the result suggests that CLHU could be found non-linear interaction that it is consequence of the interaction among different endmembers and for that reason, there is a bigger discrepancy regarding the endmembers provided by the state-of-the-art dataset. In the synthetic dataset, the linear interpretation is guaranteed and CLHU aligns with expected outcomes. However, in the other dataset this interpretation is a relaxation of the real conditions. In this way, CLHU might be found more interactions and not only the spectral variability produced by external factors such as, for instance, illumination. Table 6.3 describes that endmembers have a linear interaction among them, fitting with the first-order multilayered mixture models, the bilinear model, and for that reason, the reconstruction align with the endmembers provided by the datasets. However, it cannot be guaranteed until more controlled tests are carried out.

6.4.4 Multi-layered Scene: Biological Tissue Decomposition

HyperSpectral Imaging has been applied to numerous areas, including medical applications such as noninvasive disease diagnosis and surgical guidance [135]. It offers great potential for these applications because the absorption, fluorescence, and scattering characteristics of tissue change during the progression of disease. Therefore, the reflected, fluorescent and transmitted light from tissue captured by a hyperspectral camera carries quantitative diagnostic information about tissue pathology.

In this sense, the decomposition obtained by HSU provides a valuable additional information about the analyzed tissue. For instance, this can be useful for identifying and quantifying different substances within a tissue sample, obtaining a more in-depth understanding of its composition, which can contribute to various medical diagnostic and research purposes. Given the interest in this thesis, the CLHU framework has been specifically developed with a focus on the application of HSI in the medical field. As it has been indicated in the introduction of this chapter, the importance of HSU techniques in generating significant insights from medical hyperspectral data cannot be overstated. These techniques allow for the exploration of biochemical properties in tissues, offering a new dimension to medical analysis.

This experiment described in this section serves as an exemplification of a multilayered scene where non-linear mixture models, such as the bilinear model outlined in Eq. (6.3), are applicable. In addition, this experimental test has been designed to focusing on the differentiation of two biological tissue types: ex-vivo nerve and fat tissue. This particular test scenario was inspired based on previous research, which highlighted the challenge of distinguishing between these two tissue types using laser-induced breakdown spectroscopy and their high similarity in terms of qualitative elemental composition [331].

In this experiment, an obturator nerve and mesentery of approximately 1 mm in thickness were used. Both tissues were obtained from a pig after euthanization. The mesentery, representing fatty tissue, intentionally covers the nerve tissue, simulating challenging clinical observation conditions. For the data acquisition, an imaging system with a high-speed NIR hyperspectral camera (Compo-

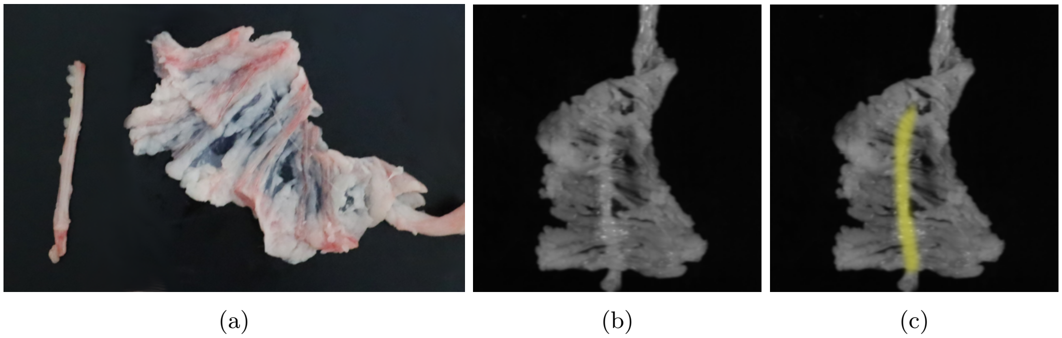


Figure 6.12: Test scenario for HSU to identify ex-vivo nerve and fat tissue. In (a), the image shows both types of tissue (nerve and fat) separated. (b) Image used for the test scenario, with fat tissue covering the nerve tissue. (c) Highlighted nerve tissue under the fat tissue (yellow).

vision, CV-N800HS; Sumitomo Electric Industries, Ltd., Osaka, Japan) was used to obtain the NIR image with a wavelength range of 1000-2350 nm and a wavelength resolution of 6.3 nm [145].

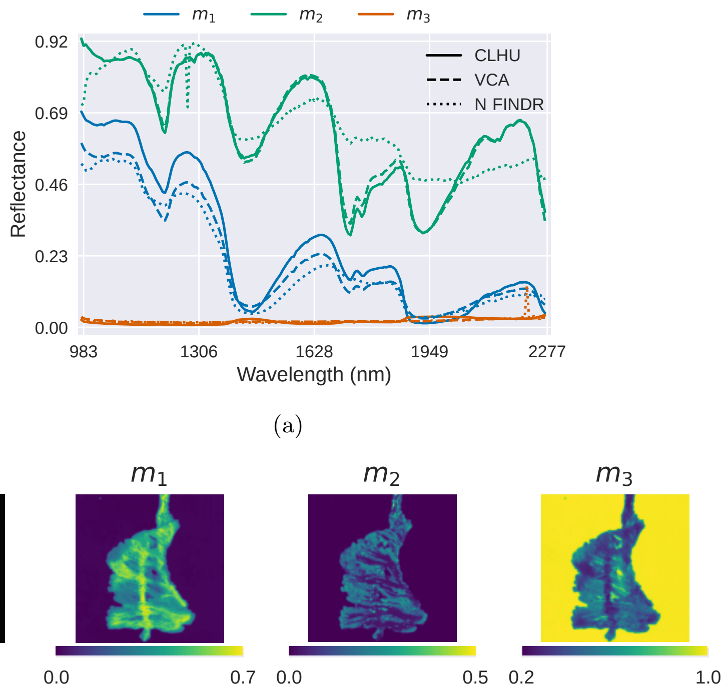
Considering the specific focus of this experiment, a qualitative evaluation approach is employed. The outcomes achieved by the CLHU model, which utilizes VCA for the initial estimation of endmembers (M_0), are showcased in Fig. 6.13a.

Figure 6.13b shows the estimated abundance matrix A. It is observed that m_2 is exclusively related to the fat tissue, achieving a successful unmixing for this component. However, for m_1, it can be observed that the nerve tissue appears among the fat tissue, suggesting a potential misclassification or overlap between these two components. On the other hand, m_3 represents the background, being clearly distinguishable.

The endmember estimation results, Fig. 6.13a, demonstrate that there are similarities in the spectral signatures of m_1 and m_2, which aligns with the expectation that distinguishing between nerve and fat tissues would be a challenging task. As mentioned previously, the two tissues exhibit similarity, and the presence of the fat tissue covering the nerve further complicates the task. It is anticipated that the spectral signature of the fat tissue dominates in the mixture, with subtle variations introduced by the nerve tissue. Therefore, the expected outcome is that the estimated endmembers would primarily reflect the spectral signature of the fat tissue, with specific variations indicative of the presence of the nerve tissue.

Given the context of the experiment, the background material's role within the entire scenario is noteworthy. As depicted in Fig. 6.13, the background's contribution is minimal, marked on the colorbar with a value of 0.2, in those pixels corresponding to nerve tissue locations, which is covered by the fat tissue. This finding supports the insights discussed in Section 6.4.3, suggesting that the CLHU framework is detecting interactions between various endmembers. This capability aligns with expectations for a multilayered scenario, illustrating the model's potential in accurately interpreting complex spectral data.

In the absence of additional experiments using different datasets, it can be concluded that the CLHU framework exhibits promising performance for medical applications. However, further evaluation is necessary to assess its effectiveness more comprehensively. Future studies should consider testing the framework on diverse medical hyperspectral datasets to validate its performance and assess its generalizability across different scenarios and tissue types. By conducting such evaluations, a



(a)

(b)

Figure 6.13: CLHU results of the test scenario using ex-vivo nerve and fat tissue. The endmember estimation (a) and the abundance estimation (b) using CLHU initialized with VCA.

more robust understanding of the capabilities of the proposed CLHU framework and its limitations in medical applications can be obtained.

6.5 HyperSpectral Unmixing by CLHU: Information-Theoretical Validation

In Section 6.3, the CLHU was described from an IT perspective to enhance its interpretability by considering how information is propagated. As in previous chapters, the IT perspective will be validated based on a visualization tool named IP [191], see Section 4.1.2.

As in the previous chapters, the IP estimation has been carried out using the kernel-based Renyi's α-order entropy estimation proposed by Giraldo *et al.* [171], which is mathematically well-defined and computationally efficient, and in Chapter 4 was validated for a scarce dataset. As kernel function, RBF was applied and the kernel-width estimation was carried out by the normalized Silverman's rule proposed in [175], Eq. (3.31). The γ in Eq. (3.31) has been set to 10^{-2}.

In contrast to the previous tests that adhered to the standard configuration outlined in Section 6.4, this section employed a specific setup to enhance result interpretation. Specifically, the encoder architecture, $f(;\theta_e)$, was simplified to three fully connected layers comprising 128, 64, and 32 neurons, respectively. As detailed in Section 6.4, each layer incorporates batch normalization, a dropout, and a ReLU activation function. The loss function, as articulated in Eq. 6.5, had λ set to 0, thereby omitting the minimum volume constraint that is commonly linked with geometric HSU methods. As noted in Section 6.2.2, within CLHU, this constraint is considered optional. For

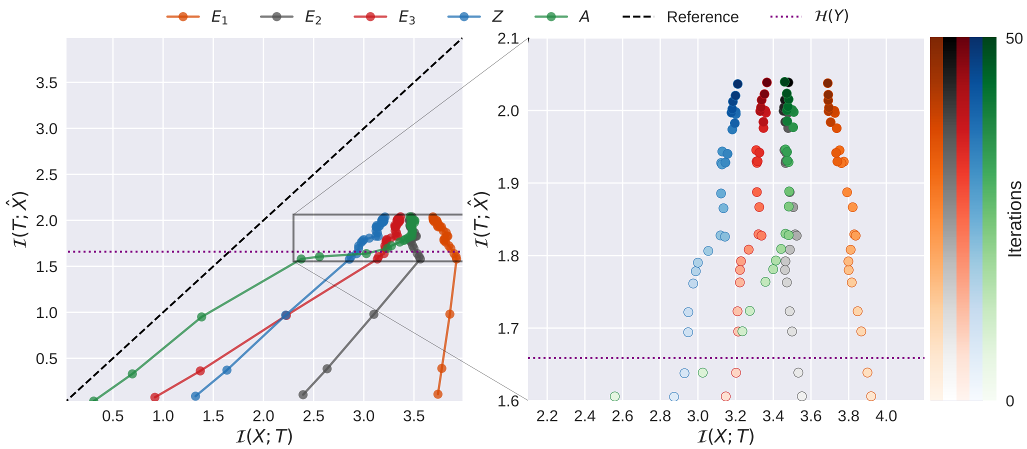


Figure 6.14: IP estimation using 3 endmembers in the synthetic dataset. Based on that it is a reconstruction task, the black dashed line represents the maximum possible value in the different range. The $\mathcal{H}(Y)$ represents the entropy value of the original ground truth.

the experiments, it has been used the synthetic dataset with an estimated SNR of 20 dB. The IP estimation was computed by the code implemented in [172].

To corroborate the IT perspective in Section 6.3, two experiments were conducted to estimate the IP using three and five endmembers. Given that the synthetic dataset was generated using three endmembers, the IP estimation must adhere to all the prerequisites outlined in Section 6.3. Key aspects include ensuring compliance with the DPI as defined by Eq. (4.3), following the typical DPI for cascaded channels (see Section 3.2.1), and Eq. (6.11), as well as a reduction in MI among the endmembers, M, as indicated by (6.12). The latter reduction is facilitated by the contrastive mechanism within the loss function, $\mathcal{L}_{Con}(M)$. In contrast, when employing five endmembers for IP estimation, potential violations of these properties are expected.

In both cases, VCA was utilized as the endmember initialization method. To improve the clarity of the visualizations, a moving average filter with a window size of 5 was applied to mitigate high-frequency noise present in the plots.

Figure 6.14 shows the outcomes of the experiment conducted with three endmembers, aligning with the scenario's requisite count. The black dashed line on the left represents the optimal solution estimated as a reconstruction problem, as the described in Section 4.3, while the purple dotted line represents $\mathcal{H}(Y)$, where Y is the original synthetic data without noise. The encoder's three outputs are labeled E_1, E_2 and E_3. The estimation of the IP reveals that the CLHU model's layers possess adequate information for X reconstruction. Nonetheless, the LMM employed for reconstruction does not sufficiently approximate the noisy input.

In the specific examination of IP estimation with three endmembers (Fig. 6.14), no violations of the DPI are detected, adhering to Eq. (4.3) and Eq. (6.11). The $\mathcal{I}(X;A)$ demonstrates a progression from minimal initial information to achieving parity with E_2 in terms of information content about X. This progression underscores that M is reducing the uncertainty for reconstructing the input X, i.e., it is injecting information as it was described in Eq. (6.8).

As it was described in Section 4.1.2, the training process consists of two distinct phases [191]:: the *fitting phase* and the *compression phase*. The compression phase should occur in the final

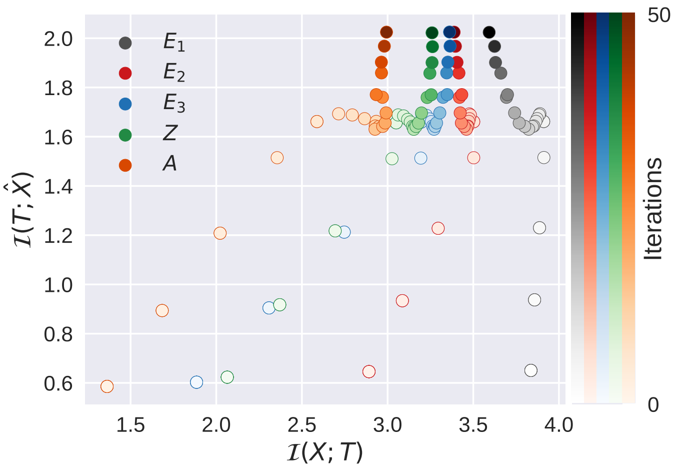


Figure 6.15: IP estimation using 5 endmembers in the synthetic dataset.

stages of training, where irrelevant information from the input X is discarded to prevent overfitting, obtaining a higher compression. In Fig. 6.14, a compression phase is noted in E_1 and E_2, suggesting the model is nearing the optimal solution, converging towards $\mathcal{I}(X;A)$. The compression in E_1 and E_2 becomes apparent as the MI in these layers approaches the entropy of the optimal linear solution, $\mathcal{H}(Y) \simeq 1.67$ bits.

During the compression phase, it's noted that the various layers tend to converge towards the quantity of information present in A. This phase is primarily evident in those layers that possess more information about X than A, while layers with lesser information, such as E_3 and Z, remain in the adaptation or fitting phase. In the context of $\mathcal{I}(X;T)$, the model appears to be converging towards a value of 3.5 bits. Nonetheless, the LMM constraints of this informational content to 2 bits. This observation suggests that A contains all the relevant function for obtaining an optimal solution, but it still contains uncertainty produced by the noise.

In contrast, Fig. 6.15 illustrates the IP estimation when the number of components is not correctly estimated, using 5 endmembers. In this case, the DPI defined by Eq. (6.11) is not satisfied, indicating that the injection of M does not adequately reintroduce the lost information from the input X. Consequently, it is inferred that M introduces uncertainty in the reconstruction rather than enhancing informative content.

However, as observed in Fig. 6.14, a compression phase is apparent in the initial layers, E_1 and E_2. Additionally, it is observed that the different layers are converging, as it was described in the three-endmembers scenario. From this observation, it can be deduced that while the reconstructions in both cases exhibit similarity, the endmember estimation is notably inferior in the five-endmember scenario. As a result, it is concluded that the problem in this scenario is the bad estimation of endmembers.

In summary, in this section it has been validated the theoretical view defined in Section 6.3. Specifically, the model explanation defined in Eq. (6.9) and Eq. (6.10) has been validated, and the problem related to the number of independent components is well-defined in Fig. 6.15. Through the estimates seen in this section, it can be concluded that the model is well explained following the IT view.

6.6 Discussion

The proposed CLHU has demonstrated in the experimental results to be competitive compared with the state-of-the-art methods. Considering the controlled experiments, specially the Section 6.4.1, it is observed that CLHU can obtain an optimal solution for HSU, which is considered an ill-posed inverse problem. Table 6.1 shows that CLHU works perfectly in conditions where the input is noisy in both task, abundance and endmember estimation.

Despite the promising results using the synthetic dataset, Section 6.4.2 can generate doubt about the use of CLHU, in specific, considering the quantitative results in Table 6.2. In this section has been different remote-sensing HSI datasets commonly used in the literature to validate HSU algorithms. In this case, CLHU shows discrepancy with the ground-truth of these datasets. However, as it was indicated in Section 6.4.2, the ground-truth provided by these datasets is not well-defined, and it has been estimated by different HSU. For this reason, the results obtained in this Section 6.4.2 should be interpreted as indicative rather than definitive. In other words, the ground-truth is not a real reference and the results are just demonstrated that the proposed CLHU is obtained a different 'interpretation' of the dataset that cannot be considered wrong given the ill-posed problem.

In general, in the datasets used in Section 6.4.2, the qualitative results about the abundance estimation are considered correct, and it does not show special discrepancy with the different methods. In the case of the Apex dataset, it is observed that the abundance of 'water' shows that this element is observed in the whole image. However, as it was indicated in Section 6.4.2, Apex dataset has not a preprocessing step in which hyperspectral bands contaminated with water vapor density and atmosphere have not been removed. For this reason, the result obtained in Apex is considered a problem because of the lack of preprocessing step. The use of methods in which the solution is regularized by a sparse factor helps to obtain a better result, compared to the proposed ground-truth, as it happens in FSSU or in the EGU-Net which was guided by SUnSAL, a method in which the sparsity is considered.

Regarding the endmember estimation, Section 6.4.2 shows results that differ much from the proposed ground-truth. In specific, CLHU obtains a solution that differs significantly from the other methods in Samson dataset. It was a consequence of the m_1 estimation, as it is observed in Fig. 6.10b. In Section 6.4.3, it was compared the 'soil' signature obtained in Samson with the obtained in another dataset, Jasper Ridge. In both case, it was observed a specific pattern that it is not observed in the different ground-truth, a peak of absorbance around 700 nm.

Given the aforementioned observation, in Section 6.4.3 was proposed to analyze what happens when CLHU has to reconstruct the estimated endmembers. Considering the *Similarity* module, Eq. (6.4) shows that the CLHU model force to find a compressed representation Z in which all endmembers contribute to the A estimation. This observation is well-defined in the IT perspective given the Eq. 6.9, where it is necessary to maximize $\mathcal{H}(M, Z)$. As a result, the reconstruction of M should be a worse estimation, but Table 6.3 shows that, in terms of SAD_M, the reconstruction fits better with the proposed ground-truth in the state-of-the-art dataset used in Section 6.4.2. Additionally, it is observed that for Jasper Ridge and Apex dataset the results obtained outperforms the estimation to the other methods. As a conclusion, it was suggested that CLHU consider the interaction among endmembers, as different NMM such as the bilinear model, Eq. 6.3.

In general, in remote sensing, the LMM is considered a good approximation of the mixture model since the mixing scale is macroscopic, see Section 6.1.1. However, even the mixture model in

CLHU is the LMM, the approximation that it has been obtained corresponds to a NMM and it is a consequence of the discrepancy of CLHU with the different methods.

Finally, for supporting the conclusions obtained from Section 6.4.3, it was considered a multilayered scene in which the endmember interaction has to be considered. The experiment was described in Section 6.4.4, and it was demonstrated that CLHU works in this scenario. In addition, this experiment allows illustrating how the HSU technique would be relevant in the medical imaging domain for a better interpretability of biological tissue data.

6.7 Conclusion

The interpretability of data is crucial for gaining more in-depth insights into its informational content. The data interpretation is essential in critical applications fields, such as autonomous driving or medical screening, since bad interpretation of the data can result in inconsistencies and potentially adverse outcomes. However, this interpretability is demanding in high-dimensional data, such as HSI, and it is necessary the use of dimensional reduction technique to address the ‘curse of dimensionality’, the challenges encountered when managing data in high-dimensional spaces challenges that are not present in lower-dimensional settings.

While common dimensionality reduction techniques, like PCA, are widely used, they may not always aid in enhancing data interpretability. This is because the new representations created by such methods are typically linear combinations of the original features, which can obscure the physical significance of the data. In contrast, HSU plays a pivotal role in improving data interpretability. Unlike PCA or ICA, HSU decomposes signals using the original feature space. This approach ensures that the outcomes of HSU methods not only enhance the physical understanding of the data but also constrict the solutions to physically meaningful ranges. This makes HSU a valuable tool in translating complex high-dimensional data into interpretable and physically relevant information.

It is worth noting that although these methods are used for HSI, it might be extended to different high-dimensional signal decomposition since, as it was described at the beginning of this chapter. The result of these methods obtained a decomposition in endmembers and the abundance map but, as it has been mentioned in this chapter, the endmember definition is problem-dependent as the components in PCA or ICA.

This chapter presents a novel approach for HSU using a DL model, the CLHU framework. Unlike existing methods that primarily rely on the DL model’s reconstruction capabilities for abundance estimation, the proposed CLHU approach leverages the relationship between the input signal and the estimated endmembers to approximate the abundance map. By incorporating the latent space and exploiting its connection to the estimated endmembers, the proposed method offers a new perspective and potential improvements in HSU performance. This method might be extended to different high-dimensional data, but it is necessary to consider that the endmember initialization is a fundamental part for obtaining an optimal solution in CLHU.

The proposed method was tested across various hyperspectral datasets, proving that the proposed method exhibits promising performance and is competitive with existing approaches. Although the LMM is utilized as a foundational mixture model within the CLHU framework, it has been noted that CLHU provides a nonlinear perspective without the necessity of adopting a NMM. This is achieved by incorporating endmember interactions, an aspect often linked with non-linearity in scenarios involving multiple layers. Furthermore, CLHU demonstrates its capability to identify patterns in the estimation of specific material endmembers across diverse datasets, as exemplified by the 'soil' endmembers from the Jasper Ridge and Samson datasets.

This thesis aims to improve the interpretability of the algorithms used in DNN models, seeking a deeper comprehension of these sophisticated mechanisms. An IT perspective has been employed to enhance the interpretability of the model, extending the discussed in previous chapters, examining the inner workings of DNNs in problem-solving. This analytical approach offers valuable insights into adapting the proposed CLHU model for various applications. For example, the IT perspective elucidates the role of the *Similarity* module, as detailed in Section 6.2.1, which ensures the interaction among different endmembers during the hyperspectral signature reconstruction process, following InfoMax principle depicted in Eq. (6.9). A validation of the IT perspective was carried out in Section 6.5. Given this IT perspective, it is observed that a bad estimation of the number of endmembers can impact the solving mechanisms employed in CLHU, potentially leading to a situation where the DPI, Eq. (6.11), might not be satisfied. Based on this analysis, it would be possible to define an algorithm to estimate the correct number of endmembers found in HSI data.

Chapter 7

Anomaly Detection by Deep Learning

> The greatest discoveries often lie not in finding new things, but in seeing familiar things in new ways.
>
> *Alexander Fleming*

Beyond the conventional applications of DL models in classification tasks, as illustrated in Chapter 4, the focus has shifted towards enhancing data interpretability in recent chapters. Chapter 5 concentrated on developing a DNN algorithm for FS, while Chapter 6 introduced a DL framework for interpretable high-dimensional signal decomposition, specifically targeting HSU. This chapter shifts attention to a crucial task for performing a coherent dataset analysis: Anomaly Detection (AD). The method based on DL will be evaluated through a toy experiment using the MNIST dataset, along with two medical datasets — pneumonia detection and brain cancer — both of which will be tested under an extremely imbalanced configuration in the training set.

The goal of AD is to identify unexpected or outlying observations within the data that deviate significantly from the norm, potentially signifying errors or noise. This task is critical for ensuring the integrity and reliability of data analysis, highlighting outliers that may influence the overall outcomes or indicate novel or rare events worthy of further investigation. However, these outliers may contain important information, so they cannot be always discarded depending on the hidden assumptions regarding the data structure.

The concept of an anomaly, or an outlying observation, is crucial for deepening our understanding of AD. Hawkins provides a foundational definition of an outlier as [332]:

> *an observation that deviates so much from other observations as to arouse suspicion that it was generated by a different mechanism.*

This definition underscores that the identification of outliers is not solely based on the data structure, but also heavily influenced by the detection methodology. In other words, the AD algorithm might be designed for a specific mechanism for generating data and, for this reason, it is worth emphasizing the need to select or design algorithms capable of identifying various types of anomalies effectively.

In the healthcare sector, the application of AD has received significant research interest. Within this context, the pathologies are typically interpreted as unusual patient health conditions [333, 30], a definition that fits with anomalies or outliers in the data. For instance, an 'anomalous' MRI image may indicate the presence of malignant tumors. This perspective underscores the potential of AD in identifying rare diseases, atypical symptoms, or unexpected responses to treatment, highlighting its importance in enhancing patient care and facilitating early intervention strategies.

Chapter 4 discussed how AI could alleviate the radiologists' workload in analyzing radiological imaging data [5], such as CT scans or MRIs. However, a significant obstacle in this endeavor is the *data challenge*, primarily the scarcity of sufficiently labeled data [180]. This scarcity arises from the difficulties in systematically collecting data and the labor-intensive nature of labeling datasets, which paradoxically could increase a radiologist's workload. For instance, in pathology, annotating diverse microscopy datasets is both time-consuming and expensive due to that a tissue slide may potentially contain millions of cells that necessitate specialized expertise from the annotator [334].

In the same chapter, it was evaluated the analysis of DNN models based on IT perspective using a limited dataset. This approach aimed to diminish the uncertainty inherent in DNNs without the need for extensive testing datasets, offering an alternative testing methodology for such scenarios. However, in these cases, a crucial question arises: could the task of labeling datasets be optimized? AD emerges as a promising solution in this context. By identifying outliers or unusual patterns automatically, AD has the potential to streamline the labeling process, highlighting data points that require human expert review. This approach not only aids in efficiently generating labeled datasets but also ensures that the focus is maintained on the most clinically significant or anomalous cases, optimizing resource allocation and potentially enhancing diagnostic processes.

Furthermore, in supervised contexts, AD methods are often intended to manage unbalanced datasets. Semi-supervised AD specifically seeks to train models using solely the normal class [29], often represented by healthy data, and subsequently applies the model to both healthy and pathological datasets to generate 'scores' [30]. It is worth noting that, in general, the samples referring to a pathology are a minority. Consequently, it is important to acknowledge that these methods can also serve classification purposes, particularly in medical screening scenarios where there is not enough labeled data.

Throughout this thesis, the discussion has centered on the ability of DL models to extract features and optimally compress data. In addition, these models have the benefits that they are easily-adaptable to large and different complex datasets. This makes them suitable for AD, aiming to learn feature representations or anomaly scores via DNN for the sake of AD. Recently, numerous DL-approaches for AD have been introduced, demonstrating significantly better performance than traditional and ML-based AD methods on addressing challenging detection problems in various real-world applications [335, 336], including the medical context [192].

The effectiveness of DL in AD can be attributed to its capability for data compression, which directly addresses the 'curse of dimensionality' [228], a challenge thoroughly examined in earlier chapters. This curse represents a significant barrier for numerous AD techniques when dealing with high-dimensional data, undermining the efficiency of traditional methods like distance-based, density-based, and clustering-based AD approaches [337]. In the healthcare sector, as highlighted in Chapter 5, high-dimensional data is increasingly prevalent [231], exemplified by high-resolution and multimodal medical imaging. Multimodal imaging, which combines various medical imaging

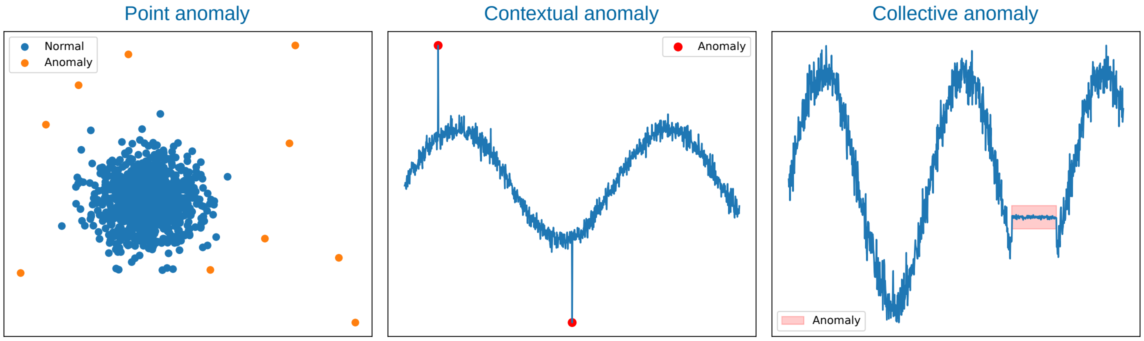


Figure 7.1: Illustration of the different types of anomalies in the state-of-the-art.

techniques [338], epitomizes the complex data types in healthcare that benefit from DL's advanced processing and analysis capabilities.

This chapter details the adaptation of DL models for AD, with a specific focus on medical imaging. This context is particularly relevant, as the definition of an anomaly often depends on the AD method employed, as discussed in previous chapters. The chapter is structured as follows: first, it introduces the different types of anomalies and the various detection methods. The next section presents the proposed DL-based method, Anomaly Detection in Latent Spaces using Entropy-based score (ADeLEn), which is built on semi-supervised learning for AD. Following this, the IT perspective is discussed, describing the mechanism based on the InfoMax principle. Finally, experimental results are presented, followed by a discussion and conclusions.

7.1 Type of Anomaly and Detection Methods

As previously mentioned, the design of an AD method is fundamentally designed by the characteristics of the anomalies it aims to detect. The current literature recognizes various ways to classify how anomalies manifest. The main distinctions are made between anomalies that are extreme yet genuine parts of the main population, such as random fluctuations at the tails of a distribution, and contaminants which come from different distributions [339]. Further distinctions are made between obvious errors and outliers that are unusual but not extremely rare, as well as between weak outliers (noise) and strong outliers (significant deviations) [339]. However, broadly, anomalies are commonly classified into three types: point anomalies, contextual anomalies, and collective anomalies [340, 339].

Point anomalies pertain to cases that deviate significantly from the rest of the data. They often represent irregularities or deviations that occur randomly and may not have a specific interpretation [340, 339]. Contextual anomalies, on the other hand, are anomalies that, while appearing normal in a general sense, become deviant when considered within a specifically selected context [340, 339]. Finally, collective anomalies involve collections of individual data points that, while not necessarily anomalous on their own, appear anomalous as a group compared to the rest of the data [340, 339]. Figure 7.1 illustrates these types of anomalies.

The kind of anomaly is closely linked to the specific task at hand. For example, in NLP contextual anomalies are prevalent. These anomalies occur when the usage of words deviates from the expected context, such as the use of technical jargon in a casual conversation. An illustrative example of collective anomalies can be seen in the case of fraudulent credit card transactions, where a single transaction may not appear anomalous, but a group of unusual transactions together can be indicative of fraud [335]. However, point anomalies are the simplest case and the majority of research are focuses on this [340].

7.1.1 Anomaly Detection Methods

Anomaly Detection has been a subject of study since at least the 19th century [341], resulting in a wide variety of AD methods. The most common techniques for AD are the statistical techniques, in which we can find the parametric and non-parametric statistical techniques. The parametric techniques assume that the normal data is generated by a parametric distribution with parameters ω and a probability density function $f(x;\omega)$, where x is an observation [340]. The non-parametric are those methods in which the model structure is not defined a priori, and it is determinate by the data.

The simplest case of parametric technique is the box plot rule using the IQR, which was applied in Section 5.4.2, where the data is assumed to follow a Gaussian distribution. Following this Gaussian assumption, there are different tests that can be used for AD, such as the Grubb's test [342], using a score based on the mean and standard deviation of the data sample, or the student t-test [343], where a normal sample is compared with a test sample using t-test and, if the test shows significant difference between them, it signifies the presence of an anomaly. The use of a linear regression is another practice of parametric statistical technique, in which is assumed a linear relationship between independent and dependent variables. The idea of this approach is to fit a linear model on the data and the residual, the part that it is not explained by the model, corresponds to the anomaly score [340, 332]. Finally, there are techniques based on a mixture of parametric statistical distributions to model the data. As an example of these methods, one approach consists on the assumption that the data and the anomaly follow the same distribution but with different parameters, for instance a Gaussian data with the same mean but with a larger variance [344]. Another approach can be considered that normal instances follow a mixture of parametric distributions, as a GMM [345], and if the test does not fit with that mixture distribution, then it is considered an anomaly.

The non-parametric statistical techniques are based on the estimation of the PDF of the normal data. Here, it can find two approaches based on histograms, frequency-based, or the use of Parzen windows [170]. For instance, in the use of histograms, requires constructing the histogram of the normal data, although it can be considered to obtain the histogram of the anomaly distribution too [346], and the size of the bin used is key for the AD [340].

The use of ML has also considered for AD, including the DL. The taxonomy of these methods is complicated, and it could be considered based on the learning process using, such as the observed in Section 2.1.1. However, in the literature it is commonly to categorize the AD method based on ML considering how the problem is carried out [347]: classification, regression, clustering, etc. For instance, the clustering-based AD technique assumes that normal data belong to a cluster while anomalies do not belong to any cluster, similar to the observed in the point anomaly in Fig. 7.1. This method is mainly an unsupervised technique, although it has been explored in the semi-supervised approach in clustering [29]. In those techniques, the clusters are generated by using ML algorithms such as K-Means [27] or DBSCAN [348].

In classification-based AD techniques using ML, it is associated with a supervised or semi-supervised technique. Supervised AD involves training a supervised binary or multi-class classifier, using labels of both normal and anomalous data instances [335]. However, there are alternatives that are considered unsupervised as the One-Class SVM (OC-SVM) [349]. This method is a kernel-method where the SVM is trained for finding a maximum margin hyperplane in feature space that best separates the mapped data from the origin. In this case, the data corresponds to the 'normal' data and the anomaly score is based on this distance to the origin, as a result, the anomaly detection is carried out by using an estimated threshold in this distance. The Support Vector Data Description (SVDD) [350] is another technique related to the OC-SVM but, instead of using a hyperplane, this method separates the data using a hypersphere. The use of SVDD then consists in detecting as a normal every point contained in the hypersphere. Another alternative but using ensembles model is the Isolation Forest [351], which employs binary Decision Trees for identifying anomalies, assuming that anomalous data points are easier to separate from the rest of the sample.

7.1.2 Deep Learning Approaches for Anomaly Detection

The use of DL in AD has become increasingly prevalent recently. Deep learning approaches have been shown to outperform traditional AD methods, including ML, especially as the scale of data grows [335]. The primary contribution of these DL approaches is related to data compression, a task at which DL is particularly effective, as has been detailed in previous chapters. For example, Hawkins *et al.* [352] proposed the use of a Replicator Neural Network for AD. This architecture, originally designed for data compression [353], utilizes a staircase activation function that enables the network to compress data by assigning it to a defined number of clusters. This concept, initially implemented with replicator neural networks, has evolved into the use of AEs, which serve a similar purpose, providing a non-linear compressed representation of data through a self-supervised learning approach [67], see Section 2.2.3.

The primary approach to using AEs for AD rests on the hypothesis that the training dataset consists exclusively of observations confirmed to be normal [354]. The underlying idea is to learn a model of what is defined as the normal class; observations that deviate from this learned model are then classified as anomalies [354, 340, 336]. Consequently, training involves compressing the data into a latent representation Z and then decompressing it to reconstruct the input x, denoted as $\hat{x}$. The reconstruction error, which can be measured as the MSE, is directly utilized as an anomaly detection score for evaluating test instances [340]. In other words, the score is based on the likelihood $p(x|z)$, see Section 2.1.3. As a result, the data instances such as anomalies that deviate from the majority of the data are poorly reconstructed. For contextual anomalies, this method can be adapted using an LSTM-based AE, which is particularly effective in handling sequence data [355, 356, 357].

Nonetheless, the assumption that the training data solely comprises normal observations cannot be guaranteed due to the possibility of annotation errors or the impracticality of thoroughly screening large datasets. While strategies like data standardization can mitigate the impact of outliers during training, they cannot completely rectify the distortion of the initial hypothesis. Consequently, if anomalies contaminate the training dataset, they might lead to a misrepresented model of normalcy, inadvertently causing these outliers to be misclassified as normal. This highlights a critical vulnerability in the reliance on AEs for anomaly detection, especially in settings where pristine training data cannot be reliably guaranteed.

Recently, several AE-based approaches have been proposed to address the issue of training data contamination in AD. Zhou and Paffenroth [358] introduced the Robust Deep AE, inspired by Robust PCA [359], which enhances the robustness of AEs for AD by iteratively separating the original data into two subsets: a normal instance set and an anomaly set. Chen *et al.* [360] aimed to further improve the AE-based approach by learning an ensemble of AEs instead of a single AE. This method was specifically designed for tabular data [336] but is adaptable to other data types as well.

Additionally, there are robust alternatives to AEs that can handle noise in the normal data. One such method is the Deep SVDD proposed by Ruff *et al.* [361]. This method integrates concepts from one-class classification approaches like the OC-SVM and SVDD. Later, Ruff *et al.* further developed this approach into a semi-supervised method known as Deep Semi-supervised Anomaly Detection (Deep SAD) [362], which utilizes a subset of labeled anomaly samples to ensure that these samples are distanced from the centroid of the normal samples, thereby enhancing model training and anomaly detection performance.

Although the use of AEs for AD is commonly associated with describing an anomaly score based on likelihood, Sakurada *et al.* [363] recognized that the latent representation in the bottleneck of the AE can distinguish between normal and anomalous data. In other words, AD can be conducted by using an anomaly score derived from the latent space Z. Probabilistic interpretations of AEs, such as those discussed by Vincent *et al.* [364], can model aspects of the data-generating process more directly. This form of regularization helps the latent representation to focus on a data manifold, encoding the most relevant features of the data [354], which enhances the effectiveness of AEs in AD by emphasizing crucial data characteristics while filtering out noise and less relevant information. Alain and Bengio [365] indicate that regularized AEs implicitly estimate the data generation process, and there is a established link between reconstruction error and data generation density.

Based on the aforementioned relationship, the generative models are a suitable option for the use of AEs for AD. The use of generative models, such as VAE has been well-established in the literature for AD. An and Cho [366] proposed to use a VAE for reconstruction and the anomaly score is based on this reconstruction. Additionally, the use of Adversarial AutoEncoder (AAE) [367], which combines the principles of generative adversarial networks introduced by Goodfellow *et al.* [56], has also been proposed for AD [368] using reconstruction error as anomaly criterion. Zhai *et al.* introduced an energy-based model for estimating the PDF and applied a criterion based on an energy score, which has demonstrated superior results compared to traditional reconstruction error criteria [369].

The application of generative models in AD has been extended to include various types of anomalies. For instance, Longyuan Li *et al.* [370] explored the use of VAEs for contextual anomalies, particularly in time-series data. They developed a sequential VAE and proposed two decision criteria for AD: reconstruction probability and reconstruction error. This illustrates the flexibility and effectiveness of generative models in adapting to different anomaly detection scenarios, reinforcing their suitability for a wide range of applications. Recently, DL-based DPM have also been considered for AD using a semi-supervised approach [371].

To enhance the robustness of generative models for AD, various conditions can be applied within the latent space Z where the PDF is estimated. For example, Beggel *et al.* [354] utilized an AAE and implemented an iterative refinement process during training. This process involves removing samples that are deemed anomalies by an OC-SVM within the latent representation Z after several training epochs. This approach ensures that by the end of the training period, the estimated PDF of

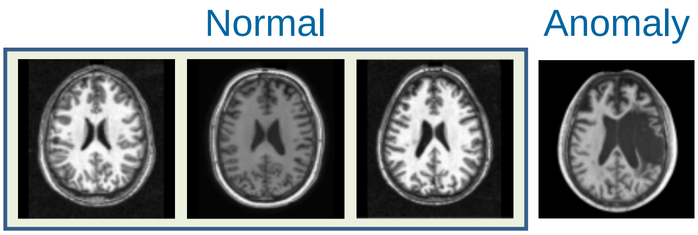


Figure 7.2: Example of anomalies in medical imaging: Brain case study.

Z more accurately models the normal data samples, significantly enhancing the model's effectiveness in distinguishing between normal and anomalous data.

For multi-class AD scenarios, Norlander and Sopasakis [372] introduced a method that incorporates a conditional aspect in the variational inference, based on the labels of different classes. This strategy allows for distinct representations of 'normal' data to be derived based on the specific classes involved. By doing so, the model can more precisely capture and model the normal behavior within each class, thereby improving its ability to detect anomalies specific to each class context.

7.2 Proposed algorithm

In this section, it is described the proposed method designed specifically for the healthcare scenario. As discussed in the introduction of this chapter, AD methods are tailored based on the data acquisition mechanism. In the healthcare sector, anomalies are typically interpreted as pathologies; for instance, an 'anomalous' MRI or CT image can indicate unusual patient health conditions [333, 30]. An illustrative example is depicted in Fig. 7.2.

As observed in Section 7.1.2, most DL-based AD methods generally use reconstruction error as the anomaly score, operating under the assumption that anomalies will be poorly reconstructed by the AE. Although there are alternatives, as mentioned in Section 7.1.2, this reconstruction-based approach is predominant. Generative models such as VAEs, AAEs, and DPMs are well-suited for this task. However, this approach requires the establishment of a threshold that is heavily dependent on the data. To sum up, an anomaly is determined by:

$$\|f(x, \theta) - y\|_F^2 > \xi, \tag{7.1}$$

where θ are the parameters of the model and ξ is the manually assigned threshold.

In this section, a method is proposed that does not rely on the reconstruction error for AD. Instead, it leverages the latent representation, Z, from the bottleneck of the AE architecture to compute a score used for AD. The method introduced in this section focuses on medical imaging, where anomalies can be categorized as point anomalies (see Section 7.1). Consequently, we propose the ADeLEn semi-supervised DL method.

7.2.1 Parametric Statistical approach for Medical Imaging

Parametric statistical AD methods are based on detecting anomalies by analyzing the probabilities of the data, $p(X)$. In the context of medical imaging, it is intuitive to consider that normal data, denoted as X^+, and anomalous data, denoted as X^-, may follow the same distribution but with different parameters. For instance, if $p(X)$ follows a Gaussian distribution, anomalies could be represented by a larger variance:

$$\begin{aligned} &X^{+} \sim \mathcal{N}(0; \sigma^{+}), \\ &X^{-} \sim \mathcal{N}(0; \sigma^{-}), \\ &s.t. \quad \sigma^{-} > \sigma^{+}. \end{aligned} \tag{7.2}$$

This idea is grounded in the fact that, in medical imaging, the images are generally similar among subjects. For instance, as observable in Fig. 7.2, brains are typically similar across subjects, but a brain with a tumor, considered an anomaly in this context, exhibits a specific pattern that is uncommon in normal brains. Moreover, apart from the fact that human brains are similar, medical imaging is conducted following standardized acquisition protocols. These protocols reduce variability in the images, which facilitates interpretation by radiologists, as the diagnostic information in these data is heavily dependent on human interpretation [6]. For this reason, the aforementioned approach is suitable under these conditions.

However, there are two key considerations when applying this approach for anomaly detection in medical imaging. First, the images are typically high-dimensional data, and due to the 'curse of dimensionality' [228], estimating the marginal distribution $p(X)$ in high-dimensional spaces is extremely complex. Secondly, the assumption that an image follows a well-known distribution, such as a Gaussian distribution, is overly simplistic. Even if the images are similar, it is intuitively clear that they are too complex to be accurately described by a normal distribution. As a result, the description in Eq. (7.2) is not suitable.

The aforementioned challenges to the use of statistical approach can be carried out using generative models by incorporating a latent variable Z. In this way, the problem framed as:

$$p_\theta(x) = \int_z p_\theta(x|z) p_\theta(z), \tag{7.3}$$

where $p_\theta(.)$ corresponds to the probability function estimated by the model with parameters θ. The form of the prior $p_\theta(z)$ and the likelihood $p_\theta(x|z)$ will depend on the data being modeled. However, as it was indicated in Section 2.1.4, this marginalization is often computationally intractable.

Assuming that the true prior $p(z) \sim \mathcal{N}(0, I)$ and $p_\theta(x|z)$ as a factorized Gaussian, the goal is to fit the model to an empirically observed subset $\hat{p}(X)$. A naive approach involves estimating the model parameters θ by MLE, Eq. (2.7), using Monte Carlo sampling to approximate the integral over z.

$$\theta^* = \operatorname*{argmax}_\theta E_{x \sim \hat{p}(x), z \sim p(z)} [\log p_\theta(x|z; \theta)]. \tag{7.4}$$

However, it is well-known that this approach does not scale well to high-dimensions of Z [373], another manifestation of the 'curse of dimensionality'. Here is where DL-based models play a fundamental role.

DL-based generative models can handle these problems effectively. As discussed throughout this thesis, from the IT perspective, it has been demonstrated that DL-based models provide efficient compression of information. In other words, a DNN is well-suited for obtaining a compressed representation of the input X, denoted by Z, with a lower dimensionality. Using VAE [55] as reference, VAE addresses this problem by sampling z from a new distribution $q_\phi(z|x)$, where ϕ represents the

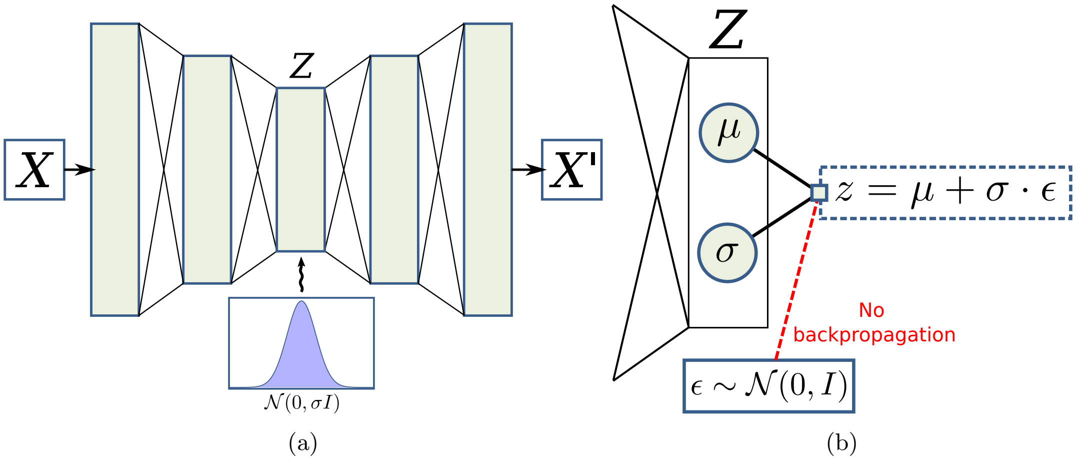

X
Z
X'
$\mathcal{N}(0, \sigma I)$
(a)
Z
μ
σ
$z = \mu + \sigma \cdot \epsilon$
No backpropagation
$\epsilon \sim \mathcal{N}(0, I)$
(b)

Figure 7.3: Illustration of the VAE. In (a), it is depicted how the latent space follows a normal distribution based on the variational distribution. In (b), the Reparameterization Trick for optimizing the model is illustrated.

parameters of the distribution, which is jointly optimized with the generative model.

The idea of VAE is to use the variational inference (see Section 2.1.4) to optimize the variational parameters, which are obtained by a neural network, so that $p_\theta(z|x) \approx q_\phi(z|x)$. This is achieved by optimizing the ELBO, described in Eq. (2.11). As a result of solving the ELBO, the generative model follows:

$$\log p_\theta(x) \geq \log p_\theta(x) - D_{KL}(q_\phi(z|x) \| p_\theta(z|x)). \tag{7.5}$$

A more detailed description of this alternative description of the ELBO can be found in Appendix A.4.1. Observing this equation, the VAE will concurrently maximize the marginal likelihood $p_\theta(x)$ of the generative model while minimizing the KL Divergence between the approximation $q_\phi(z|x)$ and the true posterior $p_\theta(z|x)$ obtained by the encode path in the AE. In other words, the solution is conditioned by the 'distance' between $q_\phi(z|x)$ and $p_\theta(z|x)$. However, VAEs address this problem by setting the prior distribution $p(z)$ (see Section 2.1.4).

As a result, VAE provides a compressed representation that follows a probability distribution established by $q_\phi(z|x)$. Figure 7.3 illustrates a VAE where the latent space $Z \sim \mathcal{N}(0, \sigma I)$, with I is the identity matrix, as shown in Fig. 7.3a. Figure 7.3b depicts the Reparameterization Trick [254], which is used for optimizing the DL model. In this approach, the neurons in the bottleneck estimate the variational parameters that define the distribution, specifically the mean (μ) and the variance (σ).

7.2.2 Semi-supervised approach for Anomaly Detection

In this section, we describe the proposed method, ADeLEn, a semi-supervised approach for AD using the parametric statistical approach outlined in Section 7.2.1.

The proposed ADeLEn method utilizes a generative model to obtain the distribution of the compressed representation Z. The goal is to extend the naive statistical approach observed in Eq. (7.2) to medical imaging using this compressed representation. As previously discussed, even if the images are similar and their variability is limited, they are too complex to be accurately described by a normal distribution. However, the compressed representation can follow a multivariate Gaussian distribution. In that way, the hypothesis follows:

$$p_\theta(z|x^+) \simeq p_\theta(z|x^-), \tag{7.6}$$

where x^+ represents an normal sample and x^- the anomalous sample. It is worth noting that it is assumed that $p_\theta(z|x^-)$ and $p_\theta(z|x^-)$ follows a normal distribution.

As a result, considering the proposed method inspired by Eq. (7.2) and Eq. (7.6), the normal and anomalous samples follow the same normal distribution, but the latter has a higher standard deviation, as shown in Eq. (7.2). Consequently, the hypothesis follows the next assumption:

$$\begin{aligned} &Z^+ \sim \mathcal{N}(0;\sigma^+ I),\\ &Z^- \sim \mathcal{N}(0;\sigma^- I),\\ &s.t. \quad \sigma^- > \sigma^+, \end{aligned} \tag{7.7}$$

where $p(z^+) = \int p(z|x^+)p(x^+)$. The assumption that both distributions follow an isotropic Gaussian distribution, with a diagonal covariance matriz I, is for simplicity, reducing computational complexity and enhancing efficiency since each dimension in the latent space is independent. However, if desired, it is possible to extend this to a Gaussian distribution with full covariance [374].

This distinction between the two distributions for normal and anomalous samples is achieved through semi-supervised learning, see Section 2.1.1. As previously discussed, it is commonly assumed that the training dataset consists exclusively of observations confirmed to be normal [354]. However, it is challenging to ensure that the dataset is free of unidentified anomalous samples in practice. This is particularly relevant in our case, where normal and anomalous samples are quite similar and follow the same distribution. Therefore, it is necessary to include a subset of labeled anomalous samples.

In this approach, the non-labeled dataset, X^+, is assumed to be normal and consists of N samples. The subset of labeled anomalous samples, X^-, is relatively small, with M samples. Consequently, the problem addressed in this chapter follows $N \gg M$. The idea is to use a VAE (although another DL-based model, such as AAE, could also be used) where the variational parameters ϕ are optimized to distinguish between both types of samples.

Loss function

This problem is addressed by the ELBO. Specifically, the loss function in a VAE is defined by:

$$\mathcal{L}_{\theta,\phi}(x) = -E_{q_\phi(z|x)}\left[\log p_\theta(x|z)\right] + D_{KL}(q_\phi(z|x)\|p(z)), \tag{7.8}$$

which corresponds to the observed in Eq. (2.11) where the KL Divergence acts as a regularization term (see Section 2.4.2). However, in this case there are two distributions to consider, as it was described in Eq. (7.7).

Recall that the latent space in a VAE aims to estimate the variational parameters of the distribution of the model, see Fig. 7.3b, and it is assumed that follows a Gaussian distribution. As it was previously mentioned, Eq. (7.7), it is assumed that $p(z)$ follows an isotropic Gaussian distribution to simplify the annotation. Thus, the KL Divergence between two isotropic Gaussian distributions is given by:

$$D_{KL}(Q_\phi \| P) = \frac{1}{2}\left[\ln\left(\frac{\sigma_p^2}{\sigma_q^2}\right) + \frac{\sigma_q^2 + (\mu_q - \mu_p)^2}{\sigma_p^2} - 1\right], \tag{7.9}$$

where $q_\phi(z|x) \sim \mathcal{N}(\mu_q, \sigma_q)$ and $p(z) \sim \mathcal{N}(\mu_q, \sigma_q)$. Observing the hypothesis described by Eq. (7.7), the KL Divergence can be simplified by removing the mean value $\mu_p = 0$.

Additionally, for optimizing the problem, it is necessary to describe the σ for both distributions, σ^+ for normal samples and σ^- for anomalous samples. It consists on hyperparameters of the model, and it is from free election. In this chapter, it will be considered that:

$$\begin{aligned} z^+ &\sim \mathcal{N}(0,1), \\ z^- &\sim \mathcal{N}(0,3). \end{aligned} \tag{7.10}$$

Given the sample distribution depends on a label, $y \in \{0, 1\}$ where 0 represents the normal samples and 1 the anomalous samples, the marginal distribution $p(z)$ can be expressed considering the label. In that way, $p(z^+) = p(z|y = 0)$ and $p(z^-) = p(z|y = 1)$.

Finally, the loss function in ADeLEn is defined considering this label. As a result, Eq. (7.8) is modified as follows:

$$\mathcal{L}_{\theta,\phi}(x) = -E_{q_\phi(z|x)}\left[\log p_\theta(x|z)\right] + D_{KL}(q_\phi(z|x)\|p(z|y)). \tag{7.11}$$

In practice, the first term corresponds to the MSE and the second term is the KL Divergence defined in Eq. (7.9) considering the Gaussian distribution depending on the label y, see Eq. (7.10).

Score

An AD method has to obtain a score which is used by the model for taking a decision. As it has been previously mentioned, the proposed ADeLEn aims to discriminate the samples in the latent space Z which follows a variational approach. In this way, following the Bayesian inference approach (see Section 2.1.4), the score can be expressed based on the confidence of the estimation carried out by the model.

As a result, it is proposed to use as a score function the uncertainty that the model has about the distribution $p_\phi(z|x)$. The entropy measures this uncertainty $\mathcal{H}_\phi(Z|x)$, given the name to the proposed ADeLEn. In this way, the score function using the Gaussian distribution assumption follows:

$$\mathcal{H}_\phi(Z|x) = E_{z\in Z}[-\log p(z|x)] = \frac{D}{2}\log(2\pi e) + \sum_i^D \log \hat{\sigma}_i^2, \tag{7.12}$$

where D is the dimension of the bottleneck and $\hat{\sigma}_i^2$ represents the variational variance estimated by the model in the bottleneck.

Training

As it can be observed in Eq. (7.11), the solution is obtained by sampling in the bottleneck L times, where L is a hyperparameter. Consequently, $z^{(i,l)} = f(x^{(i)}; \phi, \epsilon)$ corresponds to the set of l elements in the batch $x^{(i)}$ of size K. The ϵ represents the distribution of the noise injected by the Reparameterization Trick depicted in Fig. 7.3b.

Considering the batch size, K, and the sampling size, L, the loss function Eq. (7.11) for a batch is defined as:

$$\mathcal{L}^{B}_{\theta,\phi}(x^{(i)}) = -\frac{1}{L}\sum_{1}^{L} \log p_{\theta}(x^{(i)}|z^{(i,l)}) + D_{KL}(q_{\phi}(z|x^{(i)})\|p(z|y^{(i)})). \tag{7.13}$$

However, Kingma *et al.* concluded that the sampling is not necessary if the number of samples per batch is large [55]. For instance, the authors mentioned that a batch size of $K = 100$ is considered adequate. Finally, the training process is described in Algorithm 3.

Algorithm 3 ADeLEn training process

```
$\theta, \phi \leftarrow$ Initialize parameters
while not converged do
    $X_b \leftarrow$ Generates random batches of $M$ datapoints
    for Batches, $x_b^{(i)} \in X_b$ do
        $\epsilon^{(i)} \leftarrow$ Random samples from noise distribution $p(\epsilon)$        ▷ see Fig. (7.3b).
        $g \leftarrow \nabla_{\theta,\phi}\mathcal{L}^B(x_b^{(i)}; \epsilon^{(i)})$        ▷ Gradients of batch estimator.
        $\theta, \phi \leftarrow$ Update parameters        ▷ e.g. SGD or Adam.
    end for
end while
```

7.3 Information Theoretical Perspective on ADeLEn

As done in previous chapters and in line with the objectives of this thesis, this section provides an IT perspective on the proposed ADeLEn for AD in medical imaging. The study of the theoretical foundations of DL is still an ongoing research effort, where IT has emerged as a key approach for improving the interpretability of these 'black-box' models, as illustrated throughout this thesis.

In Chapter 4, it was demonstrated that AE follows the Infomax Principle (see Section 3.5) and the regularization in this architecture is described by $\mathcal{H}(Z)$, see Eq. (4.6) [190]. In other words, an AE aims to increase the MI between the input X and the compressed representation Z.

In VAEs, the InfoMax principle (see Section 3.5) follows the same condition that in the AEs. However, VAEs have an additional constraint on $\mathcal{H}(Z)$ that forces the model to obtain a manifold — a connected region that locally appears to be a Euclidean space —[21]. In summary, a VAE achieves higher compression where the data 'points' are 'compactly' arranged in a connected region. Given this 'compactness', VAEs are constrained to an entropy minimization in the latent space Z, which results in a reduction of $\mathcal{I}(X; Z)$ due to the compression.

Given that Z is constrained by the mean value and the variance ($\phi \in [\mu, \sigma]$), the $\mathcal{H}_\theta(Z)$ (the entropy of the estimated marginal distribution $p_\theta(Z)$ is subject to an upper limit that corresponds to a Gaussian distribution [157]. It fits with the assumption proposed in ADeLEn where the prior distribution $p(Z)$ is a Gaussian distribution. As a result, the upper limit is defined as:

$$\mathcal{H}_\theta(Z) = E[-\log p_\theta(Z)] \leq \frac{1}{2}\log((2\pi e)^D \cdot |\Sigma|), \tag{7.14}$$

where $|\Sigma|$ is the determinant of the covariance matrix. However, given the assumption that Z follows an isotropic Gaussian, see Eq. (7.7), the upper bound can be simplified to:

$$\mathcal{H}_\theta(Z) \leq \frac{D}{2}\log(2\pi e\sigma) \propto \log\sigma\,, \tag{7.15}$$

given $|\Sigma| = |\sigma I| = \sigma^D$, where σ is the variance and D is the dimension of the multivariate Gaussian distribution. Given the aforementioned inequality, the KL Divergence of both distribution is described as [157]:

$$D_{KL}(p_\theta(Z)\|p(Z)) = -\mathcal{H}_\theta(Z) + \mathcal{H}(Z). \tag{7.16}$$

The upper bound described in Eq. (7.15) can be intuitively understood given the ELBO used in VAEs. The regularization term in Eq. (7.8) forces a reduction in the 'distance' between $q_\phi(z|x)$ and the prior $p(z)$, which implicitly is a constraint in the $p_\theta(z|x)$ given $p_\theta(z|x) \approx q_\phi(z|x)$. Consequently, the VAE minimizes the empirical variance in ϕ to minimize the upper bound depicted in Eq. (7.15), reducing the entropy and obtaining the data manifold.

The explanation using elements of IT, in addition to the DPI that describes the AEs, is fundamental for describing a generative model such as VAE, GAN or DPM. However, it is worth noting that the proposed ADeLEn is a semi-supervised method. Therefore, it is necessary to incorporate this approach within the InfoMax principle.

InfoMax in ADeLEn

As discussed in Section 7.2.2, the proposed ADeLEn method uses a semi-supervised approach that includes a subset of labeled anomalous data. It is assumed that most of the unlabeled data corresponds to normal samples. For easier annotation, the normal samples are denoted as $p(z^+) = p(z|y=0)$, and the anomalous samples are denoted as $p(z^-) = p(z|y=1)$, where $y \in \{0,1\}$ represents the label.

Given that labels play a fundamental role, this can be addressed as a classification problem where the Information Bottleneck method provides an explanation of the problem-solving mechanism. However, considering exclusively the latent space Z, the proposed ADeLEn aims to reduce the CE in Z, discriminating between normal and anomalous samples.

As observed in Eq. (3.9), the CE — $CE(p,q)$, measures of the uncertainty of p by observing q — corresponds to a reduction of the KL Divergence. Consequently, discrimination between Z^+ and Z^- is achieved by maximizing the $D_{KL}(p(Z^+)\|p(Z^-))$. However, given that the model must satisfy Eq. (7.15) and Eq. (7.7), the KL Divergence follows the Eq. (7.16):

$$D_{KL}(p(Z^+)\|p(Z^-)) = -\mathcal{H}(Z^+) + \mathcal{H}(Z^-). \tag{7.17}$$

As a result, the proposed ADeLEn method for AD is described by the following InfoMax principle:

$$\max_{p(z|x)} \mathcal{I}(X;Z) + \beta\left(\mathcal{H}(Z^-) - \mathcal{H}(Z^+)\right). \tag{7.18}$$

Parameter initialization based from the IT perspective

As mentioned above, the VAE aims to minimize the upper bound described in Eq. (7.15) to obtain the data manifold. The estimated variational parameters ϕ in the VAE are obtained by the hidden layer, where one neuron estimates μ and another estimates σ (see Fig. 7.3b). However, it is

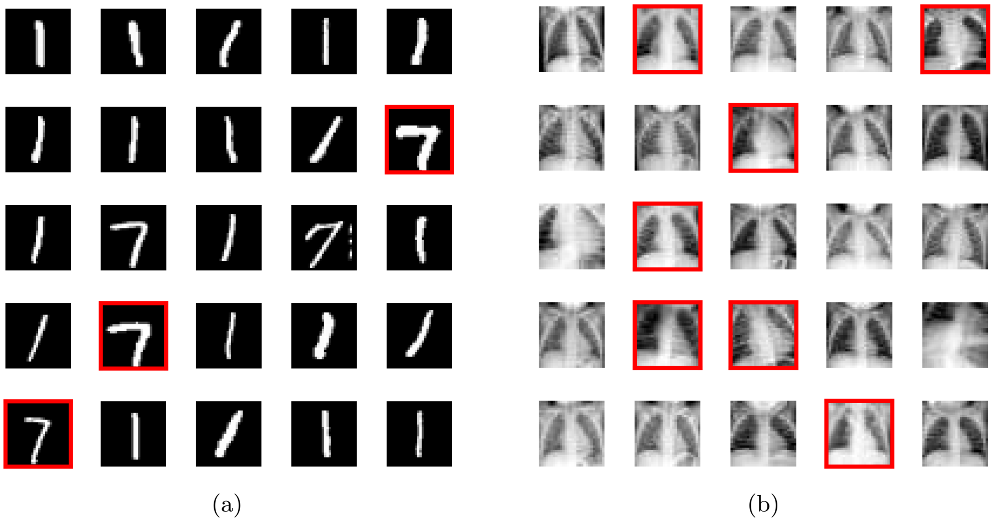

(a) (b)

Figure 7.4: Subset of samples of the dataset for evaluating ADeLEn method using MNIST (a) and the 'PneumoniaMNIST' (b). The red square represents samples that are known anomalies and, as it is easily illustrated in (a), the normal samples are contaminated with non-labeled anomalies.

intuitive to understand that initially, the upper bound should be higher and is reduced through training. Therefore, a specific initialization of the parameters of the hidden layer that generates the variational parameters is proposed.

As illustrated in Eq. (7.15), the entropy in a Gaussian distribution depends exclusively on σ. Given that the neuron to obtain the σ is described by a function $\sigma = W_\sigma^T X + b_\sigma$ (see Section 2.2.1), it is necessary to ensure a higher initial value. For this reason, the following initialization is proposed:

$$\begin{aligned} W_\sigma &\sim \mathcal{U}\left[-\left(\frac{1}{\sqrt{n}}\right), \left(\frac{1}{\sqrt{n}}\right)\right] \\ b_\sigma &= 2 \cdot \log(2\sigma^-)\,, \end{aligned} \tag{7.19}$$

where the W_σ initialization corresponds to the Xavier uniform initialization method [375], and n is the number of inputs of the neuron. As a result, the initial estimation should correspond to the double of the anomaly distribution, which has the higher entropy as it is observed in Eq. (7.17).

7.4 Experimental Results

In this section, the proposed ADeLEn will be subject to experimental evaluation by using different datasets. To understand the insights of ADeLEn, it will be used the MNIST [48] as a toy experiment and a dataset extracted from MedMNIST [376], a large-scale MNIST-like collection of standardized biomedical images. From MedMNIST, it will be used the known 'PneumoniaMNIST' [377], that corresponds to X-ray images for the diagnosis of pneumonia.

For the simplest case, using the MNIST dataset as a toy experiment, a subset of samples composed of two numbers will be considered. Using two numbers that share similarities, the aim is to illustrate how the latent space is mapped in the proposed ADeLEn. In this case, a subset of numbers composed of 1's and 7's is used, where the 7's represent the anomalies. The number of normal samples, 1's,

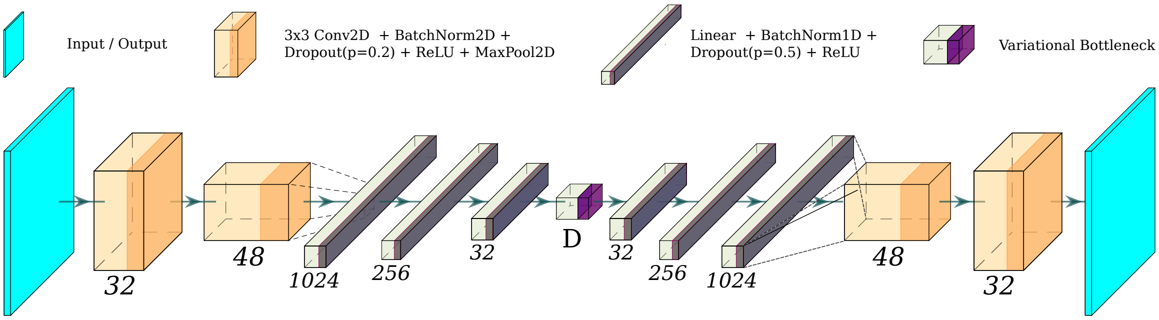

Input / Output
3x3 Conv2D + BatchNorm2D + Dropout(p=0.2) + ReLU + MaxPool2D
Linear + BatchNorm1D + Dropout(p=0.5) + ReLU
Variational Bottleneck
32
48
1024
256
32
D
32
256
1024
48
32

Figure 7.5: The proposed architecture for evaluating the ADeLEn method in the experimental setup.

has been limited to 2000 samples.

Regarding the 'PneumoniaMNIST' dataset, this dataset is composed of 5232 samples and is a binary classification dataset. Specifically, it consists of 1349 normal samples and 3883 samples characterized as pneumonia (2538 bacterial and 1345 viral, the two main leading causes of pneumonia [378, 377]). Given the purpose of this chapter, which is oriented towards AD, the experimental setup will not discriminate between bacterial or viral pneumonia. However, it is worth noting that both cases can be identified from the X-ray images. Bacterial pneumonia requires antibiotic treatment, while viral pneumonia is treated with supportive care [377].

Using the aforementioned datasets, various experiments will be conducted to analyze the behavior of the proposed ADeLEn and compare it with other approaches. The robustness of ADeLEn will be investigated regarding an increasing pollution ratio, i.e., an increase in unidentified anomalous samples contained within the 'normal' unlabeled samples. Another experiment will consider the effect of increasing the number of known anomalies by adding labeled anomalies in the training process. Additionally, the experiments will examine how increasing the bottleneck dimension D affects the performance.

The DNN used in this section remains consistent across the various experiments. It consists of an AE with an encoder path composed of two convolutional layers, with 32 and 48 channels, followed by three linear layers with 1024, 256, and 32 neurons. The bottleneck corresponds where the variational parameters described by the dimension D are estimated, see Fig. 7.3. The decode path mirrors the encode path. The proposed architecture is illustrated in Fig. 7.5. The hidden layers use ReLU as the activation function, batch normalization, and the dropout technique. The dropout rate for the convolutional layers is 0.2, while the dropout rate for the linear layers is 0.5, as shown in Fig. 7.5.

The training process of the ADeLEn, described in Algorithm 3, uses a batch size of 128 samples, as a large batch size helps to discard the of $q_\phi(z|x)$ [55] as it was previously mentioned. Regarding the optimizer, as in the previous chapters, the model training is carried out using the Adam optimizer [65] (see Section 2.2.2). Finally, the learning rate has been set to 10^{-3} and the model is trained for 50 epochs.

Additionally, an experiment with higher-resolution images was conducted using the BRAin Tumor Segmentation Dataset (BRATS) dataset, originally designed for brain tumor segmentation and adapted here as an AD task. In this experiment, the DNN architecture follows that shown in Fig. 7.5, with the addition of two initial convolutional layers and a specific bottleneck size. The training process is configured identically to that of the other experiments.

7.4.1 Pollution

In this section, the effect of pollution will be examined, which involves the injection of anomalous samples into the non-labeled set that is assumed to correspond to normal samples, within the dataset. This experiment involves using different configurations of pollution levels while maintaining a fixed number of labeled anomalous samples. Specifically, the number of labeled anomalous samples is set to 10% of the non-labeled dataset. Regarding pollution, four different cases will be considered: 0% (no pollution in the dataset), 5%, 10%, and 20% of the non-labeled samples.

The proposed ADeLEn will be compared with a well-known method used in the state-of-the-art for AD, the oneOC-SVM. Nevertheless, given that the OC-SVM is an unsupervised method and the comparison may be deemed unfair, the hyperparameters of the OC-SVM have been configured utilizing the test set. Specifically, the hyperparameters were tuned to obtain the highest Area Under the Curve (AUC) using a subset of the test set, which consists of 10% of the test set. In addition, a naive approach will be used where a supervised DNN is employed for binary classification. This DNN follows the architecture of the encode path without including the bottleneck used in the proposed architecture of ADeLEn in these experiments, see Fig. 7.5. Instead, a single neuron layer is added to the output, indicating that the model corresponds to a binary classifier.

For the three methods used in this experiment (the OC-SVM, the supervised DNN, and the proposed ADeLEn), it is necessary to apply a threshold based on the obtained score. Regarding the OC-SVM, this work uses the implementation in the Scikit-learn library [379], which has a specific method for obtaining the optimal threshold based on the score obtained during the fitting process. For the supervised DNN, where the score corresponds to the model's output, the threshold will be fixed at 0.5, the commonly used threshold in binary classifiers.

In the case of the proposed ADeLEn, which uses the entropy of the latent space as a score (see Eq. (7.12)), a fixed threshold of $\sigma = 1.2$ has been set. In this way, an anomaly is identified by ADeLEn if the following condition is satisfied:

$$\mathcal{H}(Z|x) \geq \frac{D}{2} \log(2\pi e \cdot 1.2), \tag{7.20}$$

where D is the dimension of Z. Given that the proposed ADeLEn in this section will be use a $D = 2$, the threshold corresponds to, approximately, 3.02.

MNIST

Firstly, the different methods will be evaluated using the MNIST dataset [48], utilizing the subset of samples described in Section 7.4. As previously mentioned, this represents a simple toy experiment that makes it easy to identify if the models are functioning correctly. The results presented here have been obtained using a test set that is not applied during the training of the model.

Table 7.1 depicts the quantitative results obtained by the different methods using different levels of pollution in the unlabeled set assumed to be normal. Considering the AUC, it is demonstrated that the different methods are suitable for discriminating between anomalous and normal samples with varying levels of pollution. Specifically, the worst case corresponds to the unsupervised OC-SVM with an AUC of approximately 98%. Here, the worst case is the supervised DNN, given that the threshold used during training works well during training, but using the test set reveals that this threshold does not effectively identify the anomalous samples in unobserved data. As a result,

Table 7.1: MNIST Metrics Anomaly Detection Pollution. The DNN represents a supervised DNN method.

Method	Pollution	Accuracy	Precision	Recall	F1-Score	AUC
ADeLEn	0	98.63 ± 2.22	98.44 ± 4.00	98.97 ± 0.38	98.61 ± 2.11	99.84 ± 0.10
	5	97.99 ± 0.69	99.20 ± 1.45	96.59 ± 1.39	97.86 ± 0.72	99.74 ± 0.15
	10	96.08 ± 1.83	98.86 ± 2.23	92.94 ± 4.45	95.72 ± 2.13	99.65 ± 0.15
	20	84.94 ± 6.91	99.60 ± 0.68	68.64 ± 14.84	80.32 ± 10.96	99.5 ± 0.04
OC-SVM	0	75.85 ± 1.52	100 ± 0	53.99 ± 2.91	70.07 ± 2.50	99.13 ± 0.12
	5	77.07 ± 1.63	100 ± 0	56.30 ± 3.11	72.01 ± 2.59	98.93 ± 0.11
	10	96.08 ± 1.83	98.85 ± 2.23	92.94 ± 4.45	95.71 ± 2.13	99.65 ± 0.15
	20	78.34 ± 1.91	100 ± 0	58.73 ± 3.65	73.93 ± 2.98	98.69 ± 0.09
DNN	0	99.30 ± 0.37	99.95 ± 0.01	98.57 ± 0.82	99.61 ± 0.41	99.97 ± 0.00
	5	89.67 ± 4.39	99.95 ± 0.01	78.31 ± 9.23	87.52 ± 6.02	99.90 ± 0.00
	10	78.75 ± 7.25	99.95 ± 0.00	55.32 ± 15.29	70.06 ± 12.27	99.84 ± 0.01
	20	64.99 ± 5.18	99.95 ± 0.01	26.32 ± 10.91	40.48 ± 14.35	99.61 ± 0.02

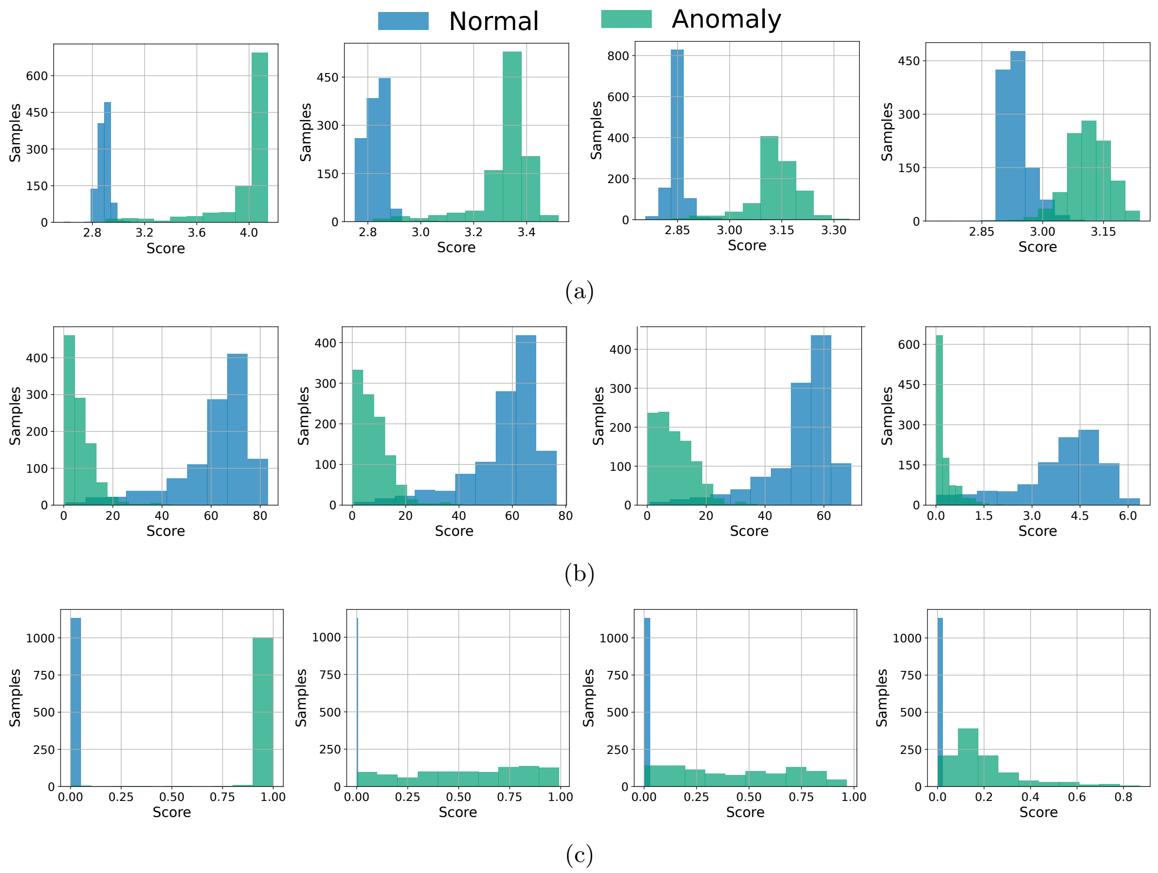
using F1-Score as reference, it is illustrated that the proposed ADeLEn obtains the better results with higher pollution.

Figure 7.6 illustrates how the pollution affects the scores obtained by the different methods. In Fig. 7.6a, it is shown that the score difference between normal and anomalous samples narrows as pollution increases in the unlabeled dataset. In other words, the KL Divergence between Z^+ and Z^- is reduced, demonstrating how pollution leads to a worse solution depicted by the Eq. (7.18) (see Section 7.3). This issue is observed across the different methods, with OC-SVM demonstrating better tolerance to pollution (see Fig. 7.6b) compared to the supervised DNN (see Fig. 7.6c).

Observing Fig. 7.6c, it can be deduced that the supervised DNN has a problem of overfitting. It is observed that when there is no pollution, the model can discriminate perfectly both classes but when there is pollution, given that the model has not able to identify the anomalies, the score is reduced. Finally, with pollution of 20%, the model can only identify around the 21% of anomalous samples. Remember that it has been used the dropout technique that it is mainly used for reducing the overfitting, although it has been demonstrated in Chapter 4 that has another purpose in the compression. However, even using dropout technique, this model cannot mitigate the problem of overfitting for this toy experiment.

Finally, Fig. 7.7 illustrates the distribution of the latent representation obtained by ADeLEn at different levels of pollution. Figure 7.7a shows the ideal case: the normal data is well-defined, and there is a clear distinction between normal and anomalous samples. Here, the data manifold is well-illustrated, with normal samples (1's, as illustrated in Fig. 7.4a) located at the center. There is a linear interpolation between the 1's and 7's, illustrating how a 1 gradually transforms into a 7 as it moves away from the center. Additionally, it is shown how two dimensions are sufficient to characterize the data manifold, resulting in a symmetric representation in the image.

However, when pollution is injected into the dataset, this symmetry disappears, and the representation of the anomalies (7's) is observed at the borders of the images. This behavior is noted in Fig. 7.7b, but it is more pronounced in the worst case with 20% pollution, as shown in Fig. 7.7c. In this figure, it is observed that the border defined by the coordinates (-6, -6) does not represent a 7 but rather a 1 with a slight disturbance at the top. Despite this disturbance being indicative



(a)

(b)

(c)

Figure 7.6: Pollution experiments with (a) ADeLEn, (b) OC-SVM and (c) a supervised DNN using the MNIST dataset. The experiments demonstrate how the pollution in the data affects the different methods. The pollution, from left to right, corresponds to 0%, 5%, 10% and 20%.

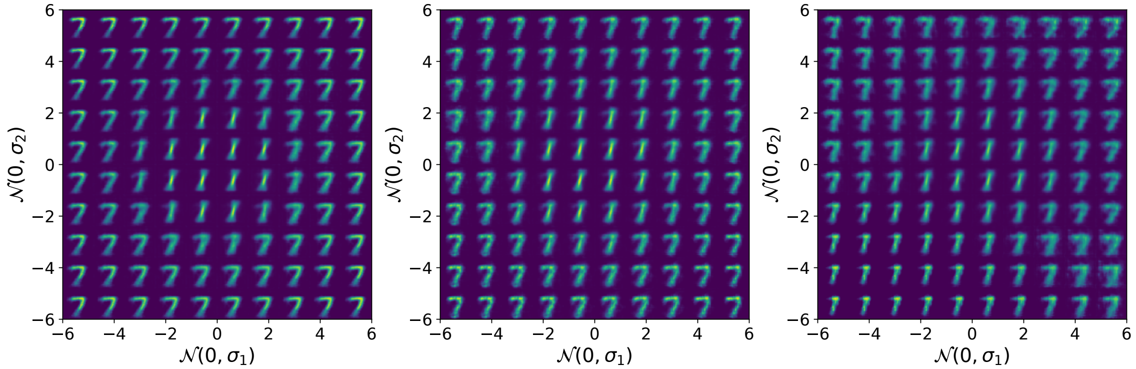
of anomalies, the model can still function for AD. However, as observed in Fig. 7.6a, the difference between normal and anomalous samples becomes less noticeable.

MedMNIST

The same comparison has been carried out using the aforementioned 'PneuomoniaMNIST' dataset, see Fig. 7.4b. As can be deduced, this problem is more complicated than the toy experiment using MNIST. For instance, observing Fig. 7.4b, it is complicated to identify the unlabled anomalous samples in the figure. Additionally, the number of normal' samples is lower than in the MNIST experiment. In this case, the number of 'normal' samples is 1349, whereas in MNIST, the number of normal samples was set to 2000.

Table 7.2 illustrates the results obtained using the different methods. Firstly, it is noticeable that the unsupervised OC-SVM is not suitable for discriminating between normal and anomalous samples, as observed in the AUC, even though the hyperparameters have been optimized using a subset of the test set as previously mentioned. Regarding the proposed ADeLEn and the supervised DNN, both methods have similar performance, with the supervised DNN performing slightly better.

Figure 7.8 illustrates the issues observed in Table 7.2. Figure 7.8b shows that OC-SVM cannot discriminate between normal and anomalous samples, as indicated by the AUC in Table 7.2. he



(a) (b) (c)

Figure 7.7: Reconstruction obtained by the compressed 2-dimensional latent space in the VAE used in ADeLEn. The results are depicted with different levels of pollution: (a) 0%, (b) 10% and (c) 20%.

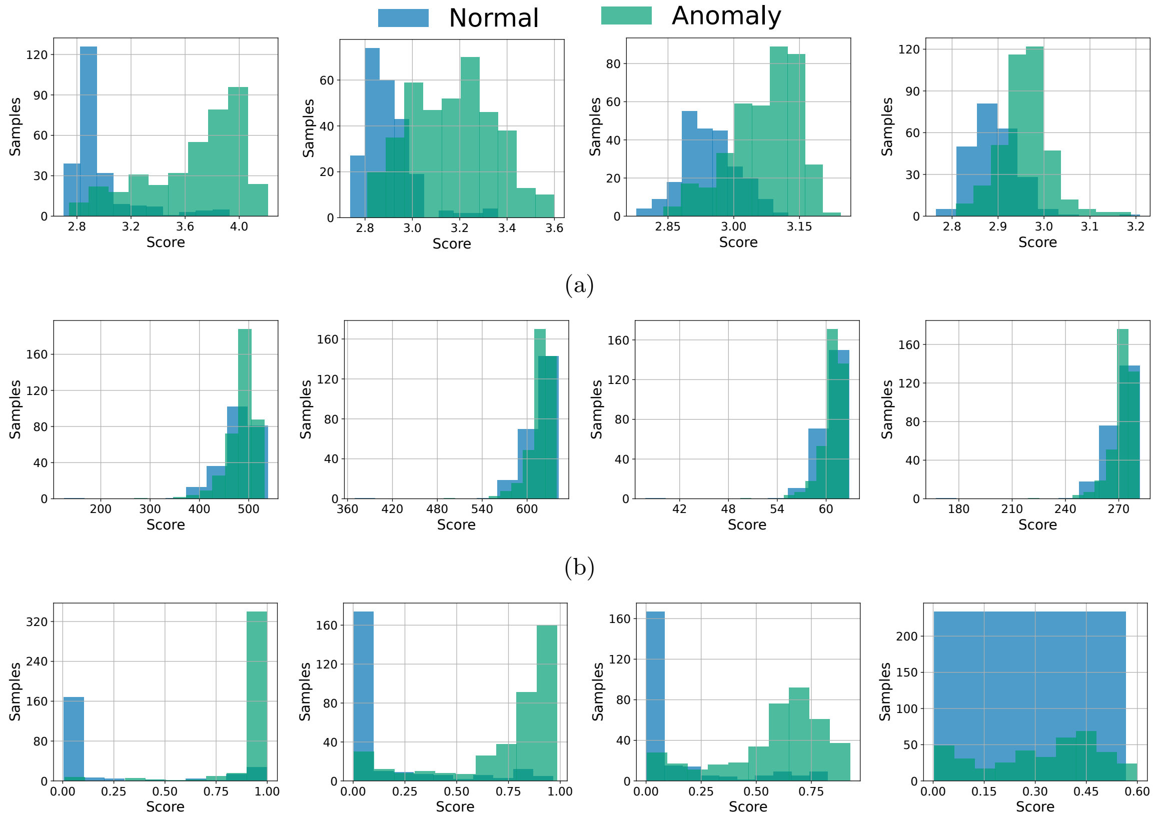
Table 7.2: MedMNIST Metrics Anomaly Detection Pollution. DNN represents a supervised DNN method.

Method	**Pollution**	**Accuracy**	**Precision**	**Recall**	**F1-Score**	**AUC**
ADeLEn	0	86.91 ± 3.19	86.36 ± 3.85	94.27 ± 1.87	90.07 ± 1.90	93.47 ± 0.74
	5	83.51 ± 2.44	91.56 ± 3.54	81.38 ± 4.90	86.01 ± 2.26	91.34 ± 1.52
	10	73.17 ± 7.29	93.45 ± 3.87	61.77 ± 13.34	73.35 ± 10.26	88.43 ± 3.39
	20	43.53 ± 5.06	87.57 ± 22.31	10.57 ± 7.99	18.12 ± 12.31	84.08 ± 4.51
OC-SVM	0	45.85 ± 8.08	36.17 ± 0.97	58.22 ± 28.93	41.81 ± 9.77	46.87 ± 1.49
	5	44.59 ± 7.07	34.45 ± 1.52	55.65 ± 27.30	40.41 ± 9.51	44.90 ± 1.72
	10	42.17 ± 5.83	33.63 ± 2.11	59.82 ± 25.27	41.42 ± 8.92	43.24 ± 1.59
	20	41.98 ± 4.92	33.03 ± 2.71	57.86 ± 26.06	40.69 ± 9.18	40.91 ± 1.59
DNN	0	88.70 ± 1.31	90.15 ± 2.46	92.13 ± 3.17	91.06 ± 1.08	94.29 ± 0.85
	5	81.23 ± 5.18	92.84 ± 3.03	76.18 ± 10.78	83.13 ± 5.92	91.82 ± 2.03
	10	71.30 ± 6.98	93.70 ± 3.34	58.36 ± 13.08	70.89 ± 10.32	90.57 ± 2.86
	20	49.35 ± 6.20	90.15 ± 8.49	21.20 ± 10.63	33.18 ± 13.31	88.51 ± 3.40

supervised DNN, shown in Fig. 7.8c, appears to avoid overfitting. The only issue arises when the pollution level is 20%, but in the other experiments, the behavior is as expected, and the samples can be discriminated.

Regarding the proposed ADeLEn, shown in Fig. 7.8a, , there are overlaps in the distributions of the normal and anomaly scores. In the case with no pollution (the first column in Fig. 7.8), some anomalous samples have the lowest scores, while some normal samples have much higher scores. This behavior appears in both DL-based approaches, ADeLEn and the supervised DNN. The overlap becomes more severe as pollution increases. This suggests that there may be incorrect annotations in the original training or test set of the 'PneumoniaMNIST' dataset [376].

In Fig. 7.9 illustrates the distribution of the latent representation obtained by ADeLEn at different levels of pollution. However, interpreting this latent representation is not as straightforward as in Fig. 7.9.In the case without pollution, Fig. 7.9a, the image shows a symmetry in the latent representation. Although the center appears well-defined, the anomalous samples are not clearly delineated, resulting in a noisy representation of anomalies.

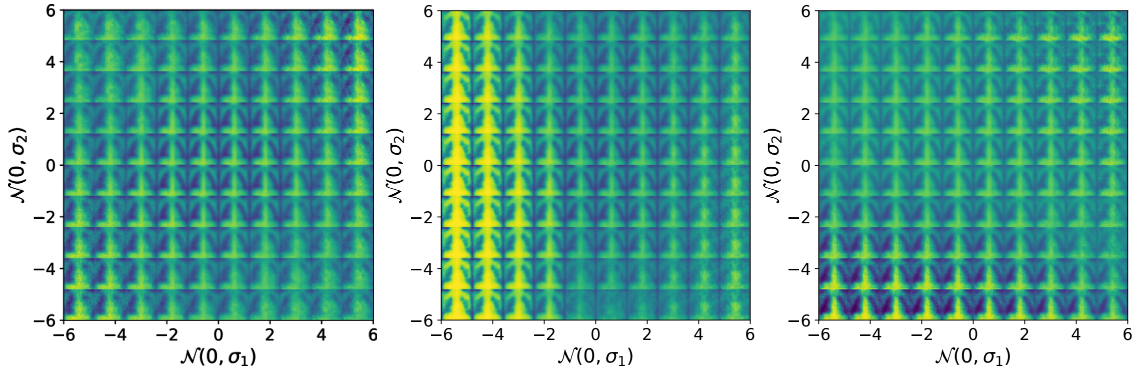


(a)

(b)

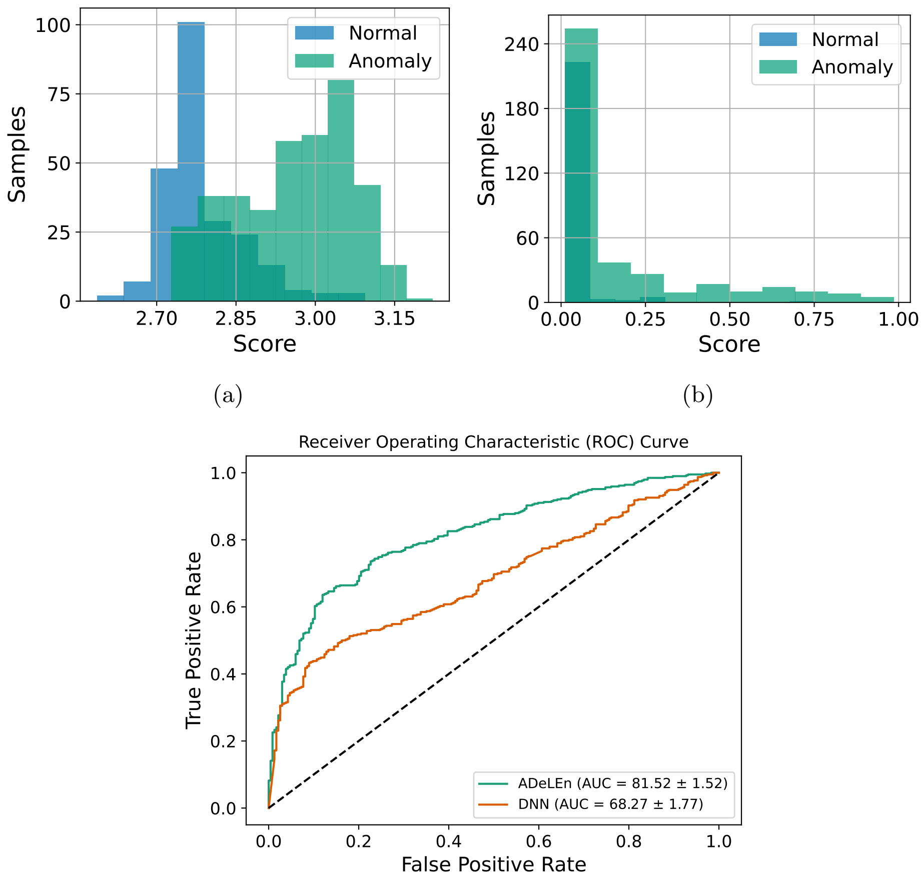
(c)

Figure 7.8: Pollution experiments with (a) ADeLEn, (b) OC-SVM and (c) a supervised DNN using the MedMNIST dataset. The experiments demonstrate how the pollution in the data affects the different methods. The pollution, from left to right, corresponds to 0%, 5%, 10% and 20%.



(a) (b) (c)

Figure 7.9: Reconstruction obtained by the compressed 2-dimensional latent space in the VAE used in ADeLEn. The results are depicted with different levels of pollution: (a) 0%, (b) 10% and (c) 20%.



(a)

(b)

(c)

Figure 7.10: Qualitative results using Monte-Carlo Dropout in ADeLEn and the supervised DNN for anomaly detection in the MedMNIST dataset. The scores obtained by ADeLEn and the supervised DNN are illustrated in (a) and (b), respectively. The ROC curve and the average AUC for the different models are illustrated in (c).

In the presence of pollution, both the 10% pollution case (Fig. 7.9b) and 20% (Fig. 7.9c) show that the symmetry is lost. Observing the center in these polluted cases, it appears that the model has not been able to clearly define the normal samples.

7.4.2 Overfitting

Although overfitting is not a significant concern due to the Bayesian approach and the application of the dropout technique primarily for data compression, a simple test will be conducted to evaluate the difference between the supervised DNN and the proposed ADeLEn. As observed in the previous section, the supervised DNN appears to achieve better results for AD on the MedMNIST dataset. This is particularly noticeable in the qualitative results shown in Fig. 7.8.

As previously mentioned in Section 2.3.2, Gal and Ghahramani [74] provide an interpretation of the dropout technique in DL models based on deep GP. As a result, they proposed Monte-Carlo Dropout [74] to estimate the uncertainty in DL models. Although not perfectly accurate, this approach provides a computationally efficient estimation of the uncertainty in the model's decisions.

Using Monte-Carlo Dropout, a simple test will be conducted to evaluate its effect on the models. Figure 7.10 illustrates the results obtained when Monte-Carlo Dropout is used, sampling 100 times. The experiments have been conducted with a pollution level of 5%, so the scores should be compared with the second columns in Fig.7.8.

Figure 7.10a shows how the 'noise' in the model affects the proposed ADeLEn. It is observed that the results do not differ significantly from those obtained without applying Monte-Carlo dropout (see Fig. 7.8a, second column). However, the effect is dramatic in the supervised DNN, as illustrated in Fig. 7.10b, which performs considerably worse than what is seen in the second columns of Fig. 7.8c. Finally, the impact on the AUC is illustrated in Fig. 7.10c.

This experiment highlights the expected problems in the supervised DNN, particularly overfitting, even after using the dropout technique. In contrast, ADeLEn, due to its Bayesian approach, can handle these issues more effectively. This is especially relevant to ensure that the model can deal with unobserved new anomalies or noisy inputs and provides and provides greater confidence as an AD method.

7.4.3 Number of Anomalies and Bottleneck Dimension

In Section 7.4.1, it has been demonstrated that the proposed ADeLEn is effective for AD. The results obtained using the MedMNIST dataset suggested that a simple supervised DNN might perform better. However, as shown in Section 7.4.2, this apparent superiority is likely a consequence of overfitting. With a larger test set, the performance of the supervised DNN would probably decline. On the other hand, ADeLEn has shown robustness and effectiveness even when noise is introduced during inference through Monte-Carlo dropout.

In this section, experiments are conducted to investigate two important features of the proposed ADeLEn method: the bottleneck size and the number of known anomalies. Previous sections used a 2-dimensional bottleneck (illustrated in Fig. 7.7 and Fig. 7.9) and set the number of known anomalies (labeled data) to 10% of the non-labeled data assumed to be normal. For instance, in the MNIST dataset, with 2000 non-labeled samples, the number of known anomalous samples was set to 200.

Table 7.3 depicts the results obtained by varying the bottleneck dimension size (set to 2, 5, and 10) and increasing the number of labeled data from 5% to 50%, progressively. In the experimental results, it has been applied pollution in the non-labeled data of 5%.

The quantitative results demonstrate that using the MedMNIST dataset for pneumonia detection, both increasing the number of labeled data and the bottleneck size improve the results. However, as expected, the number of labeled data is the most significant factor. While increasing the bottleneck size generally enhances the classification metrics, the impact is not particularly substantial. For example, with 10% labeled data, there is an improvement of around 2% in the F1-Score.

Figure 7.11 illustrates an example of scores obtained by increasing the number of labeled data (from left to right) and the bottleneck dimension, D. As observed in Table 7.3, D has been evaluated for 2 (Fig. 7.11a), 5 (Fig. 7.11b) and 10 (Fig. 7.11a). These figures confirm the findings in Table 7.3, where the most significant improvement is due to the increase in labeled data.

From Fig. 7.11, it is noticeable that using 20% labeled anomalous samples substantially improves the model, making the score more discriminative and reducing the overlap between distributions. The threshold used in the experimental results, as depicted by Eq. (7.20), corresponds to 3.02

Table 7.3: MedMNIST Metrics Anomaly Detection by increasing the number of labeled data and the dimension of the bottleneck in the proposed ADeLEn. The results are obtained by injecting pollution of 5% in the non-labeled data.

Dimension	**N. Labels**	**Accuracy**	**Precision**	**Recall**	**F1-Score**	**AUC**
2	0.05	51.25 ± 10.10	89.24 ± 17.48	24.21 ± 18.59	35.06 ± 21.17	82.46 ± 0.03
	0.10	83.36 ± 2.45	91.46 ± 3.41	81.25 ± 5.71	85.84 ± 2.49	91.15 ± 1.22
	0.20	87.50 ± 1.31	88.28 ± 2.15	92.36 ± 2.35	90.23 ± 0.99	92.71 ± 1.71
	0.25	87.51 ± 1.60	87.02 ± 2.34	94.17 ± 1.87	90.42 ± 1.12	92.93 ± 1.16
	0.50	88.94 ± 1.18	86.84 ± 1.54	97.07 ± 0.78	91.66 ± 0.81	91.88 ± 2.25
5	0.05	56.38 ± 8.48	94.78 ± 6.17	31.58 ± 14.17	45.71 ± 17.08	85.85 ± 0.02
	0.10	84.90 ± 1.99	91.20 ± 1.94	84.04 ± 4.28	87.38 ± 1.94	91.74 ± 1.38
	0.20	87.78 ± 1.74	87.71 ± 2.68	93.72 ± 2.58	90.56 ± 1.28	92.34 ± 1.83
	0.25	88.34 ± 1.49	87.60 ± 2.59	94.92 ± 2.64	91.05 ± 1.10	92.73 ± 1.58
	0.50	89.15 ± 1.16	86.90 ± 1.67	97.35 ± 0.93	91.81 ± 0.79	91.65 ± 2.17
10	0.05	57.06 ± 11.02	96.96 ± 4.50	32.22 ± 18.32	45.52 ± 21.28	87.94 ± 2.59
	0.10	85.20 ± 2.39	91.34 ± 2.49	84.46 ± 4.72	87.65 ± 2.30	91.96 ± 1.62
	0.20	87.83 ± 1.41	87.72 ± 2.16	93.76 ± 2.98	90.58 ± 1.15	92.23 ± 1.54
	0.25	88.60 ± 1.62	88.26 ± 2.66	94.47 ± 2.89	91.19 ± 1.25	92.33 ± 1.87
	0.50	89.65 ± 1.68	87.89 ± 2.43	96.88 ± 1.10	92.14 ± 1.13	91.53 ± 2.38

for a 2-dimensional bottleneck, 7.55 for a 5-dimensional bottleneck, and 15.1 for a 10-dimensional bottleneck.

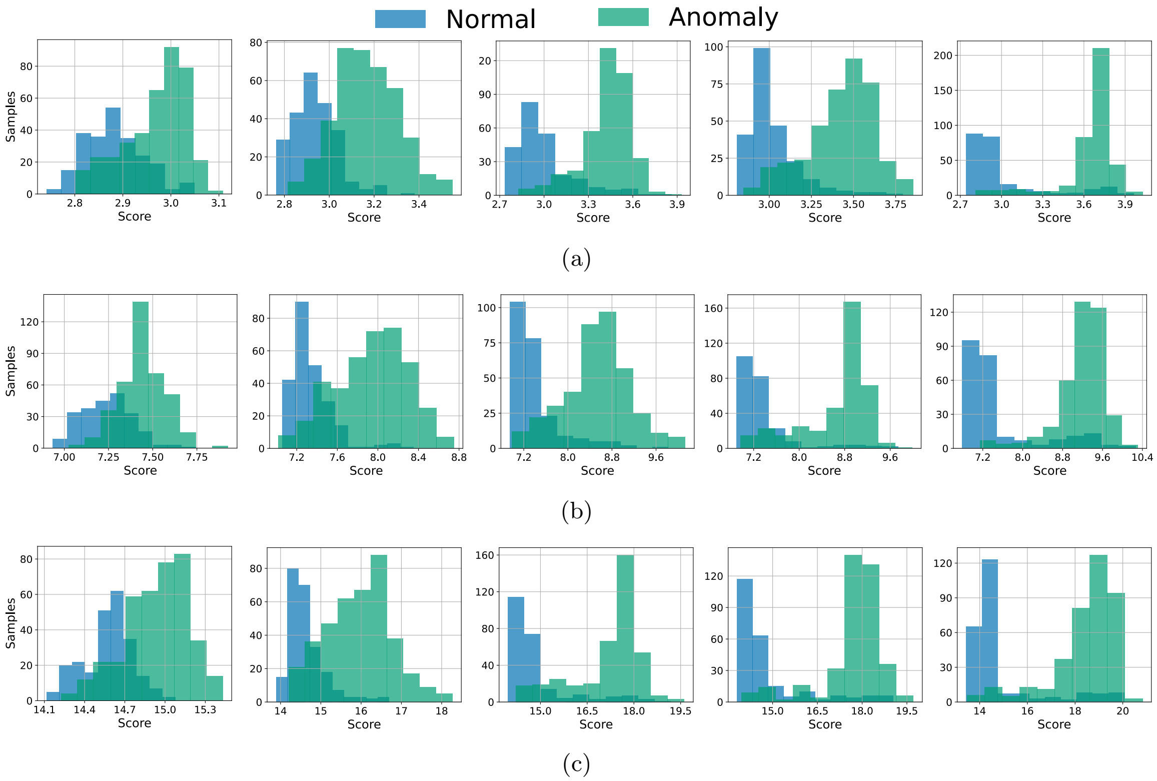
Figure 7.11 illustrates a example of scores obtained by increasing the number of labeled data, from left to right, and increasing the bottleneck dimension, D. As it was observed in Table 7.3, the D has been evaluated for 2 (Fig. 7.11a), 5 (Fig. 7.11b) and 10 (Fig. 7.11a). These figures illustrate the observed in Table 7.3, where the significant improvement is provided by the number of labeled data.

In addition, observing Fig. 7.11, it can be concluded that the threshold effectively identifies normal samples, as reflected in the precision metric in Table 7.3. However, as noted in Section 7.4.1, a 5% pollution means that using only 10% labeled data is insufficient. The model starts to perform properly when the labeled anomalous data is increased to 20%. Increasing the bottleneck dimension, in this particular case, does not appear to be significantly impactful.

7.4.4 Novelty detection

Another important aspect of AD is the ability to identify new, previously unobserved events, known as novelties. Novelty detection involves applying AD techniques and is commonly used in semi-supervised approaches. This section explores how the proposed ADeLEn can identify elements not seen during the semi-supervised training.

To study novelty detection, we will use the MNIST dataset. The experiment will involve a training set composed of images of the digits 1 and 7. The configuration will be the same as described in Section 7.4.1 Given that ADeLEn demonstrated no issues with this toy experiment, the dataset will consist of 2000 samples of the digit 1, with 10% of pollution injected. The number of anomalous labeled samples will be 10% of the normal samples. As a reference, this setup corresponds to the third column in Fig. 7.6a.



(a)

(b)

(c)

Figure 7.11: Labels and dimensional experiments with ADeLEn using D as (a) 2, (b) 5 and (c) 10. The dataset has been set using pollution of 5% in the 'normal' samples. The number of anomalous labeled samples, from left to right, corresponds to 5%, 10%, 20%, 25% and 50%.

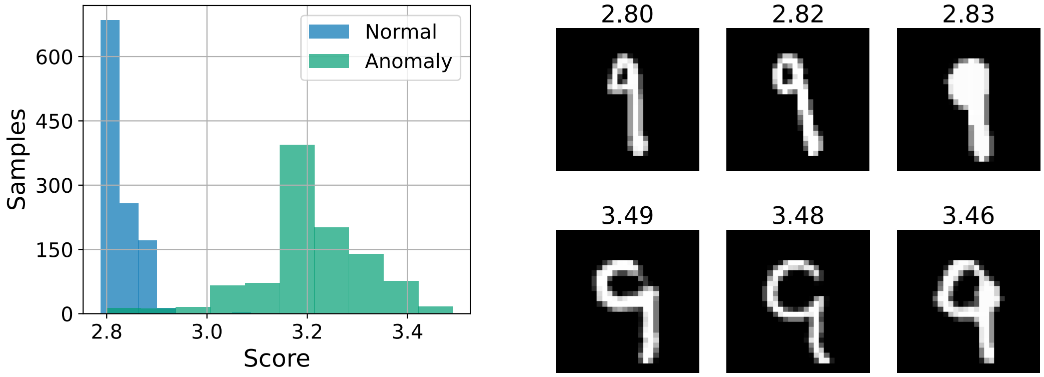
However, in the test set, the anomaly does not correspond to 7's. Instead, a test set composed of samples of the digits 1 and 9 will be used. The selection of 9's is based on the main hypothesis in the ADeLEn approach: normal and anomalous samples share similarities. This setup allows evaluating how well ADeLEn can detect novel anomalies that differ from the ones seen during training.

The quantitative results obtained has an AUC over 99% and the F1-Score is around 98%. In other words, the proposed ADeLEn can differ perfectly between normal samples and the unobserved anomalies, the 9's. A illustrative example is depicted in Fig. 7.12.

Figure 7.12a shows the scores obtained on the test set composed of 1's and 9's. Similar to the results with a test set containing anomalies observed during training (Fig. 7.6a), the model effectively discriminates between normal and anomalous samples. However, a subset of anomalies is misclassified with high confidence as normal. Some of these most confident misclassifications, with scores around 2.8, are shown in Fig. 7.12b. In these cases, the distinctive feature of the 9's (the 'disturbance') is almost indistinguishable, making them appear similar to 1's. When the 9's clearly depict this point of difference, the proposed ADeLEn successfully identifies them as anomalies. Based on these results, it can be concluded that the proposed model functions correctly.

7.4.5 Additional Experiment: Brain Cancer Detection

In this section, it will be carried out a experiment to evaluate the proposed ADeLEn using the BRATS [380] from the challenge of 2020. This dataset is composed of samples of MRI with dif-



(a)

(b)

Figure 7.12: Results using the proposed ADeLEn for novelty detection on a test set composed of anomalies not previously observed in the training set. In (a), the histogram illustrates the scores obtained by ADeLEn. An example of samples with their respective scores is shown in (b), where each sample's score is indicated in the title.

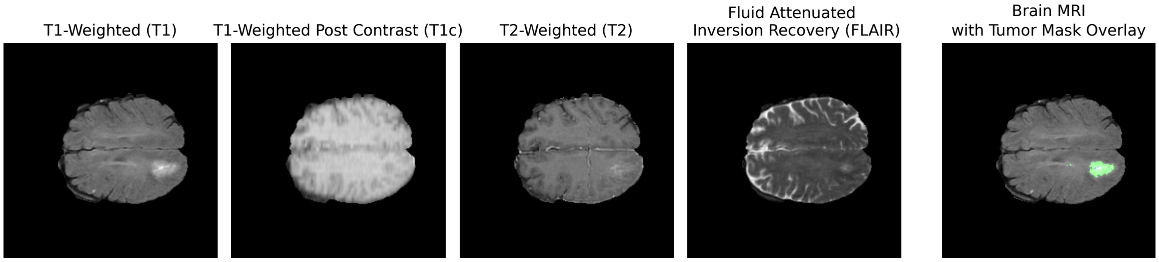
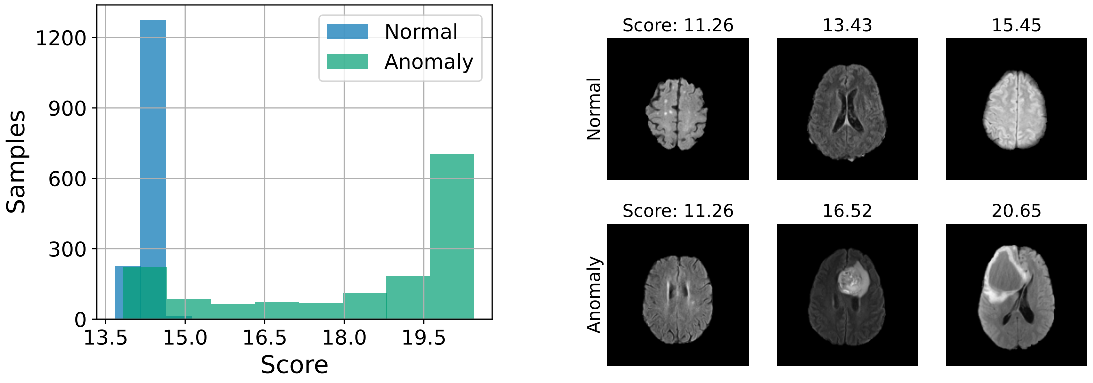


Figure 7.13: Example of MRI data in the BraTS dataset. The four modalities used are T1, T1c, T2, and FLAIR. In addition, it has been included the tumor mask overlaid on the MRI image.

ferent modalities, T1-weighted, T1-weighted with contrasts, T2-weighted and FLAIR. T1-weighted images are characterized by their ability to highlight differences in tissue composition, making them particularly useful for identifying fat and anatomical structures with high spatial resolution. T2-weighted images, on the other hand, excel at detecting fluid and edema, as they enhance the contrast of water-rich tissues, aiding in the visualization of various pathological conditions such as tumors and inflammation. FLAIR imaging is a specialized technique that suppresses the signal from free fluids like cerebrospinal fluid, making it highly effective for identifying lesions adjacent to these fluid spaces, such as in multiple sclerosis or brain infarction.

Figure 7.13 illustrates the four image modalities of the BRATS dataset alongside the tumor mask overlaid on the MRI. Given that the current architecture (refer to Fig. 7.5) is designed for 2D images, only the axial slices are considered. Each slice consists of four channels, corresponding to each image modality. These images are padded to a size of 256 × 256 and normalized between -1 and 1.

Since tumors primarily occur in the middle of the brain, slices from the lowest 80 and uppermost 26 slices are excluded. A slice is classified as healthy or normal if no tumor is present in the ground truth label mask. All other slices are labeled as diseased, i.e., anomalous samples. In total, there are extracted 17712 slices, where 6044 are normal samples and 11668 correspond to slices that contain pixels labeled as tumor. For training, it has been used the 75% of the normal samples that correspond



(a)

(b)

Figure 7.14: Results using the proposed ADeLEn for brain cancer detection in BRATS dataset. In (a), the histogram illustrates the scores obtained by ADeLEn. An example of samples with their respective scores is shown in (b), where each sample's score is indicated in the title.

Table 7.4: Classification metrics by using ADeLEn for anomaly detection in BRATS dataset.

Accuracy	**Precision**	**Recall**	**F1-Score**	**AUC**
81.58 ± 9.26	95.77 ± 5.91	65.38 ± 16.45	76.84 ± 14.40	89.34 ± 6.95

to 4533 and the number of labeled anomalies is the 10% of the normal samples for training, 453 samples. In total, 4986 samples compose the training set.

Regarding the ADeLEn architecture, given the higher resolution of the images compared to the previous experiments using MNIST and MedMNIST, it has been included two more convolutional layers with the respective activation function, dropout rate and pooling. These two layers have 8 and 16 filters, respectively. As a result, the beginning of the encode path is composed by 4 convolutional layers of 8, 16, 32 and 48 filters. The decode path has been modified too for keeping the symmetry in the architecture. The dimension of the bottleneck, d, was set to 10.

7.5 Discussion

Recently, the use of DL-based models for AD has gained significant attention (see Section 7.1.2) and, in this chapter, it has been proposed a method for AD based on DL model, referred to as ADeLEn. Similar to most DL-based AD methods, this approach utilizes an AE-based architecture. However, the main difference with most works in the state-of-the-art in DL-based AD methods lies in the anomaly score. Generally, AE-based AD methods provide a score derived from the output of the AE. In contrast, the proposed ADeLEn use a score based on the latent space Z, the compressed representation of X obtained by the AE. Essentially, while most AE-based methods use a score based on $p(x|z)$, our proposed method generates a score based on $p(z|x)$. this approach is supported by previous works demonstrating that the latent representation in the AE bottleneck can distinguish between normal and anomalous data [363] and has been used in other DL-based methods [361, 362].

Several generative DL-based approaches for AD use the reconstruction error as an anomaly score (see Section 7.1.2). As mentioned in Section 7.3, generative AE-based models help to obtain a latent representation that describes the data manifold, encoding the most relevant features of the data. This makes it intuitive to understand why these models should provide reconstructions where anomalies are depicted. This is particularly relevant for obtaining a pixel-wise anomaly map, as proposed by Wolleb *et al.* [371], who applied a DPM for pixel-wise AD in medical imaging.

In the proposed ADeLEn, a parametric statistical approach for AD is used based on the latent representation Z, see Eq. (7.7). To obtain a tractable approximation of the real distribution, variational inference via VAE is employed. The proposed ADeLEn is a semi-supervised approach, with the regularization factor conditioned by the labeled dataset, as depicted in Eq. (7.11). As a result, this variational approach allows obtaining an approximation of the conditional probability noted as $p_\phi(z|x)$, where ϕ represents the variational parameters.

However, the anomaly score in ADeLEn does not correspond to this $p_\phi(z|x)$. Instead, the proposed anomaly score is $\mathcal{H}(Z|x)$ (see Eq. (7.12)), which represents the model's uncertainty. The intuition behind this approach is that the model understands one type of data well, the 'normal' data, but disturbances in this data generate uncertainty. Using the MNIST dataset as an example, Fig. 7.7a shows that normal samples (1's) have low uncertainty. However, when the model sees a pattern similar to a 7, the uncertainty increases, considering it an abnormal (anomalous) 1. Recall that the entropy of a Gaussian distribution is proportional to σ as depicted in Eq. (7.15).

In general, the proposed ADeLEn can be considered an extension of the Deep SAD [362]. Although Deep SAD is a deterministic approach, the score for AD can be interpreted as the conditional probability $p(z|x)$. Both semi-supervised approaches can be depicted, from a IT perspective, by Eq. 7.18. However, because ADeLEn aims to obtain the probability distribution of the latent representation, it can quantify uncertainty, which Deep SAD cannot. Thus, Deep SAD uses the distance to the centroid as the anomaly score.

Although more intensive experimental are needed to evaluate ADeLEn in real medical cases, Section 7.4 provides different experiments that offer insight into the ADeLEn mechanism. These experiments demonstrate that ADeLEn can handle problems where classic AD methods, specifically the OC-SVM, fail and provide benefits over a naive DNN for classification in this case where the datasets are not balanced.

The study of the pollution ration in the dataset, i.e., the unidentified anomalies within a dataset assumed to consist solely of normal observations, offers good insight into the problem-solving mechanism in ADeLEn. Figure 7.6 or Fig. 7.8 provide illustrative examples of the explanation described in Section 7.3, where data compression is constrained by the maximization of the KL Divergence between distributions, see Eq. (7.18). Additionally, the impact of certain hyperparameters of the ADeLEn model, such as the number of labeled anomalies and the dimension of the bottleneck, has been explored.

Overall, the results presented in Section 7.4 and the IT perspective discussed in Section 7.3 provide an interpretation of the proposed ADeLEn and its benefits. However, future work should include a more exhaustive comparison with other methods.

7.6 Conclusions

Nowadays, AD has received significant research interest in the healthcare scenarios based on the interpretation that pathologies are unusual patient health conditions, i.e., anomalies. One key advantage of these methods is that they do not require extremely comprehensive datasets. In the healthcare sector, it is often easier to obtain data from healthy subjects, although this can depend on the data modality. State-of-the-art AD methods frequently do not require labeled data. However, these methods face limitations, specially in high-dimensional data. For this reason, the use of AD methods based on DL has gained attention.

In this chapter, it has been proposed a DL-based approach for AD, named ADeLEn. This method distinguishes itself from conventional DL-based approaches by leveraging the latent space Z for anomaly scoring, based on the conditional probability. The idea is to apply a parametric statistical approach in Z, assuming that the distributions of normal and anomalous samples follow the same form but with different parameters [344]. Specifically, it is assumed that the compressed representation Z corresponds to a Gaussian distribution for both normal samples and anomalies, with the distribution differing in their standard deviation, σ. The proposed anomaly score corresponds to $\mathcal{H}(Z|x)$, representing the model's uncertainty that, which, given the Gaussian assumption, is proportional to σ. It is worth noting that this AD method has been designed with medical imaging challenges in mind, where samples are similar among samples and anomalies exhibit specific patterns, such as a tumor in a human brain.

This thesis aims to enhance the interpretability of the solving mechanisms employed in DNN models. The proposed ADeLEn is well-defined compared to most state-of-the-art methods based on AEs. Following the main approach of this thesis, a IT perspective has been discussed, providing a description of the optimization problem based on the InfoMax principle. Using this IT perspective, an optimal initialization for ADeLEn has been proposed, as depicted in Section 7.3.

Although more extensive testing is needed, the proposed method could be considered an optimal approach for the case studies described in this chapter. However, a significant disadvantage of this approach is the necessity for a subset of previously labeled anomaly samples. The experiments considered unbalanced datasets, assuming that the dataset consists solely of N normal observations, with the number of anomalies M being significantly smaller ($N \gg M$).

This proposed ADeLEn is easier to interpret than other methods based exclusively on AEs and using the reconstruction error as a score. Additionally, it provides a more robust score than another semi-supervised approach, such as the Deep SAD. As a result, it can be considered more suitable for healthcare scenarios than other methods proposed in the state-of-the-art.

The semi-supervised nature of ADeLEn could be considered a disadvantage compared to unsupervised DL-based methods. However, this need for labeled data arises from the necessity of obtaining a discriminative representation between both distributions, which is challenging to achieve through unsupervised learning, as discussed in Section 2.2.5. Future work could explore the use of contrastive learning to achieve this discriminative representation or an iterative training approach, similar to those proposed in [354] or [358], where the original non-labeled dataset is iteratively separated into two subsets (normal and anomalous samples) during training.

In conclusion, this chapter has introduced a DL-based method for AD that offers better interpretability compared to other state-of-the-art methods. However, as emphasized throughout this thesis, this method has limitations and is not considered an absolute solution. The proposed ADeLEn cannot replace all existing methods; rather, it is essential to understand its advantages in certain settings and limitations in others. Recall, as introduced at the beginning of this chapter, that anomalies are not solely based on the data structure but are also influenced by the AD methodology.

Chapter 8

Conclusions and Future Lines

> In order to determine whether we can know anything with certainty, we first have to doubt everything we know.
>
> *René Descartes*

Nowadays, we are witnessing a new technological revolution driven by the rapid evolution of AI, powered by DL models. This is particularly evident in the rise of generative AI, which, beyond entertainment, is becoming increasingly relevant in enhancing human productivity. AI applications have the potential to significantly boost productivity by assisting humans in performing tasks, essentially acting as a 'copilot' in various activities. From my personal experience, one of the most impactful changes in my work has been using an AI assistant for software development. This has significantly reduced my development time, for instance, by avoiding the frequent need to consult library documentation to recall specific usage details. While it is not infallible, overall, it enables me to prototype more efficiently in a shorter time frame. As a human being, I have the ultimate responsibility for the acceptance or rejection of the AI's suggestions.

As it was mentioned in Section 1.2, there are still challenges to overcome before fully integrating the current DL-powered AI into our daily lives. Consequently, an increasing amount of research is focused on examining perceptions of AI tools, their applications, and their impact on key societal areas such as security [381] and education [382]. One critical issue in this integration is the lack of intuition regarding the mechanisms behind DL models. These models are so complex that it is almost impossible to explain how they work. In other words, in DL models, understanding how a prediction is made by simply analyzing the model's parameters is unfeasible, which is why these models are often referred to as 'black boxes'.

This problem of not being able to explain what is happening inside the 'black-box' is particularly significant in critical areas such as medical screening or autonomous driving. To address this issue, enhancing the interpretability of these models is crucial. However, as noted in Section 2.3.1, interpretability should not be considered absolute, as interpretations may vary depending on the context, the perspective of the observer, and the objectivity of the analysis.

This dissertation aims to propose DL models to solve various problems in the clinical field, particularly focusing on image processing. However, throughout this dissertation, special emphasis has been placed on interpreting the problem-solving mechanisms used in the proposed models. I recognize that the engineering approach to this implies that the interpretations presented might not please everyone. Nonetheless, this interpretation is objective and quantifiable, which provides a strong basis for validating the models in any context.

Different tasks have been addressed in this dissertation, and although the approaches may vary in practice, the underlying mechanisms of these models share key elements formalized through the IT-based interpretation used throughout this research. This chapter summarizes the main conclusions and contributions of this research and outlines future research directions that I consider should be addressed in the near future.

8.1 Practical Conclusions

In Chapter 4, the IT perspective introduced by Tishby and Zaslavsky [97] was explored and evaluated within a clinical context, focusing on the common challenge in medical problems known as the *data challenge*, which is characterized by insufficient labeled data [180]. This aims to provide a methodology based on the analysis of the IP estimation using a scarce dataset. For the MI estimation for obtaining the IP, it was applied the kernel-based entropy estimator described in Section 3.6.

As a case of study, it was analyzed in a classification task where the purpose is to obtain early-stage detection of DFU using thermograms, where the availability of labeled data is limited. The experiments aimed to identify patterns consistent with the state-of-the-art, grounded in the IT perspective, using the methodology proposed. As a result, it is demonstrated that, while the classification metrics alone did not reveal models likely to underperform, the IP analysis identified that models failed to satisfy the DPI, a critical requirement in IT-based interpretation.

Additionally, given the remarkable similarity of AEs to transmission channels, the effect of the dropout technique on DNNs was also evaluated using an AE. The results revealed that the dropout technique aligns with distortion rate theory for achieving optimal compressed representations, providing a new interpretation of this technique. Consequently, the dropout rate directly impacts the compression of input data. By formalizing supervised DL models as a trade-off between data compression and prediction through the IB principle, it was demonstrated that increasing the dropout rate emphasizes data compression within the model.

In Chapter 5, a FS method based on DL models was proposed. This method aims to select a subset of features from the input that contains the necessary information for a given task. In clinical scenarios, this would reduce the number of features, reduce analysis costs, and improve the interpretability of data, ultimately reducing variability in clinical decision-making based on high-dimensional data. The method was evaluated using a case study focused on identifying relevant wavelengths for differentiating biological tissues via HSI. The results were consistent with findings in the literature.

This subset of features can be viewed as a compressed representation of the data, a task at which DL models excel, as demonstrated in Chapter 4 using the IT perspective. The proposed architecture is based on the dropout technique, which emphasizes data compression as concluded in Chapter 4, where the dropout rate is a trainable parameter using a variational approach.

The FS method was framed from an IT perspective to offer more in-depth insight into the problem-solving mechanism and demonstrate the utility of this perspective for designing optimal architectures. Using this approach, it was deduced that classification and feature selection, in the proposed method, can be treated as independent problems, leading to the proposal of a scheduler for the regularization factor for data sparsity. Additionally, the optimal initialization of the feature selector layer, based on dropout, was highlighted. Finally, using the IB principle, the subset of features was described as the model's bottleneck, Z. This is particularly relevant because it suggests that the model may not generalize effectively, with the trade-off between data compression and prediction occurring at the first layer. An intuitive explanation about how the generalization is obtained in the first layers is provided in Section 4.1.1.

Chapter 6 focuses on another task related to data interpretability, i.e, signal decomposition. Specifically, this chapter is dedicated to the HSI decomposition, commonly known as HyperSpectral Unmixing, given that HSI is characterized as high-dimensional data. This chapter presents a novel approach for HSU using a DL model, the CLHU framework. Due to the limited availability of HSI data for medical applications, the proposed method was validated using remote sensing images typically used to evaluate unmixing algorithms. However, a notable case is presented in Section 6.4.4, where biological tissue decomposition was attempted with promising results, demonstrating potential applications such as identifying and quantifying substances within tissue samples.

The IT was instrumental in designing the CLHU method, providing insight into its functionality. Reformulating the method using the IB principle revealed its similarities with contrastive-learning methods, a technique that originally inspired this work. As a result, it was concluded that the proposed architecture could be extended to tasks beyond HSU, such as classification with uncertainty prediction, where the trainable descriptor M could describe the classes, and the similarity function could quantify model uncertainty. Additionally, it was concluded that the proposed model must satisfy a specific DPI, depicted in Eq.(6.11). This DPI was validated using the IP.

Finally, **Chapter 7** addresses a crucial task in dataset analysis: Anomaly Detection (AD). A DL-powered AD method, ADeLEn, was proposed, specifically designed for medical imaging challenges. The method adopts a semi-supervised approach, considering the 'data challenge' commonly encountered in medical contexts, and was evaluated with heavily imbalanced datasets. The experimental setup considered the number of anomalous samples and dataset "pollution", referring to unidentified anomalous samples used during training. In the experimental results, the proposed method was compared to the state-of-the-art OC-SVM and a supervised DNN. Results revealed that the proposed method is more versatile than the OC-SVM and does not suffer from overfitting, as seen in the supervised DNN approach using imbalanced datasets.

The proposed method adopts a generative approach using a VAE. Section 7.3 highlights the key features of this approach compared to models presented in previous chapters from the IT perspective. The variational approach in the generative model applies a strong restriction on the information in the bottleneck Z, establishing an upper limit defined by the variational parameters ϕ, ensuring that the real data distribution in the bottleneck contains no more information than the variational distribution. Additionally, the problem was reformulated using the InfoMax principle, where the bottleneck is designed to increase the divergence between the distributions of normal and anomalous samples, proposing an optimal initialization of the variational parameters for the ADeLEn method.

It was concluded that ADeLEn is easier to interpret than other AE-based methods, and the score based on the bottleneck is more robust than using reconstruction error as a score proposed in most AE-based methods. Moreover, it was observed that the proposed method generalizes the

Deep SAD, and the upper limit constraint in ADeLEn makes the score threshold more robust and less dependent on data, making it more suitable for general-purpose healthcare applications.

8.2 Dissertation Results

This section highlights the scientific contributions and achievements obtained during this Ph.D. thesis. These include academic publications, participation in research projects, scholarships, and research contracts. Together, they demonstrate the impact and scope of the work conducted, as well as the collaborations and recognition within the academic and professional communities.

List of Publications

The following acronyms are used to highlight the specific characteristics of each publication: **OA:** *Open Access*, **OC:** *Open Code*, **IF:** *Impact Factor*, and **IC:** *International Collaboration.*

Journal Publications

- Takamatsu, T., Fukushima, R., Sato, K., Umezawa, M., Yokota, H., Soga, K., **Hernandez-Guedes, A.,**, ... & Takemura, H. (2024). Development of a visible to 1600 nm hyperspectral imaging rigid-scope system using supercontinuum light and an acousto-optic tunable filter. Optics Express, 32(9), 16090-16102. **[IF: 3.2 — Q2, IC, OA]**

- Arteaga-Marrero, N., **Hernandez-Guedes, A.**, Ortega-Rodríguez, J., & Ruiz-Alzola, J. (2023). State-of-the-art features for early-stage detection of diabetic foot ulcers based on thermograms. Biomedicines, 11(12), 3209. **[Related with Chapter 5] [IF: 3.9 — Q2, OA, OC]**

- **Hernandez-Guedes, A.**, Arteaga-Marrero, N., Villa, E., Callico, G. M., & Ruiz-Alzola, J. (2023). Feature ranking by variational dropout for classification using thermograms from diabetic foot ulcers. Sensors, 23(2), 757. **[Related with Chapter 5] [IF: 3.4 — Q2][OA, OC]**

- **Hernandez-Guedes, A.**, Santana-Perez, I., Arteaga-Marrero, N., Fabelo, H., Callico, G. M., & Ruiz-Alzola, J. (2022). Performance evaluation of deep learning models for image classification over small datasets: Diabetic foot case study. IEEE Access, 10, 124373-124386. **[Related with Chapter 4] [IF: 3.4 — Q2, OA, OC]**

- Arteaga-Marrero, N., Bodson, L. C., **Hernandez-Guedes, A.**, Villa, E., & Ruiz-Alzola, J. (2021). Morphological foot model for temperature pattern analysis proposed for diabetic foot disorders. Applied Sciences, 11(16), 7396. **[IF: 2.8 — Q2, OA]**

- Arteaga-Marrero, N., **Hernandez-Guedes, A.**, Villa, E., Gonzalez-Perez, S., Luque, C., & Ruiz-Alzola, J. (2021). Segmentation approaches for diabetic foot disorders. Sensors, 21(3), 934. **[IF: 3.4 — Q2, OA, OC]**

In Peer-reviewed Articles (Preprints)

- **Hernandez-Guedes**, A., Fukushima, R., Takamatsu, T., Fabelo, H., Ruiz-Alzola, J., Takemura, H., & Callico, G. M. (2023). Contrastive Learning approach for blind Hyperspectral Unmixing (CLHU). Authorea Preprints. ***Under Review: IEEE Access.*** [**Related with Chapter 6**] [**IC, OA, OC**]

International Conference

- Fukushima, R., Takamatsu, T., Sato, K., **Hernandez-Guedes, A.**, Callico, G. M., Okubo, K., ... & Takemura, H. (2024, January). Detection of Exposed Nerves in Two Individuals In Vivo and Unexposed Nerves Ex Vivo with Near-Infrared Hyperspectral Laparoscope. In 2024 IEEE/SICE International Symposium on System Integration (SII) (pp. 19-24). IEEE. [**IC**]

- Castro-Fernández, M., **Hernandez-Guedes, A.**, Fabelo, H., Balea-Fernández, F. J., Ortega, S., & Callicó, G. M. (2022, August). Towards skin cancer self-monitoring through an optimized MobileNet with coordinate attention. In 2022 25th Euromicro Conference on Digital System Design (DSD) (pp. 607-614). IEEE.

- **Hernandez-Guedes, A.**, Arteaga-Marrero, N., Villa, E., Fabelo Gómez, H. A., Marrero Callicó, G. I., & Ruiz Alzola, J. B. (2019). Automatic segmentation based on deep learning techniques for diabetic foot monitoring through multimodal images. Lecture Notes in Computer Science.

Book Chapter

- León Martín, S. R., **Hernandez-Guedes, A.**, Fabelo Gómez, H. A., Ortega Sarmiento, S., Balea Fernandez, F. J., & Marrero Callicó, G. I. (2021). SWIR Hyperspectral Imaging to Assess Neurocognitive Disorders Using Blood Plasma Samples. [**IC**]

Participation in Research Projects

European Projects

- **Ref. 101017385:** WARIFA (Artificial intelligence and the prevention of chronic conditions – GA: 101017385). European Union's Horizon 2020 research and innovation programme. PI: Gustavo M. Callico and Ana Wägner. ULPGC. 01/01/2021 – 31/12/2024. 6,726,468.75 €. Task: HSI processing.

- **Ref. MAC/1.1b/098:** MACbioIDi (Contribuyendo a la cohesión e internacionalización de la Macaronesia para impulsar los Objetivos del Desarrollo Sostenible con las TICs y la I+D+i biomédica). Interreg MAC 2014-2020. PI: Juan Ruiz Alzola. ULPGC. 01/01/2017 - 31/12/2019. 2,354,206.58 €. Task: Thermal imaging processing and ML researcher.

National/Regional Projects

- **Ref. PID2020-116417RB-C42:** TALENT (HypErsPEctRal Imaging for Artificial intelligence applications). Spanish Government and European Union (FEDER funds). PI: Gustavo

M. Callico and Sebastian Lopez. ULPGC. 01/09/2021 - 01/09/2024 175,813.00 €. Task: HSI processing.

- **Ref. ProID2017010164:** ITHACA (IdenTificacion Hiperespectral de tumores CerebrAles). Gobierno de Canarias (Canary Islands) Programa de Apoyo a la Investigación María del Carmen Betancourt y Molina. PI: Gustavo M. Callico. ULPGC. 01/01/2018-30/09/2019. 69,914.45 €. Task:

Grants Obtained

- **Pre-doctoral grant** given by the "Agencia Canaria de Investigacion, Innovacion y Sociedad de la Información (ACIISI)" of the "Consejería de Economía, Conocimiento y Empleo" of the "Gobierno de Canarias", which is part-financed by the European Social Fund (FSE) (POC 2014-2020, Eje 3 Tema Prioritario 74 (85%)).

- **Research intership movility grant** given by the "Agencia Canaria de Investigacion, Innovacion y Sociedad de la Información (ACIISI)" of the "Consejería de Economía, Conocimiento y Empleo" of the "Gobierno de Canarias", which is part-financed by the European Social Fund (FSE) (POC 2014-2020, Eje 3 Tema Prioritario 74 (85%)).

Research Contracts

- Researcher in the MACbioIDi project (October 2017 – March 2020).

8.3 General Conclusions

Throughout this dissertation, various DL architectures have been proposed and evaluated in relation to the specific tasks. However, this work does not aim to surpass state-of-the-art results. Instead, the primary focus is on developing more interpretable solutions using DL. For comparison purposes, the proposed methods have been assessed using several state-of-the-art algorithms to ensure that the models yield competitive results. Additionally, key concepts have been addressed in a broader context, aiming to provide a general intuition about the problems at hand and the mechanisms required to solve them.

This dissertation aims to emphasize the importance of striving for more interpretable models. It is not a criticism of using a 'black box' to solve problems. There are critical scenarios where the complexity of the problems makes it necessary to use a 'black box' because some problems are unapproachable by other methods. Nevertheless, the success of this work lies in shifting the focus away from the assumption that accurate predictions can only be achieved by increasing the number of parameters in the 'universal function', the DNN, and towards exploring alternative architectures that balance accuracy with interpretability.

As highlighted in Section 2.3, interpretability is a domain-specific concept, and interpretations that work in one context may not extend to another. However, throughout this dissertation, IT is presented as a foundational framework for achieving more general-purpose interpretability that can be used in any context. By conceptualizing models as systems that process and transmit information, we can shift our focus away from the detailed structure of the data or feature explanations at various layers. Instead, the primary concern becomes the amount of information captured and transmitted through the model's feature representations.

Information is a key point and, in Chapter 1, it was dedicated a section for describing the importance of this and how it plays a central role in almost every scientific discipline and how the skepticism about the information is a consequence of confusing 'meaning' with 'information'. Shannon points out that 'meaning' was "*irrelevant to the engineering aspects*" [10], as it introduces subjectivity and human psychology into the equation.

The IT perspective used in this work is applicable to any context, as it abstracts away from the specific meaning of the data. However, it can be further extended by using the post-hoc human interpretable explanations used in 'explainable ML' (see Section 2.3.3), where the objective is to describe how DL behaved, and why. While these post-hoc interpretations are subjective and cannot be universally applied, the IT-based interpretability framework used in this thesis must always be satisfied. As Warren Weaver noted [9]

> *It is this, undoubtedly, that Shannon means when he says, "the semantic aspects of communication are irrelevant to the engineering aspects". But this does not mean that the engineering aspects are necessarily irrelevant to the semantic aspects.*

A good example of this can be found in Chapter 5, where the results obtained about the FS for band selection in HSI are presented. The validation of the results presented in Section 5.4.3 were partially validated because of the selection of specific wavelength that it is considered in the state-of-the-art such as the selection of a range around 1200 nm for the identification of fat and muscle tissue because of the absence of lipids in muscle tissue where the absorption coefficients of water (abundant in muscle) and lipids are comparable [270]. The interpretation obtained from the data here is context-dependent, and cannot be extrapolated to another context, but the IT validation used in Section 5.5 can be used in any context.

This post-hoc human interpretable explanation can be used in this specific case because the FS method has a mechanism used for obtaining a better interpretation about the behavior of the model, but this explanation would be invalid if the IT perspective is not satisfied in the model. For this reason, the IT interpretability used in this dissertation should be considered not only for the design of DL architectures, but also for the validation of the model.

As introduced in Section 1.1, the reason that DL models outperform other ML methods is that they operate at multiple levels of abstraction, creating representations that are appropriate to the problem being addressed and acting as 'information processors'. A DNN does not process images or signals directly; instead, it processes the information contained in the data. Consequently, the information from the source must be effectively propagated across the different levels of abstraction, reflecting the behavior discussed throughout this dissertation. This is a fundamental characteristic necessary for describing intelligence.

Therefore, satisfying this IT interpretation is a necessary, though not sufficient, requirement for defining intelligence. Similar to how the human brain works, these models aim to reduce the uncertainty in the decision-making process — specifically, by minimizing the conditional entropy $\mathcal{H}(Y|X)$. In humans, the information from sensory organs is transmitted to specific parts of the brain responsible for decision-making, with information passing through various levels of abstraction. The same principle applies to DL models, which propagate information across layers of abstraction to support decision-making.

Throughout this dissertation, we have explored how well DL models transmit information. Just as an intelligent system must effectively process and transmit information, I trust that the information presented in this document has been communicated accurately and clearly.

8.4 Future Work

The methodologies and findings presented in this thesis offer multiple avenues for future exploration and application. These ideas are inherently adaptable and can be extended to various contexts, paving the way for innovative solutions in different domains. Furthermore, the key concepts introduced throughout the chapters have the potential to be combined, yielding more refined and effective results. For instance, both Chapter 5 and Chapter 6 focus on HSI data, and a natural progression would be the integration of the approaches developed in these chapters. An alternative direction involves utilizing the proposed DL-based AD method in Chapter 7 to eliminate outliers in the dataset for band selection used in Chapter 5, replacing the IQR approach. Moreover, the applicability of these methods could be expanded to other fields beyond the scope of this work.

Nevertheless, several conceptual areas remain unexplored within this dissertation, offering significant opportunities for further research. The following directions are particularly worthy for further exploration and have the potential to drive meaningful advancements within the scientific community:

- **Proposed FS method as a quantization technique**: The dropout technique, traditionally considered as a regularization method, can be reinterpreted through the lens of distortion-rate theory as a compression mechanism. This perspective opens the door to explore dropout as a quantization method aimed at reducing the size and complexity of DL models.

- **Uncertainty estimation using the proposed CLHU framework**: From an IT perspective, the concept of endmembers, grounded in the IB principle, can be extended beyond their original purpose as class descriptors. The similarity function used for abundance estimation could serve as a valuable tool for quantifying model uncertainty, providing a deeper insight into the confidence of predictions.

- **Enhancing robustness in AD by normalizing flows in ADeLEn**: While the proposed ADeLEn method demonstrates significant potential, especially in handling imbalanced datasets, further improvements can be achieved by incorporating normalizing flows. This enhancement could be particularly beneficial for high-resolution image datasets, offering a more robust framework for AD by leveraging the expressive power of normalizing flows.

This dissertation has explored a range of learning paradigms (see Fig. 2.1), including supervised, unsupervised, and semi-supervised learning; however, reinforcement learning remains an area for future investigation. One of the guiding hypotheses of this dissertation was to establish a common framework for exploring different tasks, a goal achieved through the consistent application of IT as a lens for exploring the interpretability of DL models and understanding problem-solving mechanisms. A valuable extension of this work would be to design a reinforcement learning algorithm inspired by the IT perspective, using IT principles to enhance interpretability and provide a more comprehensive understanding of the model's learning processes.

Part III

Appendix

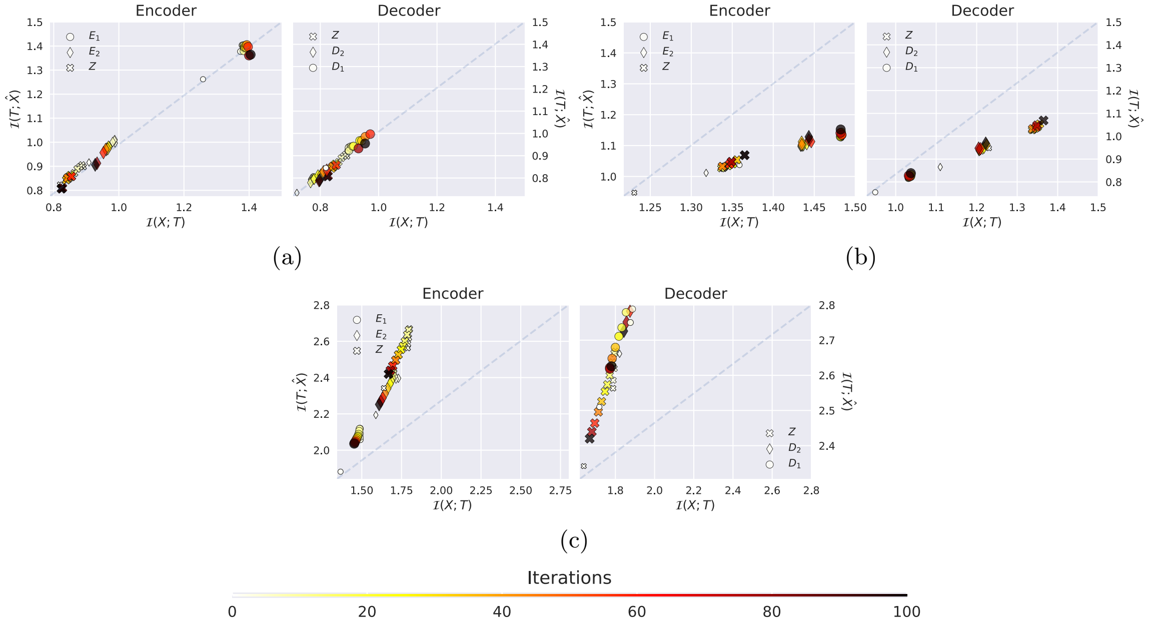
Appendix A

Supplementary Material

A.1 Information-theoretical Evaluation in Deep Learning

A.1.1 Information Plane during DFU fitting phase

In Section 4.2.2, we outlined the two-step process of using the AE: first, pretraining the AE with the FMNIST dataset, and then fine-tuning this pretrained model with the DFU dataset (detailed in Section 4.2.1). This section presents the IP estimation during the second stage, where we train the AE, which was initially pretrained with FMNIST, using the DFU dataset.



(a)

(b)

(c)

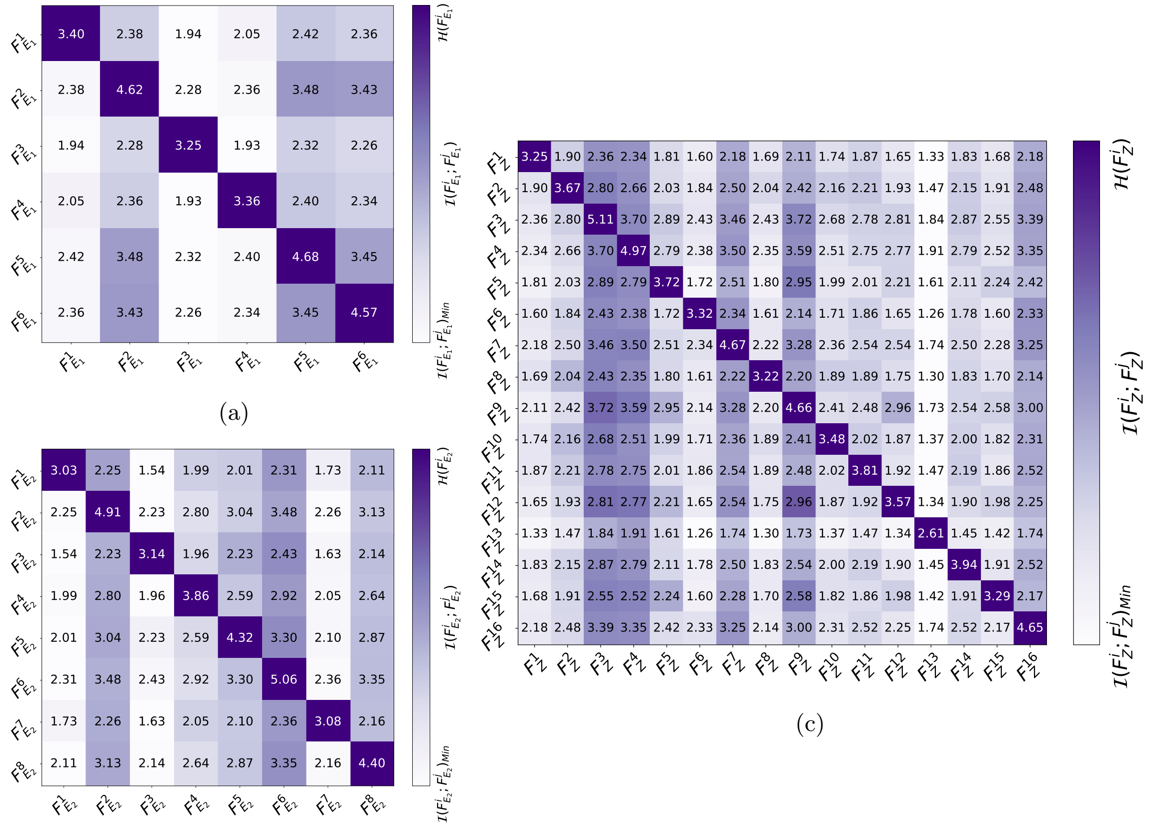
Figure A.1: IP estimation using dropout technique in the AE among different dropout rates (ρ_d) fitting with DFU dataset: (a) $\rho_d = 0.5$, (b) $\rho_d = 0.2$ and (c) No dropout ($\rho_d = 0$).

Figure A.1 reveals a relatively stable compression pattern in the IP estimation during the fitting phase on the DFU dataset across most cases. Notably, when a dropout rate of $\rho_d = 0.2$ is applied, the model appears to still be in a fitting phase, as there is no observable compression in the IP estimation (Fig. A.1b). In the third case, there is no dropout (Fig. A.1c), there is an evident DPI violation, so it is concluded that this model has an evident overfitting although the MI estimation in the latent space ($\mathcal{I}(X;Z)$) is higher than in the other cases. Apparently, this overfitting is clearly evident because of the use of batch normalization in the convolutional layers.

However, when these results are compared to the observations in Fig. 4.12, it becomes evident that the learned representation may not be as effective. This implies that the dropout rate has a significant impact on the learning process and the quality of the acquired representation. In summary, Fig. A.1 essentially confirms the conclusions outlined in Section 4.3.2.

A.1.2 Encode path without dropout technique

This section illustrates the example of Section 4.3.2 where the dropout technique is not applied. As was concluded in that section, this AE exhibits a behavior that indicates it suffers from overfitting. This outcome can be attributed to the use of batch normalization, a technique that standardizes layer inputs within each mini-batch, which stabilizes the learning process and significantly reduces the number of training epochs needed for training DNN.



(a)

(c)

(b)

Figure A.2: MI estimation between different filters in the encode path of the AE with a dropout rate of 0.5 at the end of each layer, denoted as F_k^i where i represents the index of the filter and k the encoder layer. The encode path is composed by (a) E_1, (b) E_2 and (c) Z.

Nevertheless, this AE exhibits some characteristics that were previously described in Section 4.3.2. Figure A.2 presents the MI estimation among the different filters in the various layers of the decode path: E_1, E_2, and the bottleneck Z. In Fig. A.2c, it is apparent that the amount of information in these layers is higher compared to the other AEs where dropout techniques were employed. This is evident through the entropy estimation of each layer, which is higher due to the absence of the dropout technique.

A.2 Feature Selection with Deep Learning: Hyperspectral Band Selection

A.2.1 Mutual Information

Following the proposed method for FS in Section 5.2.1, the optimization problem based on IT-learning is indicated in Section 5.3.1. From this perspective, it involves maximizing the MI, as it is depicted in Eq. (5.9). In this section, it will be described the $\mathcal{I}(X_s;Y)$ expressed in the equation.

The optimization problem considers three variables: the input X, the output Y, and the subset indices S. As the S only affects to the input X, Eq. (5.2), this operation is represented by $p(X,S)$. The goal is to maximize the MI between these variables along a Markov chain $(X \to S \to Y)$, i.e., $\mathcal{I}(X \odot S;Y) = \mathcal{I}(Y;X,S)$. It can be expressed as follows:

$$\begin{aligned}\mathcal{I}(Y;X,S) &= \sum_{X,Y,S} p(x,y,s)\log\frac{p(x,y,s)}{p(y)p(x,s)} = \\ &= \sum_{X,Y,S} p(x,y,s)\log\frac{p(y|x,s)p(x,s)}{p(y)p(x,s)}.\end{aligned} \tag{A.1}$$

However, it is imported to note that S follows a Bernoulli distribution, and $p(x,s)$ can be expressed as the product of independent variables X and S. Thus, the joint probability is defined as $p(x,s) = p(x)p(s)$. Therefore, the Eq. (A.1) is modified as follows:

$$\begin{aligned}\mathcal{I}(Y;X,S) &= \sum_{X,Y,S} p(y|x)p(x)p(s)\log\frac{p(y|x)p(x)p(s)}{p(x)p(y)p(s)} = \\ &= \sum_{S} p(s)\sum_{X,Y} p(x,y)\log\frac{p(x,y)}{p(x)p(y)},\end{aligned} \tag{A.2}$$

where the second expression corresponds to $\mathcal{I}(X;Y)$, so it may be expressed as:

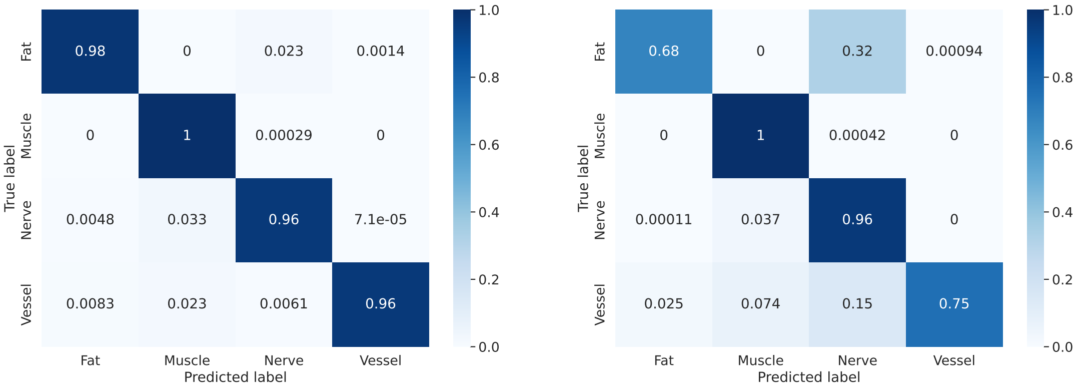
$$\mathcal{I}(Y;X,S) = \sum_{S} p(s)\,\mathcal{I}(X;Y) = E_S[\mathcal{I}(X;Y)]. \tag{A.3}$$

As a result, the MI estimation can be expressed as the expected value over S of the $\mathcal{I}(X;Y)$, as it was indicated in Eq. (5.10).

A.2.2 Conditional Entropy

As it was indicated in Section 5.3.1, the objective consists on maximizing the MI between X_s and Y, $\mathcal{I}(X_s;Y)$. Considering that Eq. (5.2), $p(X_s) = p(X,S)$ and the MI is estimated as:

$$\mathcal{I}(Y;X,S) = \mathcal{H}(X,S) + \mathcal{H}(X,S|Y), \tag{A.4}$$



(a)

(b)

Figure A.3: Quantitative evaluation based on SVM for the downsampling methods. Figure A.3a represents the dimensional reduction of the subset of samples selected by the method based on K-Means and Fig. A.3b represents the case where OCSP was applied.

where X is the input, Y corresponds to the output, and S is the subset of indices. In this way, the maximization of the MI corresponds to a minimization of the conditional entropy, $\mathcal{H}(X, S|Y)$.

Given the Eq. (3.6), the conditional entropy may be expressed as:

$$\mathcal{H}(X, S|Y) = \mathcal{H}(X, Y, S) - \mathcal{H}(Y), \tag{A.5}$$

and, as the $\mathcal{H}$(Y) is fixed, the minimization of the conditional entropy might be expressed as a minimization of the joint entropy $\mathcal{H}(X, Y, S)$.

Given that S, is defined by an independent Bernoulli distribution $p(X, S) = p(X)p(S)$ and there is dependent between X and Y, $p(X, Y) = p(Y|X)p(X)$, the joint probability $p(X, Y, S) = p(Y|X, S)p(X, S) = p(Y|X)p(X)p(S)$. In this way, the joint entropy is defined as follows:

$$\begin{aligned}\mathcal{H}(X, Y, S) = \sum_{x,y,s} p(y|x)p(x)p(s) \, \log \left(p(y|x)p(x)p(s)\right) = \\ = \sum_{x,y,s} p(x, y)p(s) \, \log \left(p(x, y)p(s)\right).\end{aligned} \tag{A.6}$$

As a result, this expression might be simplified as:

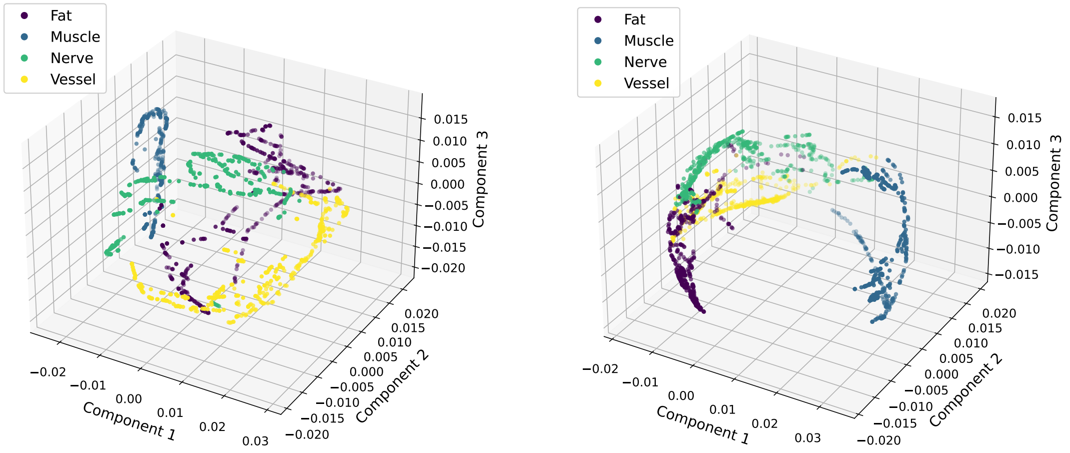
$$\begin{aligned}\mathcal{H}(X, Y, S) = \sum_{X,Y} p(x, y) \log p(y|x) + \sum_{S} p(s) \log p(s) = \\ = \mathcal{H}(X, Y) + \mathcal{H}(S).\end{aligned} \tag{A.7}$$

A.2.3 Laparoscopy Dataset Downsampling

As it was indicated in Section 5.4.2, the downsampling of the dataset was evaluated by a quantitative evaluation and a qualitative result based on PHATE. The results obtained indicated that the most interesting downsampling techniques for this purpose were the one based on K-Means [260] and the another based on OCSP [266]. Figures A.3 and A.4 represent the quantitative and qualitative evaluation, respectively.

In terms of quantitative evaluation, a cross-evaluation was conducted using 10 folds. Simultaneously, the SVM used for evaluation was optimized, obtaining specific hyperparameter tuning for each method. For the K-Means approach, the SVM uses an RBF kernel with γ set to 0.0811, and the parameter C to 1.52. The SVM for the OCSP method used an RBF kernel with γ set to 0.0811, and the parameter C to 1. In both methods, the kernel and γ, which controls the spread of the Gaussian center, are the same. However, the hyperparameter C, which controls the trade-off between decision boundary and misclassification, is different.

Considering Fig. A.3, it is observed that the method based on K-Means has almost a perfect score in cross-validation (see Fig. A.3a). For this reason, it can be deduced that there is overfitting associated with this model. On the other hand, the method based on OCSP shows a clear deficiency in identifying fat tissue, confusing it with nerve tissue, as it is illustrated in Fig. A.3b. However, it can be deduced that the subset of samples obtained by OCSP has more variability compared to K-Means, and for this reason, this subset might be more representative than the one obtained by K-Means.



(a)

(b)

Figure A.4: Qualitative evaluation based on PHATE for the downsampling methods. Figure A.4a represents the dimensional reduction of the subset of samples selected by the method based on K-Means and Fig. A.4b represents the case where OCSP was applied.

In terms of qualitative evaluation, a clustering pattern is observed in the K-Means approach, especially noticeable in local structures (see Fig. A.4a). This pattern is a consequence of obtaining samples that are closer to different centroids, as proposed in [260]. On the other hand, OCSP shows a global pattern where samples are well dispersed, but in the local structure, it exhibits more continuity than the K-Means-based method.

Both evaluations lead to the same conclusions. The clustering pattern in K-Means, observed by PHATE, might be the reason for the nearly perfect cross-validation results in the quantitative evaluation. This clustering pattern reduces the variability observed in the subset. Consequently, the OCSP method for downsampling is selected based on these observations.

A.3 Signal Decomposition: HyperSpectral Unmixing

A.3.1 Multiple channel contribution

The CLHU model proposed in Chapter 6, has the injection of two different variables in the bottleneck, Z and M for obtaining A, as it is observed in Fig. 6.3. In Section 6.3.1, it was described as there are two channels for providing information of A. In this way, it is interesting to obtain the amount of information provided by the different channels.

Considering $p(a_i|m_j)$, where $a_i \in A$ and $m_j \in M$, as the posteriori distribution given m_j and $p(a_i|m_j, z_k)$, with $z_k \in Z$, as the posteriori distribution considering both m_j and z_k, the posteriori entropy can be defined as:

$$\mathcal{H}(A|m_j, z_k) = \sum_{a \in A} p(a|m_j, z_k) \log \frac{1}{p(a|m_j, z_k)}. \tag{A.8}$$

In this way, the uncertainty of A with respect M and Z is defined as follows:

$$\mathcal{H}(A|M, Z) = \sum_{m \in M} \sum_{z \in Z} p(m, z) \mathcal{H}(A|m, z). \tag{A.9}$$

Given the Eq. (A.9) and following Eq. (3.10), $\mathcal{I}(A; M, Z) = \mathcal{H}(A) - \mathcal{H}(A|M, Z)$ provide the amount of MI among the different variables. Remember, this MI formulation can be decomposed by the additivity of the information provided for the different channels, AZ and AM, as it has been called in Section 6.3.1. Considering that $\mathcal{I}(A; M)$ is the amount of information in the channel AM, the $\mathcal{I}(A; M, Z)$ can be decomposed as:

$$\begin{aligned} \mathcal{I}(A; M, Z) &= \{\mathcal{H}(A) - \mathcal{H}(A|M)\} + \{\mathcal{H}(A|M) - \mathcal{H}(A|Z, M)\} = \\ &= \mathcal{I}(A; M) + \mathcal{I}(A; Z|M), \end{aligned} \tag{A.10}$$

where the first term is the amount of information provided by M and the second term is the additional information provided by Z.

A.4 Anomaly Detection by Deep Learning

A.4.1 Alternative Expression for ELBO

The VAE uses the variational inference described in Section 2.1.4 to obtain a DL-based generative model. Although the ELBO corresponds to the Eq. (2.11) and the optimization process is solved by Eq. (2.12), in this section it will be provided an alternative description, without using the Jensen inequality, that considers the encode path in the AE and provides insights about the ELBO.

The encode path describes $p_\theta(z|x)$, where θ are the parameters of the DNN. Based on the description in Section 2.1.4, the variational inference in a VAE aims to approximate $p_\theta(z|x)$ to an easy-to-sample parametric distribution $q_\phi(z|x)$. Consequently, for any choice of inference model

$q_\phi(z|x)$, we have:

$$\begin{aligned}\log p_\theta(x) = E_{q_\phi(z|x)}[\log p_\theta(x)] =& \\ =&E_{q_\phi(z|x)}\left[\log\left(\frac{p_\theta(x,z)}{p_\theta(z|x)}\right)\right] = \\ =&E_{q_\phi(z|x)}\left[\log\left(\frac{p_\theta(x,z)\,q_\phi(z|x)}{q_\phi(z|x)\,p_\theta(z|x)}\right)\right] = \\ =&E_{q_\phi(z|x)}\left[\log\frac{p_\theta(x,z)}{q_\phi(z|x)}\right] + D_{KL}(q_\phi(z|x)\|p_\theta(z|x)).\end{aligned} \tag{A.11}$$

In Eq. (A.11), the first term in the final result corresponds to the ELBO:

$$\begin{aligned}E_{q_\phi(z|x)}\left[\log\frac{p_\theta(x,z)}{q_\phi(z|x)}\right] = E_{q_\phi(z|x)}\left[\log\frac{p_\theta(z|x)\,p_\theta(x)}{q_\phi(z|x)}\right]& \\ = \log p_\theta(x) - D_{KL}(q_\phi(z|x)\|p_\theta(z|x))&\end{aligned} \tag{A.12}$$

Appendix B

Feature Ranking for Diabetic Foot Ulcers Thermograms

In this study, the proposed method for FS depicted in Chapter 5 will be applied to determine relevant features extracted from the DFU dataset used in Chapter 4 (see Section 4.2.1). Unlike the band selection problem that was reformulated considering the hyperspectral signature as a feature vector, in this case a feature extraction process will be carried out from the thermograms. It is a summary of the work presented in [226].

The subset of features of the feature extracted from thermograms was employed as input for a SVM [173] classifier. The SVM classifier was used as a reference, with the aim of assessing the performance of these features. Finally, for comparison purposes, features previously reported as state-of-the-art were also fed to the classifier.

B.1 Feature Extraction

Following the workflow proposed in the INAOE dataset, the IT images from the local dataset were processed to automatically segment the angiosomes, a composite unit of tissues supplied by an artery, as previously described in [383, 206]. By considering these angiosomes, the foot was divided into four regions: Medial Plantar Artery (MPA), Lateral Plantar Artery (LPA), Medial Calcaneal Artery (MCA), and Lateral Calcaneal Artery (LCA), as illustrated in Figure B.1.

This segmentation step was required to extract the features for each angiosome [197], which included the Thermal Change Index (TCI) [383], Estimated Temperature (ET), Estimated Temperature Difference (ETD), the Hot Spot Estimator (HSE) [197], as well as the summarizing statistics (mean, standard deviation, maximum, minimum, skewness, and kurtosis). In addition, these features also extracted for the entire foot, and following previous approaches [197, 384], a class, based on the Normalized Temperature Ranges (NTR), was assigned to each foot.

Regarding the extraction of the TCI feature, despite the extended database containing more control subjects, the average control temperature was kept unchanged, being the values previously reported [383, 206] considered as reference. These values are displayed in Table B.1, and, for comparative purposes, the mean values corresponding to the healthy subjects, from the internal dataset, are also listed.

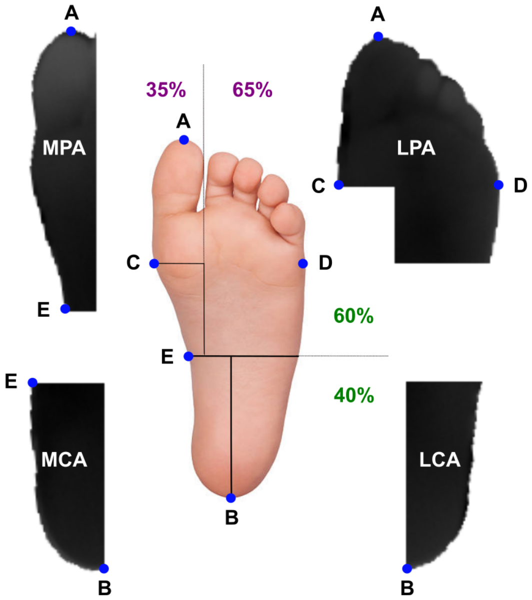


Figure B.1: Graphical illustration of the defined angiosomes. The main reference points considered, and the proportional foot division are also specified [383, 206]. Reference point A was located at the tip of the innermost toe, whereas B at the center of the calcaneal base. Points C and D corresponded to the wider part of the foot. Point E corresponded to the 60% height of the foot.

Table B.1: Mean temperature values per angiosome in the control group for the INAOE database [383] and the internal dataset described in Section 4.2.1. SD indicates standard deviation.

Angiosome	**INAOE**		**Local**	
	$\overline{T}$ (ºC)	SD	$\overline{T}$ (ºC)	SD
MPA	25.8	1.4	25.0	3.0
LPA	25.7	1.3	24.5	3.1
MCA	26.4	1.3	24.5	2.5
LCA	26.1	1.4	24.5	2.6

Table B.2: Classes defined according to the temperature of the foot's thermograms.

Class	**NTR Class**	**Interval (ºC)**	**Classmark (ºC)**
C_1	NTR_C_1	[18,22)	20.0
C_2	NTR_C_2	[22,26)	24.0
C_3	NTR_C_3	[26,27)	26.5
C_4	NTR_C_4	[27,28)	27.5
C_5	NTR_C_5	[28,29)	28.5
C_6	NTR_C_6	[29,30)	29.5
C_7	NTR_C_7	[30,31)	30.5
C_8	NTR_C_8	[31,32)	31.5
C_9	NTR_C_9	[32,33)	32.5
C_{10}	NTR_C_{10}	[33,37)	35.0

Furthermore, in order to extract the ET as well as the subsequently associated parameters, ETD and HSE, thermograms were clustered into classes based on temperature ranges. In the original study [197], Peregrina et al. used a dataset in which the feet were not segmented, so the background objects and their respective temperatures were present in the images. As a consequence, the classes were defined, from C_0 to C_7, whose temperatures were within the interval [25, 35) ºC and, excluding C_0, each class covered 1 °C. To avoid high temperatures from heat sources unrelated to the feet, temperatures between [25, 28) ºC were considered cold and associated with the background (C_0). The other classes were selected according to previously reported data [385], in which subjects with diabetes had a mean temperature of 30.2 ± 1.3 ºC, whereas for healthy subjects, the mean temperature was 26.8 ± 1.8 ºC. Because the dataset used in this work was previously segmented, the range of temperatures to be considered was extended, covering the interval [18, 37) ºC; therefore, the classes were redefined as listed in Table B.2. In this way, the complete range of temperatures in the dataset was taken into consideration. As can be observed, the number of classes was extended to 10, covering approximately 1 ºC, except C_1, C_2, and C_{10}. Furthermore because the considered temperature interval was extended, the number of NTR classes was subsequently adjusted regarding the original study [197, 384]. Finally, the mean value of the established intervals, the classmarks, were used for ET, ETD, and HSE feature extraction.

The nomenclature employed to name the aforementioned extracted features consisted of using a letter to specify the foot, 'L' for left and 'R' for right, followed by the name of the corresponding angiosome. For the features extracted using the entire foot, this second descriptor was discarded. Then, the variable was set using lowercase letters such as mean, std, max, min, skew, or kurtosis. Capital letters were employed for TCI, HSE, ET, and ETD, as well as for NTR followed by the subsequent class.

B.2 Feature Selection

In this study, the number of input variables extracted from thermograms was as high as *188*, and a detailed investigation was proposed to detect the most relevant features based on different approaches. These included some classical methods, RF and LASSO, as well as the proposed approach for FS (see Section 5.2.1) using both Bernoulli relaxation, the Concrete and Gaussian approach (see Section 5.2.2).

Firstly, the original input set was optimized by removing highly correlated variables. The correlation estimation was carried out using the well-known Pearson correlation coefficient; therefore, those features with a correlation $r > 0.95$ were considered highly correlated. For instance, a high correlation between the mean value and the ET was observed, allowing a reduction in the number of features. Then, a feature ranking based on logistic regression was developed to select the most informative variables among them. These redundant variables were ranked based on an AUC_{ROC} analysis using the logistic model as an estimator. As a result, the number of features or input variables was reduced a ∼25.5%, *from 188 to 140.*

Five-fold cross-validation was employed, dividing the dataset into five folds (80% training and 20% testing set), and the performance metric for the testing set was computed five times. Therefore, around 196 samples were used for training and 48 for validation, see Table 4.1. The relevance of the features was the average value resulting from the five iterations during the cross-validation.

B.3 Results

The performance of the proposed DL-based FS methods was evaluated by applying $\phi_i > 0.9$ in the Concrete approach to obtain the sparse representation of the original input space. Figure B.2 shows the sparse rate during the training phase of the respective model in each iteration of the cross-validation. As can be observed, the Concrete approach obtained a sparse rate of around 50%, and the Gaussian relaxation approach obtained a sparse rate of around 60% in most of the cases. This means that, in general, more than half of the features were considered irrelevant. Additionally, the proposed FS method started to become sparser in an early epoch, whereas the concrete approach required a higher λ in the regularization, which increases using the scheduler Eq. (5.15). According to the sparse representation, using the test set in each fold, average accuracies were 89.1% and 85.7% for concrete and variational dropout, respectively. In addition, we noticed that using the variational parameter, ϕ, as feature ranking, the most important features were roughly the same in all the experiments. In comparison, the LASSO approach received a sparse rate of 44%, using a lower number of features than the DL approaches, and with an approximate accuracy of 90%. These results were not reliable for comparison purposes because the models were fully optimized, including the hyperparameters, and the test set was not large enough to reject a possible overfitting.

B.3.1 Feature Selection

Following the workflow described in Section B.2, the most relevant features, listed in Table B.3, were extracted for all the approaches considered: LASSO, RF, and the proposed DL-based FS method using the Concrete relaxation and Gaussian relaxation. For the LASSO approach, the feature ranking was estimated by the absolute value of its coefficient. In relation to Concrete and Gaussian proposed FS approach, the variational parameter was used as feature ranking.

The 10 first features extracted for each approach were considered the most relevant and are highlighted in bold in Table B.3. Therefore, approximately 0.05% of the total features extracted were considered relevant. Regarding the distribution of these features by angiosome, MPA and LPA presented the largest number of features with a total of nine and six associated features, respectively. LCA and MCA angiosomes had, respectively, three and four associated features each. For the entire foot, only two associated features were found.

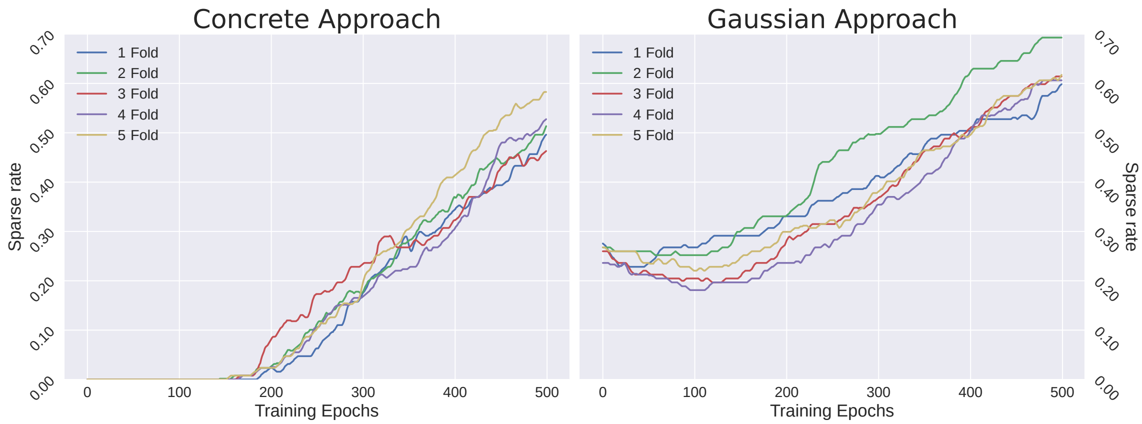


Figure B.2: The sparse rate obtained using the proposed DL-based feature selector in the different cross-validation iterations. The threshold in Concrete approach is $\phi_i > 0.9$.

Table B.3: The 30 most relevant features extracted, listed according to rank, for all the approaches considered: LASSO, Random Forest, and the DL-based feature selection approach using Concrete and Gaussian relaxation. The 10 first features are highlighted in bold as the most relevant for each method. The nomenclature employed is defined in Section B.1.

Rank	LASSO	Random Forest	Concrete Approach	Gaussian Approach
1	**R_LPA_min**	**L_MPA_min**	**R_LPA_min**	**R_LPA_min**
2	**L_LPA_std**	**R_LPA_min**	**R_MCA_std**	**R_MPA_HSE**
3	**Foot_ETD**	**L_MPA_NTR_C$_3$**	**Foot_ETD**	**MCA_ETD**
4	**L_MPA_min**	**R_MCA_std**	**R_LCA_kurtosis**	**L_kurtosis**
5	**L_MPA_skew**	**R_LPA_std**	**R_LPA_std**	**L_MPA_skew**
6	**L_LCA_NTR_C$_4$**	**L_MPA_std**	**L_MPA_min**	**L_MCA_skew**
7	**R_LPA_NTR_C$_3$**	**L_LPA_NTR_C$_2$**	**LPA_ETD**	**R_LCA_kurtosis**
8	**R_MPA_NTR_C$_4$**	**L_LPA_std**	**L_LPA_std**	**R_LPA_std**
9	**L_MPA_HSE**	**R_LCA_NTR_C$_2$**	**L_MCA_skew**	**L_MPA_NTR_C$_4$**
10	**R_MCA_std**	**R_MPA_NTR_C$_2$**	**L_MPA_HSE**	**L_LCA_std**
11	LCA_ETD	R_MPA_std	R_LCA_skew	R_LPA_HSE
12	MCA_ETD	R_LCA_mean	MCA_ETD	R_LCA_std
13	R_LCA_kurtosis	L_MCA_min	L_MCA_std	Foot_ETD
14	R_MPA_NTR_C$_3$	L_LCA_NTR_C$_2$	MPA_ETD	R_MCA_std
15	R_LCA_NTR_C$_3$	L_MCA_mean	LCA_ETD	LPA_ETD
16	L_kurtosis	R_MPA_ET	R_MCA_skew	L_MCA_std
17	R_std	R_std	L_MPA_skew	R_MCA_skew
18	R_LCA_HSE	L_MCA_NTR_C$_2$	R_kurtosis	L_MCA_NTR_C$_5$
19	R_skew	L_LPA_ET	R_HSE	R_MCA_HSE
20	L_HSE	L_LPA_NTR_C$_1$	L_LCA_kurtosis	R_kurtosis
21	R_LPA_NTR_C$_5$	L_LCA_NTR_C$_3$	R_LCA_std	LCA_ETD
22	L_max	Foot_ETD	L_MPA_std	R_LCA_skew
23	L_MCA_std	LPA_ETD	L_LCA_std	R_MPA_std
24	L_LPA_NTR_C$_4$	L_MPA_NTR_C$_4$	R_LCA_HSE	R_MPA_NTR_C$_4$
25	L_MCA_NTR_C$_3$	L_NTR_C$_3$	R_skew	L_MPA_NTR_C$_3$
26	LPA_ETD	L_std	R_MPA_NTR_C$_4$	L_LPA_NTR_C$_2$
27	R_MCA_HSE	L_kurtosis	R_MPA_HSE	R_LCA_HSE
28	L_MCA_skew	R_MPA_NTR_C$_3$	L_MCA_kurtosis	L_MPA_HSE
29	L_LPA_NTR_C$_5$	L_MCA_std	L_LCA_NTR_C$_4$	L_MCA_HSE
30	R_NTR_C$_5$	L_LCA_max	L_kurtosis	L_skew

Table B.4: Most relevant features that coincided in all the approaches considered, listed according to rank.

Rank	Features in Coincidence
Rank < 10	R_LPA_min
Rank < 20	R_MCA_std
Rank < 30	Foot_ETD
	LPA_ETD
	L_MCA_std
	L_kurtosis
Rank < 50	L_LPA_std
	R_kurtosis
	R_LCA_std
	R_LCA_kurtosis

Table B.5: Performance metrics of the optimized SVM classifier using all available features as input.

Input Dataset	Accuracy	Precision	Recall	F1 Score
All features	0.9099 ± 0.0613	0.9473 ± 0.0705	0.8535 ± 0.1016	0.8965 ± 0.0837

Furthermore, the ten first features found to appear in all the implemented approaches are listed by rank in Table B.4. The ranks of these features changed according to the approach employed. Thus, the lowest rank of each feature, among the different approaches, was assigned as its final rank. The search for coincidence was restricted to the first 30 ranked features provided for each approach. However, as observed in Table B.4, the assigned ranks are listed in intervals ranging from ten units. Features found up to a rank lower than 50 were considered. As noticed, if only the 10 first features in coincidence were considered, the angiosomes with more associated features were LPA and the entire foot, with three associated features both, whereas MCA and LCA had two associated features each. No associated features were found for the MPA angiosome in this case.

An SVM classifier was used with all the features as input to provide a reference aiming to quantify the performance of the extracted features, their rank, and selected combination. Initially, using the available features as input, the SVM classifier was optimized using a randomized search [386] to obtain the best parameters. As a result, the RBF kernel, was used. The RBF kernel has a hyperparameter, γ, that controls the spread of the Gaussian center. In addition, the hyperparameter C in SVM is used for directing the L_2 penalty, which controls the trade-off between decision boundary and misclassification. The best performance, displayed in Table B.5, was achieved with a γ value of 0.0035 and a C value of 7.743.

B.3.2 Evaluation of Features by SVM Classifier

Several experimental settings were considered to evaluate the extracted features for the chosen classification task, which was to distinguish between healthy and diabetic patients. In this case, the SVM classifier was not optimized; that is, standard hyperparameters were chosen to offer a fair comparison between the proposed approaches to rank the features. For the different experiments described in this section, γ was set to 0.1, motivated by the low dimensional space of the input data.

In addition, the hyperparameter C was set to 1. This configuration was the same for the different selected features, trying to avoid bias in the conclusions due to well-fitted settings for the indicated features. The average value resulting from five-fold cross-validation, testing the models five times, was used for the metrics estimation depicted in Table B.6, as previously reported [384].

First, the SVM was fed with the ten first features extracted for each approach, LASSO, RF, and the proposed DL-based FS method using the Concrete relaxation and Gaussian relaxation (features highlighted in bold in Table B.3). Second, the ten first features in coincidence, this is, those that appeared in all the approaches and are listed in Table B.4, were also employed to feed the classifier. Finally, to compare the features extracted and the subsequent classification task with those from a previous study [384], the following ten ranked features were also considered: TCI, NTR_C$_4$, NTR_C$_3$, MPA_mean, LPA_mean, LPA_ET, LCA_mean, highest temperature, NTR_C$_2$, and NTR_C$_1$. These features were among the top ten features resulting from testing several techniques, which included Pearson, chi square, RFE, logistics, RF, and LightGBM. The metrics of the performance for each approach, according to the experimental settings described, are listed in Table B.6.

Notice that, contrary to the setup employed in the present study in which all features were extracted by foot, L or R, the foot to which the previously mentioned features were associated was not specified in [384]. Therefore, the mean value between both feet was calculated to match these features and offer a fair comparison. In addition, the NTR class definition considerably differed from the one considered previously; thus, the equivalent class, based on temperature values, was used instead. NTR_C$_4$ and NTR_C$_3$ in the original study corresponded to ranges between 31 and 32 ºC as well as 30 and 31 ºC, respectively [197, 384]. In the present study, the closest approximations were NTR_C$_8$ and NTR_C$_7$, for which the respective ranges coincided with the ranges mentioned above.

Considering the features extracted for each approach and the subsequent classification task, all approaches provided good metric values. However, the best scores, except for the recall parameter, were observed for the concrete dropout approach. When the set of relevant features were those common to all the approaches, although at different rank positions, the performance in this experimental setting provided the best recall. Furthermore, as noticed, the recall values were lower in comparison with the other parameters of the performance metrics. This may have been due to the imbalance between healthy and diabetic samples from the original dataset because a low recall score is associated with a high number of false negatives. A relevant number of healthy samples was generated for balancing using SMOTE, which performed a linear interpolation between samples. Therefore, recall was penalized because it was exclusively dependent on the diabetic samples. In this case, considering the precision-recall tradeoff, a lower recall was preferred due to the associated implications.

As shown in Table B.6, the performance of all the models when using the corresponding first 10 features as well as when using the first 10 features in coincidence, was quite similar to those considered as reference values (shown in Table B.5). However, the classical LASSO approach and DL-based FS using the Concrete approach exhibited a slightly better performance with only 10 features.

Table B.6: Performance metrics of the approaches considered, according to the selected input features, in each experimental setting. The highest value for each performance metric is highlighted in bold.

Input Dataset	**Approach**	**Accuracy**	**Precision**	**Recall**	**F1 Score**
First 10 features	LASSO	0.8975 ± 0.073	0.9533 ± 0.079	0.8361 ± 0.130	0.8908 ± 0.107
	Random Forest	0.8893 ± 0.070	0.9703 ± 0.080	0.8033 ± 0.118	0.8789 ± 0.103
	Concrete Approach	**0.9098 ± 0.069**	**0.9808 ± 0.057**	0.8361 ± 0.131	**0.9027 ± 0.104**
	Gaussian approach	0.8934 ± 0.054	0.9615 ± 0.049	0.8197 ± 0.104	0.8850 ± 0.081
First 10 features in coincidence	-	0.9057 ± 0.066	0.9626 ± 0.052	**0.8442 ± 0.135**	0.8995 ± 0.102
Selected features from [384]	-	0.7951 ± 0.075	0.8750 ± 0.136	0.6885 ± 0.089	0.7706 ± 0.103

B.4 Discussion

Several approaches were considered to select relevant features used for DFU detection based on infrared thermograms. Classical approaches, LASSO and RF, were tested versus the proposed DL-based FS using the Concrete and Gaussian relaxation. The outputs of these approaches were analyzed to extract a new set of features considered relevant to classify whether a thermogram corresponded to a healthy or diabetic person.

Regarding the performance of the traditional approaches in comparison with that of the DL-based ones, LASSO provided results close to those of the Concrete approach, according to the F1 score, although the latter exhibited a slightly better performance compared with the established reference values. However, LASSO is limited to linear solutions, while the Concrete approach does not suffer from this limitation. No fine-tuning of the models was implemented to increase the respective performance because a comparison between extracted input features was intended. Thus, when a few features were used, i.e., 10, both methods produced promising performance. In this particular case, the LASSO approach would be an easy-to-implement and faster alternative to concrete dropout, as comparable performance was achieved. Furthermore, considering the most relevant features of each approach, six of the selected features matched for these two approaches, see Table B.3. Thus, the similarity in performance may have been due to this coincidence of features.

A previous study [384] was considered as a reference to quantify the performance of the extracted features for the classification task. This reference study employed a stacking classifier using gradient boost, XGBoost and RF considering previously ranked features as input: TCI, NTR_C_4, NTR_C_3, MPA_mean, LPA_mean, LPA_ET, LCA_mean, and highest temperature. Using these proposed features, the values reported in the present study, around 77% in the F1 score, are considerably lower than those reported previously (see Table B.6). However, although the definition of the features was slightly modified and the classifier employed considerably differed; the same input features exhibited a roughly 15% lower performance in comparison with the features extracted in this study. This difference may be explained by the use of an extended dataset as well as their proposed new labeling in the INAOE dataset for distinguishing between mild, moderate, and severe cases in the diabetic foot domain, which was not tested in this study.

Before feature extraction, our initial study focused on establishing a balanced dataset of diabetic and healthy subjects by fusing a publicly unbalanced available dataset [206] with a local dataset composed of healthy subjects. Furthermore, the preprocessing of the thermograms was also carefully considered, extracting the features previously reported [206, 384]. Among the set of features, considered within the state-of-the-art features, were the highest temperature, TCI, HSE, ET, NTR, and several statistical variables such as the mean value, being associated with the entire foot as well as some defined angiosomes (see Section B.1). Notably, in the present study, these features were extracted by foot, and considered separately, unlike previous studies in which an average between the R and L foot was assumed due to the lack of specific information on the procedure. For this reason, the number of features extracted is considerably much higher than in previous reports [384], 188 versus 37 features.

Of the most important state-of-the-art features, TCI is especially relevant. The TCI is focused on providing a quantification of the thermal change, independent of the observed distribution, and a difference of 1ºC is considered enough to notice a significant difference between the classes proposed [383]. In this study, reference values were not modified, in comparison with the original study, to calculate the TCI values despite more healthy subjects being considered in the extended dataset. Regardless, none of the features related to TCI were considered relevant among all the implemented approaches or within those features in coincidence.

Regarding NTR, the number of classes based on the thermogram temperatures was extended to 10 because the range of relevant temperatures considered was increased from 18 to 37 ºC compared wtih the original study, which was from 25 to 35 ºC [197]. These modifications were motivated by, first, the exclusion of temperature values characteristic of healthy patients, which were excluded in the original study to avoid the background, because foot segmentation was not available. The private thermograms from healthy volunteers, fused with the INAOE database for balancing purposes, showed that the temperature distributions were below 28 °C for many subjects. Second, these private thermograms were previously segmented; thus, excluding other heat sources within the background was not required. As a result, we did not discard any NTR class as was proposed in [197] for removing the background. A better-performing classifier was expected by extending the temperature ranges. However, most of the features related to the NTR were not considered relevant and, similar to that observed for the TCI, none were obtained within the features in coincidence among all the approaches. Furthermore, according to the F1 score, the best-performing approach was the Concrete approach, and not a single feature related to the NTR was among those considered relevant.

Opposite to that described above, in the present study, among those specially designed features for DFU, only HSE and ETD seem to be relevant. Furthermore, as described, the sole was divided into four different angiosomes, and their individual features were extracted. The extraction demonstrated an unbalanced significance of the angiosomes and, therefore, the division of the foot into angiosomes seemed a determinant factor for feature extraction and played an important role in the analysis. In particular, the LPA angiosome appeared as the most predictive, with more associated features than the other angiosomes, followed by LCA.

Appendix C

Code Availability

> Talk is cheap. Show me the code.
>
> *Linus Torvalds*

The majority of the code developed for this thesis is publicly accessible on my GitHub account, organized into dedicated folders within the 'Thesis' repository. Each chapter-specific folder contains the corresponding code, facilitating reproducibility of the contributions discussed throughout the thesis.

In the 'modules' folder, custom libraries developed for specialized tasks are available. For example, the IPDL Library supports experimental methodologies rooted in the IT perspective, offering MI estimation via a kernel-based entropy estimator and providing visualizations of the IP. Integrated with PyTorch, this library includes kernel optimizers specifically designed to be compatible with PyTorch's architecture. Additionally, the HySpecLab library provides a suite of tools for HSI processing, with several modules designed to be compatible with the Scikit-Learn interface, streamlining the analysis and processing workflows.

Appendix D

Resumen en español

D.1 Introducción y motivación

La inteligencia artificial (IA) ha realizado recientemente avances significativos en percepción (la interpretación de información sensorial), lo que permite a las máquinas representar e interpretar mejor datos complejos. Dicho avance ha estado motivado principalmente por los modelos de aprendizaje profundo (AP), la red neuronal (RN). Recientemente, el impacto de las RNs ha sido particularmente notable en varios aspectos de nuestras vidas. No obstante, las aplicaciones actuales donde la IA está siendo especialmente relevante no son esenciales en nuestras vidas; la mayoría de ellas están asociadas con el entretenimiento. Excluyendo las razones no técnicas, existe una razón específica por la cual la IA basada en RNs aún no se considera adecuada para su integración en campos cruciales: estos modelos no son confiables. Una RN, sigue considerándose una ‘caja negra’, ya que son modelos difíciles de interpretar donde no es posible entender qué ocurre en las capas ocultas.

Por ejemplo, la conducción autónoma y el diagnóstico médico son campos de aplicación críticos donde ser inexacto implica consecuencias desastrosas. La salida de los modelos de AP presenta una mayor incertidumbre que la salida obtenida por métodos clásicos, y más simples. Es más fácil entender los métodos clásicos, y en modelos complejos como las RNs, es mucho más complicado entender por qué un pequeño cambio en la entrada puede afectar drásticamente el rendimiento del modelo. De esta forma, los métodos clásicos podrían considerarse más robustos.

Es evidente que el uso de algoritmos de IA basados en AP son especialmente relevante en la investigación científica contemporánea, con muchos autores incorporando estos algoritmos para resolver problemas que anteriormente eran inabordables. Sin embargo, la falta inherente de ‘transparencia’ en estos modelos introduce un alto nivel de incertidumbre debido a sus complejos mecanismos de resolución de problemas. Esta incertidumbre genera preocupaciones sobre su aplicación en escenarios críticos.

Esta tesis doctoral tiene como objetivo mejorar la interpretabilidad de los modelos de AP, proporcionando una comprensión e intuición más profunda de cómo funcionan estos modelos. La tesis aborda diversos problemas comunes en el contexto médico, donde la IA puede ser beneficiosa. En distintos capítulos se proponen varios métodos para obtener resultados consistentes en estos problemas, ofreciendo además una perspectiva sobre los mecanismos de resolución de problemas empleados por los modelos de AP.

Los resultados obtenidos durante esta tesis han sido posibles gracias a la la estrecha colaboración entre el Instituto Universitario de Microelectrónica Aplicada (IUMA) y el Instituto Universitario de Investigación Biomédica y Sanitaria (IUIBS) de la Universidad de Las Palmas de Gran Canaria (ULPGC) y las siguientes instituciones de investigación:

- Grupo de Tecnología Médica de IACTEC, Instituto de Astrofísica de Canarias (España)
- Centro de Investigación y Ensayos Clínicos en Oncología Experimental, Centro Nacional de Cáncer (Japón)
- Departamento de Ingeniería Eléctrica e Informática, Universidad de Ciencias de Tokio (Japón)

Además, esta investigación se llevó a cabo como parte del proyecto ITHaCA (Identificación Hiperespectral de Tumores Cerebrales), financiado por el Gobierno de Canarias bajo el Acuerdo de Subvención ProID2017010164.

Finalmente, esta tesis doctoral ha sido cofinanciada por la Agencia Canaria de Investigación, Innovación y Sociedad de la Información del Ministerio de Universidades, Ciencia, Innovación y Cultura y por el Programa Operativo Integrado del Fondo Social Europeo Plus (FSE+) de Canarias 2021-2027, Eje 3, Tema Prioritario 74 (85%).

D.2 Evaluación e Interpretabilidad en Aprendizaje Profundo

Los modelos de AP presentan ventajas, pero su implementación en el ámbito médico sigue enfrentando retos, especialmente en términos de interpretabilidad. Este capítulo aborda cómo la naturaleza de ‘caja negra’ de las RNs limita su transparencia y reduce la confianza en estos métodos.

La tesis emplea un enfoque basado en teoría de la información (TI), considerando los modelos de AP como procesadores de información que operan en distintos niveles de abstracción. Así, se propone una interpretación de las RNs que facilita el entendimiento de su mecanismo de resolución de problemas, permitiendo la creación de metodologías para su validación.

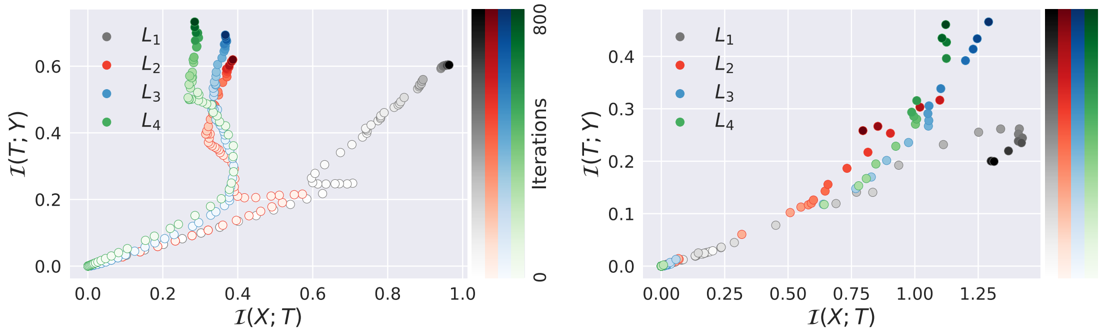
Desde esta perspectiva, las RNs se modelan como cadenas de Markov en cascada, obteniendo representaciones progresivas entre la entrada y la salida. Esto se interpreta como un medio de transmisión en cascada. Mediante el *information bottleneck* (IB), el modelo se describe en función de un punto de equilibrio entre la compresión de la información de entrada X y la predictibilidad de la salida Y, siguiendo la cadena de Markov $X \rightarrow Z \rightarrow \hat{Y}$, siendo $\hat{Y}$ la salida estimada y Z el cuello de botella (el punto de equilibrio). No obstante, debido a sus propiedades Markovianas, una RN con L capas ocultas óptimas debe cumplir las siguientes desigualdades:

$$\mathcal{I}(X;T_1) \geq \mathcal{I}(X;T_2) \geq ... \geq \mathcal{I}(X;T_L), \tag{D.1}$$

$$\mathcal{I}(T_1;Y) \leq ... \leq \mathcal{I}(T_{L-1};Y) \leq \mathcal{I}(T_L;Y). \tag{D.2}$$

Estas desigualdades describen el flujo de información considerando la entrada X (D.1) y la salida Y en la dirección inversa D.2. El operador $\mathcal{I}(.;.)$ representa la información mutua.

No obstante, otro desafío relevante en entornos médicos es la escasez de datos para poder entrenar y validar los modelos de AP. Esto se debe a la dificultad de realizar una recolección sistemática de datos para crear grandes conjuntos de datos para entrenar modelos de AP. Como resultado, es



(a) (b)

Figure D.1: *Information Plane* in the different experiments: (a) Modelo con técnicas para evitar el sobreajusta, y (b) modelo sin técnicas para evitar overfitting.

posible obtener conjuntos de datos de entrenamiento poco representativos, lo que puede introducir sesgos en los algoritmos. A su vez, usando métodos clásicos de evaluación basados en métricas de clasificación como 'exactitud' o 'precisión' también se encuentran sesgados.

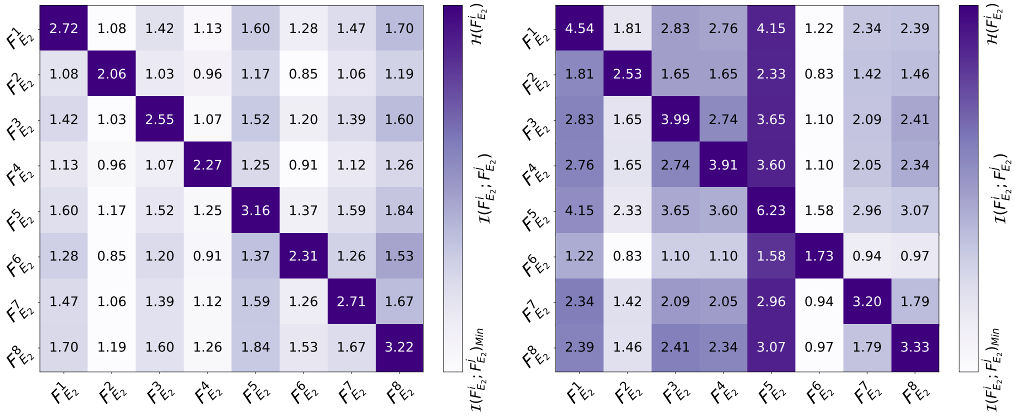
Como resultado, se propone un método de evaluación cuantitativa y cualitativa basada en las propiedades de las desigualdades (D.1) y (D.2) que no requiere un número elevado de muestras de *test*. Esto se logra mediante el uso de un método de estimación de entropía basado en *kernels*. Estos estimadores basados en núcleos no solo son matemáticamente robustos, sino también computacionalmente eficientes, lo que los hace particularmente adecuados para RNs donde las salidas frecuentemente se manifiestan en espacios de alta dimensión.

Partiendo de un dataset con pequeño, tan solo 241 muestras, se divide el dataset para tener 32 muestras para evaluar el modelo, teniendo tan solo 209 muestras de entrenamiento. Este reducido número de muestras conlleva posibles problemas de sobre-ajustes de los modelos. Finalmente, se realizan experimentos para comprobar el método de evaluación propuesto, midiendo la propagación de información a lo largo de las distintas capas de la RN y empleando una herramienta de visualización llamada *Information Plane* (IP).

La Fig. D.1 ilustra los resultados obtenidos; en uno de los modelos se emplean distintas técnicas de regularización para evitar el sobre-ajuste (Fig. D.1a), mientras que el otro se entrena sin regularización (Fig. D.1b). A pesar de que ambos modelos exhiben el mismo rendimiento en términos de exactitud, superior al 90%, se observa que las desigualdades (D.1) y (D.2) solo se satisfacen en la Fig. D.1a, como era de esperar.

Además, dada la similitud de los AutoEncoders (AEs) con canales de comunicación, se evalúa el efecto del dropout en la compresión de información en RNs. Como muestra la Fig. D.2, el dropout ayuda a homogeneizar la información entre filtros y aumentar la redundancia, reduciendo la entropía por filtro al incrementar el ratio de dropout. Esto demuestra cómo el dropout contribuye a una mayor compresión y robustez en las representaciones de los datos.

Los resultados revelaron que la técnica de dropout se alinea con la *distortion-rate theory* (teoría de tasa-distorsión) para lograr representaciones comprimidas óptimas, proporcionando una nueva interpretación de esta técnica. En consecuencia, la tasa de dropout impacta directamente en la compresión de información de los datos de entrada.



	$F^1_{E_2}$	$F^2_{E_2}$	$F^3_{E_2}$	$F^4_{E_2}$	$F^5_{E_2}$	$F^6_{E_2}$	$F^7_{E_2}$	$F^8_{E_2}$
$F^1_{E_2}$	2.72	1.08	1.42	1.13	1.60	1.28	1.47	1.70
$F^2_{E_2}$	1.08	2.06	1.03	0.96	1.17	0.85	1.06	1.19
$F^3_{E_2}$	1.42	1.03	2.55	1.07	1.52	1.20	1.39	1.60
$F^4_{E_2}$	1.13	0.96	1.07	2.27	1.25	0.91	1.12	1.26
$F^5_{E_2}$	1.60	1.17	1.52	1.25	3.16	1.37	1.59	1.84
$F^6_{E_2}$	1.28	0.85	1.20	0.91	1.37	2.31	1.26	1.53
$F^7_{E_2}$	1.47	1.06	1.39	1.12	1.59	1.26	2.71	1.67
$F^8_{E_2}$	1.70	1.19	1.60	1.26	1.84	1.53	1.67	3.22

(a)

	$F^1_{E_2}$	$F^2_{E_2}$	$F^3_{E_2}$	$F^4_{E_2}$	$F^5_{E_2}$	$F^6_{E_2}$	$F^7_{E_2}$	$F^8_{E_2}$
$F^1_{E_2}$	4.54	1.81	2.83	2.76	4.15	1.22	2.34	2.39
$F^2_{E_2}$	1.81	2.53	1.65	1.65	2.33	0.83	1.42	1.46
$F^3_{E_2}$	2.83	1.65	3.99	2.74	3.65	1.10	2.09	2.41
$F^4_{E_2}$	2.76	1.65	2.74	3.91	3.60	1.10	2.05	2.34
$F^5_{E_2}$	4.15	2.33	3.65	3.60	6.23	1.58	2.96	3.07
$F^6_{E_2}$	1.22	0.83	1.10	1.10	1.58	1.73	0.94	0.97
$F^7_{E_2}$	2.34	1.42	2.09	2.05	2.96	0.94	3.20	1.79
$F^8_{E_2}$	2.39	1.46	2.41	2.34	3.07	0.97	1.79	3.33

(b)

Figure D.2: Estimación de información mutua entre diferentes filtros en la capa de convolución del AE con distintos ratios de dropout, denotado como F^i_k donde i representa el índice del filtro y k la capa del codificador. Los valores de dropout utilizados son (a) 0.5 y (b) 0.2. La diagonal del mapa de calor representa la entropía del filtro.

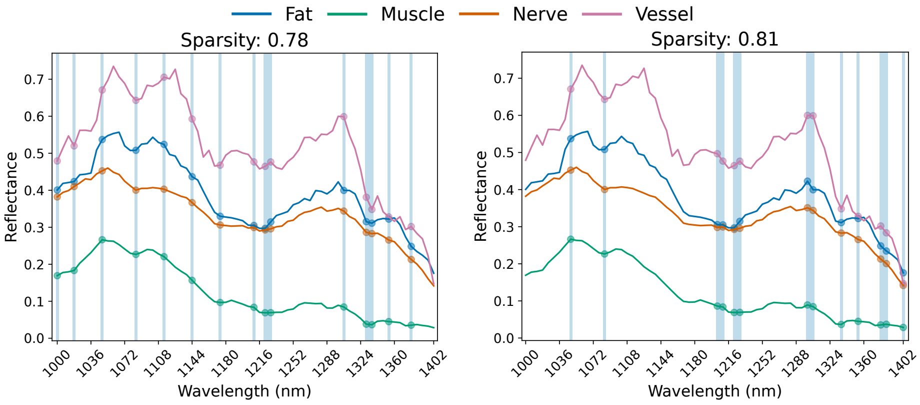
D.3 Selección de Bandas Hiperespectrales con Aprendizaje Profundo

Inspirado por el enfoque basado en TI, se propone un método de selección de características (SC) basado en AP. Este método tiene como objetivo seleccionar un subconjunto de características de la entrada que contenga la información necesaria para una tarea dada. En escenarios clínicos, esto reduciría el número de características, disminuiría los costos de análisis y mejoraría la interpretabilidad de los datos, reduciendo así la variabilidad en la toma de decisiones clínicas basadas en datos de alta dimensión.

Para tareas de SC se propone una arquitectura de RN donde, basándose en el principio de IB, el cuello de botella Z se encuentra en la primera capa de la RN. Es decir, la selección de subconjunto de características corresponde al cuello de botella del modelo. Para ello, la capa de selección de características, emplea la técnica de dropout, la cual puede ser formulada mediante la teoría tasa-distorsión. Como resultado, la tasa de dropout es un parámetro del modelo, que se ajusta durante el entrenamiento utilizando un enfoque variacional.

Usando el enfoque basado en TI, se dedujo que la clasificación y la SC, en el método propuesto, pueden tratarse como problemas independientes, lo que llevó a proponer un gestor del factor de regularización para la dispersión de datos. Además, se destacó la inicialización óptima de la capa de selección de características, basada en dropout.

El método se evaluó utilizando un estudio de caso enfocado en identificar longitudes de onda relevantes para diferenciar tejidos biológicos mediante una modalidad de imagen conocida como *hyperspectral imaging* (HSI) o imagen hiperespectral. Los resultados fueron consistentes con los hallazgos de la literatura, mostrando patrones en la selección de bandas que se fundamentan en las propiedades espectrales conocidas de los distintos tejidos. Además, se comparó el método con modelos encontrados en el estado del arte, logrando subconjuntos más pequeños, lo cual ayuda a reducir el costo de análisis y mejora la interpretabilidad de los datos. La Fig. D.3 ilustra una



(a)

(b)

Figure D.3: Selección de bandas en datos HSI mediante (a) LASSO y (b) el método propuesto en esta tesis basada en DL.

comparación de la selección de bandas más relevantes usando el método más popular del estado del arte, LASSO (Fig. D.3b), y el obtenido por el método propuesto en esta tesis (Fig. D.3a).

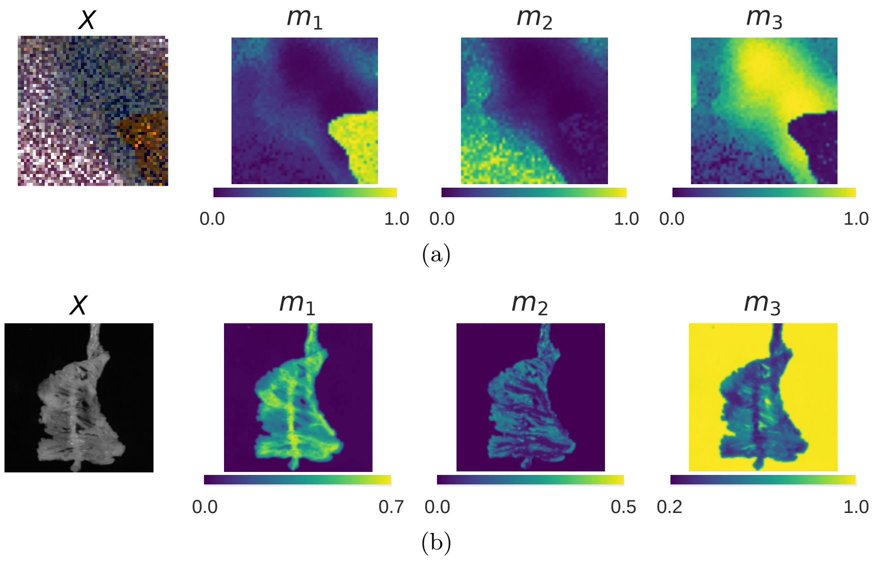
Los resultados fueron validados, además de con métricas de clasificación, mediante la comparación con las longitudes de onda específicas propuestas en el estado del arte, como la elección de un rango alrededor de 1200 nm para identificar tejido graso y muscular. Esto es relevante porque en el tejido muscular no hay presencia de lípidos, y los coeficientes de absorción del agua (abundante en el músculo) y de los lípidos son comparables. Otro ejemplo es la selección de la banda correspondiente a 1084 nm por el método propuesto, ya que en la literatura se menciona que el colágeno del tejido nervioso puede presentar ligeros picos de absorción alrededor de los 1100 nm.

Además de los resultados presentados, se ha llevado a cabo un análisis exhaustivo de los hiperparámetros asociados al método propuesto. Esto demuestra la simplicidad en la selección de hiperparámetros en comparación con otras implementaciones de SC basadas en AP, como, por ejemplo, LassoNet.

D.4 Desmezclado Hiperespectral mediante Aprendizaje Profundo

A lo largo de este capítulo se presenta un método de *hyperspectral unmixing* (HSU) o desmezclado hiperespectral mediante un modelo de AP inspirado en métodos de *contrastive learning* (CL), aprendizaje contrastivo. Al igual que en la tarea de SC, el HSU es de especial interés para mejorar la interpretabilidad de los datos, descomponiendo en las señales en un conjunto de ‘firmas puras’ (*endmembers*) y sus proporciones correspondientes. El método propuesto, titulado *Contrastive Learning for Blind Hyperspectral Unmixing* (CLHU), presenta una innovadora madera para resolver los problemas de HSU mediante AP.

Aunque el método no sigue la estructura clásica de los métodos CL, el enfoque basado en TI permite ilustrar cómo el método emplea un enfoque contrastivo. Adicionalmente, debido a la arquitectura del modelo, sse identifican patrones teóricos que deben cumplirse para garantizar el funcionamiento correcto del modelo. Como resultado, se plantean nuevas desigualdades que deben



(a)

(b)

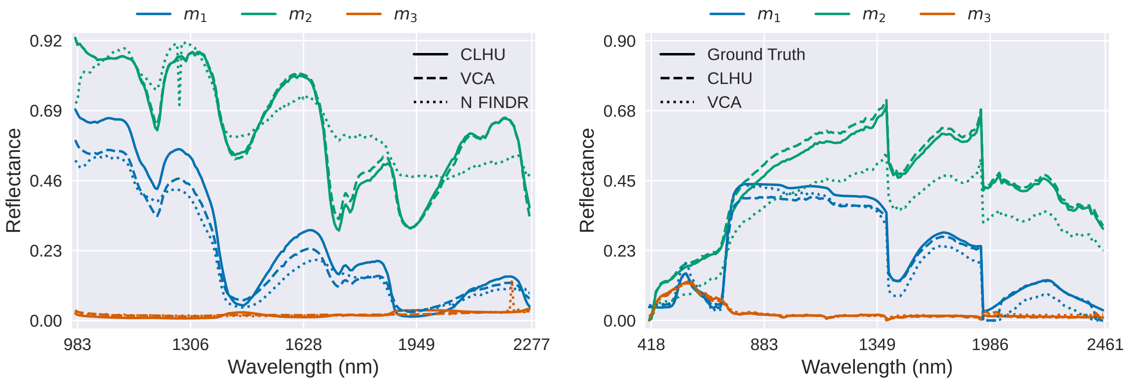
Figure D.4: Mapas de abundancia obtenidos mediante el método propuesto, CLHU, usando dos escenarios, donde (a) la descomposición de la escena corresponde a un caso lineal con ruido blanco aditivo y (b) la escena representa un caso multicapa no lineal. La entrada X corresponde a una representación de la imagen hiperespectral a descomponer.

satisfacerse, partiendo de una interpretación basada en un sistema de transmisión de información de múltiples canales, lo cual proporciona una mejor comprensión del mecanismo empleado para resolver el problema. Finalmente, se concluye que la arquitectura propuesta podría extenderse a tareas más allá del HSU, como la clasificación con estimación de incertidumbre, donde el descriptor entrenable describe las clases y la función de similitud cuantifica la incertidumbre del modelo.

Para la validación del modelo, se realizan diferentes experimentos que incluyen datos sintéticos, datos reales utilizados en el estado del arte y datos orientados a aplicaciones médicas. Los resultados obtenidos son prometedores y compiten con los métodos actuales en el estado del arte. No obstante, dado que el método propuesto representa una nueva perspectiva para abordar las tareas de HSU, los resultados obtenidos tienen margen de mejora.

Las Fig. D.4 y D.5 ilustran resultados obtenidos considerando un escenario donde (a) la descomposición es lineal con ruido blanco aditivo, y (b) la descomposición es no lineal. Con respecto a Fig. D.4, esta muestra una representación de la imagen hiperespectral a descomponer, X, y los mapas de abundancia obtenidos para los distintos componentes o firmas puras (*endmembers*) usando el método propuesto. Finalmente, Fig. D.5 muestra la estimación de los *endmembers* mediante CLHU y los obtenidos por otros métodos del estado del arte.

Al igual que en los capítulos anteriores, se empleó un estimador de entropía basado en *kernels* para hacer una estimación del IP en tiempo de entrenamiento. Los resultados obtenidos garantizan que las propiedades descritas mediante TI se cumplen cuando las condiciones son correctas. Por ejemplo, si al realizar el HSU se indica un número erróneo de firmas puras, se observa que las desigualdades establecidas no se satisfacen. De esta forma, se valida la interpretabilidad basada en comunicación multicanal y se plantean posibles estrategias futuras para estimar el número de endmembers presentes en la escena.



(a) (b)

Figure D.5: Estimación de las firmas puras, o *endmembers*, obtenidas mediante el método propuesto, CLHU, y comparado con los obtenidos con algoritmos del estado del arte. Ambos resultados corresponden a escenarios donde (a) la descomposición de la escena corresponde a un caso lineal con ruido blanco aditivo y (b) la escena representa un caso multicapa no lineal. La Fig. D.4 muestra los datos de entrada y la interpretación de las correspondientes firmas puras.

D.5 Detección de Anomalías por Aprendizaje Profundo

En este capítulo se aborda una tarea crucial en el análisis de conjuntos de datos: la detección de anomalías (DA). Se propone un método de DA basado en modelos de AP, denominado *Anomaly Detection in Latent Spaces using Entropy-based* (ADeLEn), diseñado específicamente para los desafíos de la imagen médica. El método adopta un enfoque semi-supervisado, teniendo en cuenta el 'desafío de datos' (es decir, la carencia de suficientes datos para entrenar modelos basados en AP), que es común en contextos médicos. Fue evaluado con conjuntos de datos fuertemente desbalanceados, considerando tanto el número de muestras anómalas como la 'contaminación' del conjunto de datos, que se refiere a muestras anómalas no identificadas usadas durante el entrenamiento. En los resultados experimentales, el método propuesto se comparó con el *One-Class Support Vector Machine* (OCSVM) del estado del arte y una RN supervisada. Los resultados revelaron que el método propuesto es más versátil que el OCSVM y no sufre de sobreajuste, como sucede con el enfoque supervisado de RN en conjuntos de datos desbalanceados.

El objetivo es extender enfoques estadísticos paramétricos a datos complejos. Dado que las imágenes médicas suelen compartir patrones semejantes (por ejemplo, las imágenes del cerebro presentan características comunes entre pacientes), se asume que la distribución de muestras normales y anómalas sigue patrones similares. En el caso del cerebro, un cerebro con tumor (una muestra anómala) seguiría una distribución similar a la de un cerebro sano. Por lo tanto, se propone usar una RN para obtener una compresión de los datos cuyas distribuciones sean más sencillas de modelar.

No obstante, obtener las distribuciones marginales, incluso de la representación comprimida en el espacio latente, es generalmente intratable. Por esa razón, el método ADeLEn adopta un enfoque generativo, empleando un *Variational AE* (VAE), para aplicar el enfoque estadístico sobre la distribución de los datos en el cuello de botella. De esta forma, se asume que la representación comprimida de los datos normales y anómalos sigue una distribución normal, diferenciándose por la

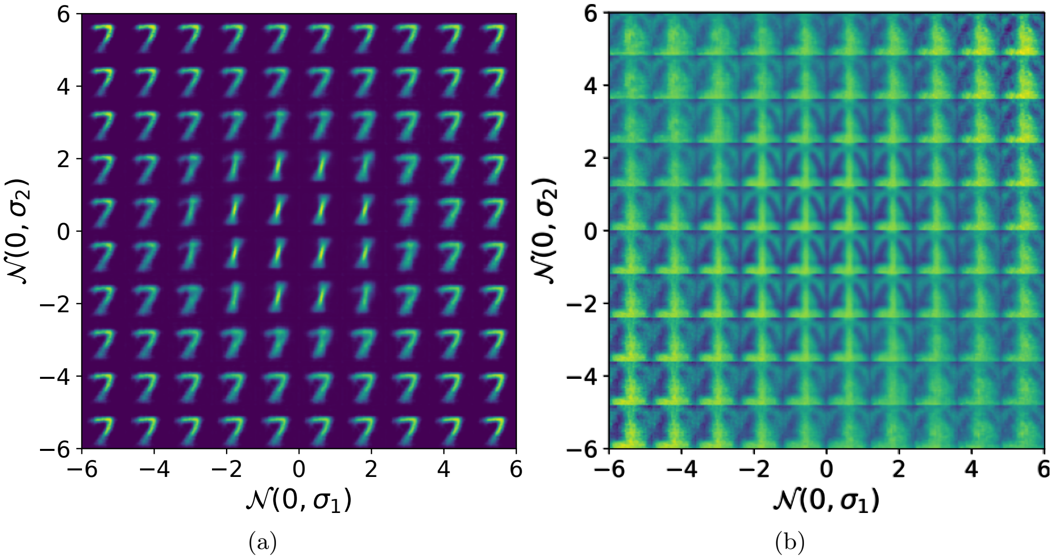


Figure D.6: Reconstrucciones obtenidas por ADeLEn partiendo de un cuello de botella representado por una distribución de dos variables. Los resultados corresponden a dos experimentos distintos, (a) usando un subconjunto de MNIST donde el '1' representa muestras normales y '7' anomalies (siendo estas últimas un subconjunto más pequeño) y (b) usando un dataset de detección de neumonía que presenta las mismas condiciones.

desviación estándar de ambas distribuciones, pero teniendo misma media:

$$\begin{aligned} Z^+ &\sim \mathcal{N}(0; \sigma^+ I), \\ Z^- &\sim \mathcal{N}(0; \sigma^- I), \\ s.t. \quad & \sigma^- > \sigma^+, \end{aligned} \tag{D.3}$$

donde Z^+ representa la representación comprimida de las muestras normales y Z^- la representación de las muestras anómalas. La detección de anomalías consiste en aplicar un umbral del valor de la entropía, el cual en este caso es directamente proporcional al valor de σ.

Como se explicó anteriormente, el algoritmo fue diseñado específicamente para imágenes médicas, que comparten una gran cantidad de patrones entre muestras normales y anómalas, y en las que el número de muestras suele estar desbalanceado. Se realizaron diversos experimentos para demostrar la robustez del método frente a la contaminación del dataset, asumiendo que este contiene muestras anómalas no identificadas, y se evaluó cómo actúa ante anomalías con patrones distintos a los observados por el modelo. Finalmente, se comprobó que el método es más robusto frente al sobreajuste que un simple modelo entrenado de forma supervisada.

La Fig. D.6 presenta un ilustrativo ejemplo de los resultados obtenidos para dos casos específicos. La Fig. D.6a muestra un experimento usando el dataset MNIST, en el que las muestras de '1' representan normales y las de '7' representan anomalías, siendo el número de muestras anómalas considerablemente inferior al de muestras normales. En el caso de la Fig. D.6b emplea un dataset para la detección de neumonía bajo condiciones similares. Aunque la Fig. D.6a facilita una interpretación más clara, en ambas figuras se observa cómo el modelo reconstruye muestras más 'anómalas' al tener un valor de σ mayor.

Finalmente, se emplea la TI para dar una mayor intuición sobre el mecanismo usado en el método propuesto para la DA. En este caso, el enfoque variacional en el modelo generativo aplica una fuerte restricción sobre la información en el cuello de botella Z, estableciendo un límite superior definido por los parámetros variacionales ϕ, asegurando que la distribución de los datos reales en el cuello de botella no contenga más información que la distribución variacional. Además, el problema fue reformulado utilizando el principio de InfoMax, en el que el cuello de botella está diseñado para maximizar la divergencia entre las distribuciones de muestras normales y anómalas, proponiendo una inicialización óptima de los parámetros variacionales para el método ADeLEn.

D.6 Conclusiones

A lo largo de esta tesis, se han propuesto y evaluado varias arquitecturas de modelos basados en AP en relación con tareas específicas. Sin embargo, el objetivo de este trabajo no es superar los resultados del estado del arte, sino explorar y desarrollar soluciones más interpretables utilizando AP. Para efectos comparativos, los métodos propuestos se han evaluado frente a varios algoritmos novedosos para asegurar que los modelos produzcan resultados competitivos. Además, se han abordado conceptos clave en un contexto más amplio, con el propósito de proporcionar una intuición general sobre los problemas estudiados y los mecanismos necesarios para abordarlos.

Esta tesis busca enfatizar la importancia de trabajar hacia modelos más interpretables. No pretende criticar el uso de una 'caja negra' para resolver problemas, ya que existen escenarios críticos en los que, debido a la complejidad de los problemas, es necesario utilizar una 'caja negra', pues ciertos problemas no se pueden abordar con otros métodos. Sin embargo, el valor de este trabajo radica en cambiar el enfoque desde la suposición de que solo se pueden lograr predicciones precisas aumentando el número de parámetros en la 'función universal' de la red neuronal, hacia la exploración de arquitecturas alternativas que equilibren precisión e interpretabilidad.

Aunque soy consciente de que la interpretabilidad presentada en esta tesis sigue una perspectiva basada en un enfoque orientado a la ingeniería, es importante entender que la interpretabilidad es un concepto específico del dominio, y las interpretaciones que funcionan en un contexto pueden no ser aplicables en otro. No obstante, a lo largo de esta tesis, la TI se presenta como un marco fundamental para lograr una interpretabilidad de propósito general que pueda usarse en cualquier contexto. Al conceptualizar los modelos como sistemas que procesan y transmiten información, podemos desviar nuestro enfoque de la estructura detallada de los datos o de las explicaciones de características en diversas capas ocultas. En su lugar, la preocupación principal se centra en la cantidad de información capturada y transmitida a través de las representaciones de características del modelo.

Bibliography

[1] François Chollet. “On the measure of intelligence”. In: *arXiv preprint arXiv:1911.01547* (2019).

[2] Alan M Turing. *Computing machinery and intelligence*. 1950.

[3] Richard M Friedberg. “A learning machine: Part I”. In: *IBM Journal of Research and Development* 2.1 (1958), pp. 2–13.

[4] Ledley Rs and LUSTED LB. “Reasoning foundations of medical diagnosis; symbolic logic, probability, and value theory aid our understanding of how physicians reason.” In: *Science (New York, NY)* 130.3366 (1959), pp. 9–21.

[5] Ahmed Hosny et al. “Artificial intelligence in radiology”. In: *Nature Reviews Cancer* 18.8 (2018), pp. 500–510.

[6] Xiaoxuan Liu et al. “A comparison of deep learning performance against health-care professionals in detecting diseases from medical imaging: a systematic review and meta-analysis”. In: *The lancet digital health* 1.6 (2019), e271–e297.

[7] Michael Kühler. “Exploring the phenomenon and ethical issues of AI paternalism in health apps”. In: *Bioethics* 36.2 (2022), pp. 194–200.

[8] Cynthia Rudin. “Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead”. In: *Nature machine intelligence* 1.5 (2019), pp. 206–215.

[9] Rafael Capurro and Birger Hjørland. “The concept of information”. In: (2003).

[10] Claude Elwood Shannon. “A mathematical theory of communication”. In: *The Bell system technical journal* 27.3 (1948), pp. 379–423.

[11] James Gleick. *The information: A history, a theory, a flood*. Vintage, 2011.

[12] Aditya Ramesh et al. “Zero-shot text-to-image generation”. In: *International Conference on Machine Learning*. PMLR. 2021, pp. 8821–8831.

[13] *Midjourney*. https://www.midjourney.com/. Accessed on: January 22, 2024. 2022.

[14] OpenAI. *ChatGPT*. https://chat.openai.com/. Accessed on: January 22, 2024. 2023.

[15] Iqbal H Sarker. “Machine learning: Algorithms, real-world applications and research directions”. In: *SN computer science* 2.3 (2021), p. 160.

[16] Vesela Trajkoska and Gjorgji Dimeski. “Analysis and Comparison of Chess Algorithms”. In: Ss Cyril and Methodius University in Skopje, Faculty of Computer Science and . . . 2023.

[17] Arthur L Samuel. “Some studies in machine learning using the game of checkers”. In: *IBM Journal of research and development* 3.3 (1959), pp. 210–229.

[18] Manuel Castells. “The information age: Economy, society and culture (3 volumes)”. In: *Blackwell, Oxford* 1997 (1996), p. 1998.

[19] Abdu Shaalan, David Baglee, and Michael Knowles. “Are We Ready for Industry 4.0?” In: *Advances in Asset Management and Condition Monitoring: COMADEM 2019*. Springer, 2020, pp. 99–113.

[20] Tom Mitchell. “Machine Learning”. In: *Publisher: McGraw Hill* (1997).

[21] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning*. http://www.deeplearningbook.org. MIT Press, 2016.

[22] Jiawei Han, Jian Pei, and Hanghang Tong. *Data mining: concepts and techniques*. Morgan kaufmann, 2022.

[23] Richard S Sutton, Andrew G Barto, et al. “Reinforcement learning”. In: *Journal of Cognitive Neuroscience* 11.1 (1999), pp. 126–134.

[24] Leslie Pack Kaelbling, Michael L Littman, and Andrew W Moore. “Reinforcement learning: A survey”. In: *Journal of artificial intelligence research* 4 (1996), pp. 237–285.

[25] Yuxi Li. “Deep reinforcement learning: An overview”. In: *arXiv preprint arXiv:1701.07274* (2017).

[26] Leo Breiman. “Random forests”. In: *Machine learning* 45 (2001), pp. 5–32.

[27] David Pollard. “Quantization and the method of k-means”. In: *IEEE Transactions on Information theory* 28.2 (1982), pp. 199–205.

[28] Miin-Shen Yang, Chien-Yo Lai, and Chih-Ying Lin. “A robust EM clustering algorithm for Gaussian mixture models”. In: *Pattern Recognition* 45.11 (2012), pp. 3950–3961.

[29] Miryam Elizabeth Villa-Pérez et al. “Semi-supervised anomaly detection algorithms: A comparative summary and future research directions”. In: *Knowledge-Based Systems* 218 (2021), p. 106878.

[30] Maximilian E Tschuchnig and Michael Gadermayr. “Anomaly detection in medical imaging-a mini review”. In: *Data Science–Analytics and Applications: Proceedings of the 4th International Data Science Conference–iDSC2021*. Springer. 2022, pp. 33–38.

[31] Tom Dietterich. “Overfitting and undercomputing in machine learning”. In: *ACM computing surveys (CSUR)* 27.3 (1995), pp. 326–327.

[32] Douglas M Hawkins. “The problem of overfitting”. In: *Journal of chemical information and computer sciences* 44.1 (2004), pp. 1–12.

[33] Yun Xu and Royston Goodacre. “On splitting training and validation set: a comparative study of cross-validation, bootstrap and systematic sampling for estimating the generalization performance of supervised learning”. In: *Journal of analysis and testing* 2.3 (2018), pp. 249–262.

[34] Davide Anguita et al. “The’K’in K-fold Cross Validation.” In: *ESANN*. Vol. 102. 2012, pp. 441–446.

[35] James Gareth et al. *An introduction to statistical learning: with applications in R.* Spinger, 2013.

[36] Matthias Seeger. “Gaussian processes for machine learning”. In: *International journal of neural systems* 14.02 (2004), pp. 69–106.

[37] José M Bernardo and Adrian FM Smith. *Bayesian theory.* Vol. 405. John Wiley & Sons, 2009.

[38] Tristan Needham. “A visual explanation of Jensen’s inequality”. In: *The American mathematical monthly* 100.8 (1993), pp. 768–771.

[39] Solomon Kullback and Richard A Leibler. “On information and sufficiency”. In: *The annals of mathematical statistics* 22.1 (1951), pp. 79–86.

[40] Jürgen Schmidhuber. “Deep learning in neural networks: An overview”. In: *Neural networks* 61 (2015), pp. 85–117.

[41] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. “Deep learning”. In: *nature* 521.7553 (2015), pp. 436–444.

[42] David H Hubel and Torsten N Wiesel. “Receptive fields, binocular interaction and functional architecture in the cat’s visual cortex”. In: *The Journal of physiology* 160.1 (1962), p. 106.

[43] Rina Dechter. “Learning while searching in constraint-satisfaction problems”. In: (1986).

[44] Norbert Wiener. *Cybernetics or Control and Communication in the Animal and the Machine.* MIT press, 2019.

[45] Warren S McCulloch and Walter Pitts. “A logical calculus of the ideas immanent in nervous activity”. In: *The bulletin of mathematical biophysics* 5 (1943), pp. 115–133.

[46] Aleksei Grigorevich Ivakhnenko, Valentin Grigorevich Lapa, et al. “Cybernetic predicting devices”. In: (1966).

[47] Istvan SN Berkeley. “The curious case of connectionism”. In: *Open Philosophy* 2.1 (2019), pp. 190–205.

[48] Yann LeCun et al. “Gradient-based learning applied to document recognition”. In: *Proceedings of the IEEE* 86.11 (1998), pp. 2278–2324.

[49] Sepp Hochreiter and Jürgen Schmidhuber. “Long short-term memory”. In: *Neural computation* 9.8 (1997), pp. 1735–1780.

[50] Rob Farber. *CUDA application design and development.* Elsevier, 2011.

[51] Kumar Chellapilla, Sidd Puri, and Patrice Simard. “High performance convolutional neural networks for document processing”. In: *Tenth international workshop on frontiers in handwriting recognition.* Suvisoft. 2006.

[52] Kyoung-Su Oh and Keechul Jung. “GPU implementation of neural networks”. In: *Pattern Recognition* 37.6 (2004), pp. 1311–1314.

[53] Alex Lamb. “A brief introduction to generative models”. In: *arXiv preprint arXiv:2103.00265* (2021).

[54] Ashish Vaswani et al. “Attention is all you need”. In: *Advances in neural information processing systems* 30 (2017).

[55] Diederik P Kingma and Max Welling. "Auto-encoding variational bayes". In: *arXiv preprint arXiv:1312.6114* (2013).

[56] Ian Goodfellow et al. "Generative adversarial nets". In: *Advances in neural information processing systems* 27 (2014).

[57] Jonathan Ho, Ajay Jain, and Pieter Abbeel. "Denoising diffusion probabilistic models". In: *Advances in neural information processing systems* 33 (2020), pp. 6840–6851.

[58] Jia Deng et al. "Imagenet: A large-scale hierarchical image database". In: *2009 IEEE conference on computer vision and pattern recognition*. Ieee. 2009, pp. 248–255.

[59] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. "Imagenet classification with deep convolutional neural networks". In: *Advances in neural information processing systems* 25 (2012).

[60] Kunihiko Fukushima. "Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position". In: *Biological cybernetics* 36.4 (1980), pp. 193–202.

[61] Vincent Dumoulin and Francesco Visin. "A guide to convolution arithmetic for deep learning". In: *arXiv preprint arXiv:1603.07285* (2016).

[62] Ilya Sutskever et al. "On the importance of initialization and momentum in deep learning". In: *International conference on machine learning*. PMLR. 2013, pp. 1139–1147.

[63] John Duchi, Elad Hazan, and Yoram Singer. "Adaptive subgradient methods for online learning and stochastic optimization." In: *Journal of machine learning research* 12.7 (2011).

[64] Alex Graves. "Generating sequences with recurrent neural networks". In: *arXiv preprint arXiv:1308.0850* (2013).

[65] Diederik P Kingma and Jimmy Ba. "Adam: A method for stochastic optimization". In: *arXiv preprint arXiv:1412.6980* (2014).

[66] Alexey Dosovitskiy et al. "An image is worth 16x16 words: Transformers for image recognition at scale". In: *arXiv preprint arXiv:2010.11929* (2020).

[67] Geoffrey E Hinton and Ruslan R Salakhutdinov. "Reducing the dimensionality of data with neural networks". In: *science* 313.5786 (2006), pp. 504–507.

[68] Pascal Vincent et al. "Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion." In: *Journal of machine learning research* 11.12 (2010).

[69] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. "U-net: Convolutional networks for biomedical image segmentation". In: *International Conference on Medical image computing and computer-assisted intervention*. Springer. 2015, pp. 234–241.

[70] Andrew Howard et al. "Searching for mobilenetv3". In: *Proceedings of the IEEE/CVF international conference on computer vision*. 2019, pp. 1314–1324.

[71] Dušan Variš and Ondřej Bojar. "Sequence length is a domain: Length-based overfitting in transformer models". In: *arXiv preprint arXiv:2109.07276* (2021).

[72] Geoffrey E Hinton et al. “Improving neural networks by preventing co-adaptation of feature detectors”. In: *arXiv preprint arXiv:1207.0580* (2012).

[73] Nitish Srivastava et al. “Dropout: a simple way to prevent neural networks from overfitting”. In: *The journal of machine learning research* 15.1 (2014), pp. 1929–1958.

[74] Yarin Gal and Zoubin Ghahramani. “Dropout as a bayesian approximation: Representing model uncertainty in deep learning”. In: *international conference on machine learning*. PMLR. 2016, pp. 1050–1059.

[75] Herbert Bay et al. “Speeded-up robust features (SURF)”. In: *Computer vision and image understanding* 110.3 (2008), pp. 346–359.

[76] David G Lowe. “Distinctive image features from scale-invariant keypoints”. In: *International journal of computer vision* 60 (2004), pp. 91–110.

[77] Raia Hadsell, Sumit Chopra, and Yann LeCun. “Dimensionality reduction by learning an invariant mapping”. In: *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR’06)*. Vol. 2. IEEE. 2006, pp. 1735–1742.

[78] Dosovitskiy Alexey et al. “Discriminative unsupervised feature learning with exemplar convolutional neural networks”. In: *IEEE TPAMI* 38.9 (2016), pp. 1734–1747.

[79] Ashish Jaiswal et al. “A survey on contrastive self-supervised learning”. In: *Technologies* 9.1 (2020), p. 2.

[80] Ting Chen et al. “A simple framework for contrastive learning of visuals representations”. In: *International conference on machine learning*. PMLR. 2020, pp. 1597–1607.

[81] Ting Chen et al. “Big Self-Supervised Models are Strong Semi-Supervised Learners”. In: *arXiv preprint arXiv:2006.10029* (2020).

[82] Kaiming He et al. “Momentum contrast for unsupervised visual representation learning”. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2020, pp. 9729–9738.

[83] Jean-Bastien Grill et al. “Bootstrap your own latent-a new approach to self-supervised learning”. In: *Advances in neural information processing systems* 33 (2020), pp. 21271–21284.

[84] Mathilde Caron et al. “Unsupervised learning of visual features by contrasting cluster assignments”. In: *Advances in neural information processing systems* 33 (2020), pp. 9912–9924.

[85] Kihyuk Sohn. “Improved deep metric learning with multi-class n-pair loss objective”. In: *Advances in neural information processing systems* 29 (2016).

[86] Kota Dohi, Takashi Endo, and Yohei Kawaguchi. “Disentangling physical parameters for anomalous sound detection under domain shifts”. In: *2022 30th European Signal Processing Conference (EUSIPCO)*. IEEE. 2022, pp. 279–283.

[87] Clément Chadebec and Stéphanie Allassonnière. “Data augmentation with variational autoencoders and manifold sampling”. In: *Deep Generative Models, and Data Augmentation, Labelling, and Imperfections: First Workshop, DGM4MICCAI 2021, and First Workshop, DALI 2021, Held in Conjunction with MICCAI 2021, Strasbourg, France, October 1, 2021, Proceedings 1*. Springer. 2021, pp. 184–192.

[88] Moshe Leshno et al. “Multilayer feedforward networks with a nonpolynomial activation function can approximate any function”. In: *Neural networks* 6.6 (1993), pp. 861–867.

[89] Andrew R Barron. “Universal approximation bounds for superpositions of a sigmoidal function”. In: *IEEE Transactions on Information theory* 39.3 (1993), pp. 930–945.

[90] George EP Box. “Science and statistics”. In: *Journal of the American Statistical Association* 71.356 (1976), pp. 791–799.

[91] Vanessa Buhrmester, David Münch, and Michael Arens. “Analysis of explainers of black box deep neural networks for computer vision: A survey”. In: *Machine Learning and Knowledge Extraction* 3.4 (2021), pp. 966–989.

[92] Finale Doshi-Velez and Been Kim. “Towards a rigorous science of interpretable machine learning”. In: *arXiv preprint arXiv:1702.08608* (2017).

[93] Been Kim, Rajiv Khanna, and Oluwasanmi O Koyejo. “Examples are not enough, learn to criticize! criticism for interpretability”. In: *Advances in neural information processing systems* 29 (2016).

[94] Tim Miller. “Explanation in artificial intelligence: Insights from the social sciences”. In: *Artificial intelligence* 267 (2019), pp. 1–38.

[95] Christoph Molnar. *Interpretable machine learning*. Lulu. com, 2020.

[96] Ajay Thampi. *Interpretable AI: Building explainable machine learning systems*. Simon and Schuster, 2022.

[97] Naftali Tishby and Noga Zaslavsky. “Deep learning and the information bottleneck principle”. In: *2015 ieee information theory workshop (itw)*. IEEE. 2015, pp. 1–5.

[98] Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. “Deep inside convolutional networks: Visualising image classification models and saliency maps”. In: *arXiv preprint arXiv:1312.6034* (2013).

[99] Mukund Sundararajan, Ankur Taly, and Qiqi Yan. “Gradients of counterfactuals”. In: *arXiv preprint arXiv:1611.02639* (2016).

[100] Avanti Shrikumar, Peyton Greenside, and Anshul Kundaje. “Learning important features through propagating activation differences”. In: *International conference on machine learning*. PMLR. 2017, pp. 3145–3153.

[101] Pascal Sturmfels, Scott Lundberg, and Su-In Lee. “Visualizing the Impact of Feature Attribution Baselines”. In: *Distill* (2020). https://distill.pub/2020/attribution-baselines. DOI: 10.23915/distill.00022.

[102] Maria Castro-Fernandez et al. “Towards Skin Cancer Self-Monitoring through an Optimized MobileNet with Coordinate Attention”. In: *2022 25th Euromicro Conference on Digital System Design (DSD)*. IEEE. 2022, pp. 607–614.

[103] Ramprasaath R Selvaraju et al. “Grad-cam: Visual explanations from deep networks via gradient-based localization”. In: *Proceedings of the IEEE international conference on computer vision*. 2017, pp. 618–626.

[104] Pantelis Linardatos, Vasilis Papastefanopoulos, and Sotiris Kotsiantis. “Explainable ai: A review of machine learning interpretability methods”. In: *Entropy* 23.1 (2020), p. 18.

[105] Christian Szegedy et al. "Intriguing properties of neural networks". In: *arXiv preprint arXiv:1312.6199* (2013).

[106] Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. "Explaining and harnessing adversarial examples". In: *arXiv preprint arXiv:1412.6572* (2014).

[107] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. "" Why should i trust you?" Explaining the predictions of any classifier". In: *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining.* 2016, pp. 1135–1144.

[108] Damien Garreau and Ulrike Luxburg. "Explaining the explainer: A first theoretical analysis of LIME". In: *International conference on artificial intelligence and statistics.* PMLR. 2020, pp. 1287–1296.

[109] Matthew D Zeiler and Rob Fergus. "Visualizing and understanding convolutional networks". In: *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part I 13.* Springer. 2014, pp. 818–833.

[110] Per Christian Hansen. *Discrete inverse problems: insight and algorithms.* SIAM, 2010.

[111] Adel Faridani et al. "Introduction to the mathematics of computed tomography". In: *Inside Out: Inverse Problems and Applications* 47.1-46 (2003), p. 12.

[112] Hákon Gudbjartsson and Samuel Patz. "The Rician distribution of noisy MRI data". In: *Magnetic resonance in medicine* 34.6 (1995), pp. 910–914.

[113] Saeed V Vaseghi and PJW Rayner. "Detection and suppression of impulsive noise in speech communication systems". In: *IEE Proceedings I (Communications, Speech and Vision)* 137.1 (1990), pp. 38–46.

[114] Alenrex Maity et al. "A comparative study on approaches to speckle noise reduction in images". In: *2015 International Conference on Computational Intelligence and Networks.* IEEE. 2015, pp. 148–155.

[115] H Mesgarani and P Parmour. "Application of numerical solution of linear Fredholm integral equation of the first kind for image restoration". In: *Mathematical Sciences* 17.4 (2023), pp. 371–378.

[116] Albert Tarantola. *Inverse problem theory and methods for model parameter estimation.* SIAM, 2005.

[117] Dmitry Ulyanov, Andrea Vedaldi, and Victor Lempitsky. "Deep image prior". In: *Proceedings of the IEEE conference on computer vision and pattern recognition.* 2018, pp. 9446–9454.

[118] Jacques Hadamard. "Sur les problèmes aux dérivées partielles et leur signification physique". In: *Princeton university bulletin* (1902), pp. 49–52.

[119] Andrei Nikolaevich Tikhonov. "On the solution of ill-posed problems and the method of regularization". In: *Doklady akademii nauk.* Vol. 151. 3. Russian Academy of Sciences. 1963, pp. 501–504.

[120] Jianyi Lin. "Sparse Models for Machine Learning". In: *Engineering Mathematics and Artificial Intelligence.* CRC Press, pp. 107–146.

[121] Jacob John. "Discrete cosine transform in JPEG compression". In: *arXiv preprint arXiv:2102.06968* (2021).

[122] Jing-jing Zong and Tian-shuang Qiu. “Medical image fusion based on sparse representation of classified image patches”. In: *Biomedical Signal Processing and Control* 34 (2017), pp. 195–205.

[123] Peg Shippert. “Introduction to hyperspectral image analysis”. In: *Online Journal of Space Communication* 2.3 (2003), p. 8.

[124] Alexander FH Goetz. “Three decades of hyperspectral remote sensing of the Earth: A personal view”. In: *Remote sensing of environment* 113 (2009), S5–S16.

[125] V Sowmya, KP Soman, and M Hassaballah. “Hyperspectral image: Fundamentals and advances”. In: *Recent Advances in Computer Vision: Theories and Applications* (2019), pp. 401–424.

[126] Pedram Ghamisi et al. “Advances in hyperspectral image and signal processing: A comprehensive overview of the state of the art”. In: *IEEE Geoscience and Remote Sensing Magazine* 5.4 (2017), pp. 37–78.

[127] Samuel Ortega Sarmiento. “Automatic classification of histological hyperspectral images: algorithms and instrumentation”. PhD thesis. Universidad de Las Palmas de Gran Canaria, 2021.

[128] PK Garg. “Effect of contamination and adjacency factors on snow using spectroradiometer and hyperspectral images”. In: *Hyperspectral remote sensing*. Elsevier, 2020, pp. 167–196.

[129] Himar Antonio Fabelo Gómez. “Contributions to the design and implementation of algorithms for the classification of hyperspectral images of brain tumors in real-time during surgical procedures”. PhD thesis. Universidad de Las Palmas de Gran Canaria, 2019.

[130] Yantao Wei et al. “Multiscale principle of relevant information for hyperspectral image classification”. In: *Machine Learning* (2021), pp. 1–26.

[131] David Landgrebe. “Hyperspectral image data analysis”. In: *IEEE Signal processing magazine* 19.1 (2002), pp. 17–28.

[132] Mary B Stuart et al. “High-resolution hyperspectral imaging using low-cost components: Application within environmental monitoring scenarios”. In: *Sensors* 22.12 (2022), p. 4652.

[133] Shen-En Qian. “Hyperspectral satellites, evolution, and development history”. In: *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing* 14 (2021), pp. 7032–7056.

[134] Sima Peyghambari and Yun Zhang. “Hyperspectral remote sensing in lithological mapping, mineral exploration, and environmental geology: an updated review”. In: *Journal of Applied Remote Sensing* 15.3 (2021), pp. 031501–031501.

[135] Baowei Fei. “Hyperspectral imaging in medical applications”. In: *Data Handling in Science and Technology*. Vol. 32. Elsevier, 2019, pp. 523–565.

[136] Kristiane de Cássia Mariotti, Rafael Scorsatto Ortiz, and Marco Flôres Ferrão. “Hyperspectral imaging in forensic science: an overview of major application areas”. In: *Science & Justice* (2023).

[137] Wenyang Jia et al. “Hyperspectral Imaging (HSI) for meat quality evaluation across the supply chain: Current and future trends”. In: *Current Research in Food Science* (2022).

[138] Yan Zheng et al. “A discrimination model in waste plastics sorting using NIR hyperspectral imaging system”. In: *Waste Management* 72 (2018), pp. 87–98.

[139] Bartosz Grabowski et al. “Automatic pigment identification from hyperspectral data”. In: *Journal of Cultural Heritage* 31 (2018), pp. 1–12.

[140] Steven L Jacques. “Optical properties of biological tissues: a review”. In: *Physics in Medicine & Biology* 58.11 (2013), R37.

[141] PRGDJ Graves and D Gardiner. “Practical raman spectroscopy”. In: *Springer* 10 (1989), pp. 978–3.

[142] Sonia Raquel Leon Martin et al. “SWIR Hyperspectral Imaging to Assess Neurocognitive Disorders Using Blood Plasma Samples”. In: (2021).

[143] Jason G Dwight et al. “Hyperspectral image mapping spectrometry for retinal oximetry measurements in four diseased eyes”. In: *International ophthalmology clinics* 56.4 (2016), pp. 25–38.

[144] Martin Halicek et al. “In-vivo and ex-vivo tissue analysis through hyperspectral imaging techniques: revealing the invisible features of cancer”. In: *Cancers* 11.6 (2019), p. 756.

[145] Daiki Sato et al. “Distinction of surgically resected gastrointestinal stromal tumor by near-infrared hyperspectral imaging”. In: *scientific reports* 10.1 (2020), p. 21852.

[146] Daniel E Johnson et al. “Head and neck squamous cell carcinoma”. In: *Nature reviews Disease primers* 6.1 (2020), p. 92.

[147] Marco La Salvia et al. “AI-based segmentation of intraoperative glioblastoma hyperspectral images”. In: *Hyperspectral Imaging and Applications II*. Vol. 12338. SPIE. 2023, pp. 67–76.

[148] Raquel Leon et al. “Hyperspectral imaging benchmark based on machine learning for intra-operative brain tumour detection”. In: *NPJ Precision Oncology* 7.1 (2023), p. 119.

[149] Karen Simonyan and Andrew Zisserman. “Very deep convolutional networks for large-scale image recognition”. In: *arXiv preprint arXiv:1409.1556* (2014).

[150] Roberto Togneri and JS Christopher. *Fundamentals of information theory and coding design*. CRC Press, 2003.

[151] Jose C Principe. *Information theoretic learning: Renyi’s entropy and kernel perspectives*. Springer Science & Business Media, 2010.

[152] Norbert Wiener. *The human use of human beings: Cybernetics and society*. 320. Da capo press, 1988.

[153] Ralph VL Hartley. “Transmission of information 1”. In: *Bell System technical journal* 7.3 (1928), pp. 535–563.

[154] Olivier Rioul and José Carlos Magossi. “Shannon’s formula and Hartley’s rule: A mathematical coincidence?” In: *AIP Conference Proceedings*. Vol. 1641. 1. American Institute of Physics. 2015, pp. 105–112.

[155] Arieh Ben-Naim. *Entropy Demystified: The Second Law Reduced To Plain Common Sense (Revised Edition)*. World Scientific, 2008.

[156] Harold Jeffreys. *The theory of probability*. OuP Oxford, 1998.

[157] Thomas M Cover. *Elements of information theory.* John Wiley & Sons, 1999.

[158] Jianhua Lin. “Divergence measures based on the Shannon entropy”. In: *IEEE Transactions on Information theory* 37.1 (1991), pp. 145–151.

[159] Imre Csiszár and János Körner. *Information theory: coding theorems for discrete memoryless systems.* Cambridge University Press, 2011.

[160] Naftali Tishby, Fernando C Pereira, and William Bialek. “The information bottleneck method”. In: *arXiv preprint physics/0004057* (2000).

[161] Claude E Shannon et al. “Coding theorems for a discrete source with a fidelity criterion”. In: *IRE Nat. Conv. Rec* 4.142-163 (1959), p. 1.

[162] Yochai Blau and Tomer Michaeli. “Rethinking lossy compression: The rate-distortion-perception tradeoff”. In: *International Conference on Machine Learning.* PMLR. 2019, pp. 675–685.

[163] Zhou Wang et al. “Image quality assessment: from error visibility to structural similarity”. In: *IEEE transactions on image processing* 13.4 (2004), pp. 600–612.

[164] Zhou Wang and Alan C Bovik. “Mean squared error: Love it or leave it? A new look at signal fidelity measures”. In: *IEEE signal processing magazine* 26.1 (2009), pp. 98–117.

[165] Peter Harremoës and Naftali Tishby. “The information bottleneck revisited or how to choose a good distortion measure”. In: *2007 IEEE International Symposium on Information Theory.* IEEE. 2007, pp. 566–570.

[166] Ralph Linsker. “Self-organization in a perceptual network”. In: *Computer* 21.3 (1988), pp. 105–117.

[167] Anthony J Bell and Terrence J Sejnowski. “An information-maximization approach to blind separation and blind deconvolution”. In: *Neural computation* 7.6 (1995), pp. 1129–1159.

[168] Gustavus J Simmons. “A survey of information authentication”. In: *Proceedings of the IEEE* 76.5 (1988), pp. 603–620.

[169] Kristoffer Wickstrøm et al. “Information plane analysis of deep neural networks via matrix-based Renyi’s entropy and tensor kernels”. In: *arXiv preprint arXiv:1909.11396* (2019).

[170] Emanuel Parzen. “On estimation of a probability density function and mode”. In: *The annals of mathematical statistics* 33.3 (1962), pp. 1065–1076.

[171] Luis Gonzalo Sanchez Giraldo, Murali Rao, and Jose C Principe. “Measures of entropy from data using infinitely divisible kernels”. In: *IEEE Transactions on Information Theory* 61.1 (2014), pp. 535–548.

[172] Abian Hernandez-Guedes et al. “Performance Evaluation of Deep Learning Models for Image Classification Over Small Datasets: Diabetic Foot Case Study”. In: *IEEE Access* 10 (2022), pp. 124373–124386.

[173] Corinna Cortes and Vladimir Vapnik. “Support-vector networks”. In: *Machine learning* 20 (1995), pp. 273–297.

[174] John Shawe-Taylor and Nello Cristianini. *Kernel methods for pattern analysis.* Cambridge university press, 2004.

[175] Nicolás I Tapia and Pablo A Estévez. "On the information plane of autoencoders". In: *2020 International Joint Conference on Neural Networks (IJCNN)*. IEEE. 2020, pp. 1–8.

[176] Bernard W Silverman. *Density estimation for statistics and data analysis*. Routledge, 2018.

[177] Nello Cristianini et al. "On kernel-target alignment". In: *Advances in neural information processing systems* 14 (2001).

[178] Jack Tulloch, Reza Zamani, and Mohammad Akrami. "Machine learning in the prevention, diagnosis and management of diabetic foot ulcers: a systematic review". In: *IEEE Access* 8 (2020), pp. 198977–199000.

[179] Effy Vayena, Alessandro Blasimme, and I Glenn Cohen. "Machine learning in medicine: addressing ethical challenges". In: *PLoS medicine* 15.11 (2018), e1002689.

[180] S Kevin Zhou et al. "A review of deep learning in medical imaging: Imaging traits, technology trends, case studies with progress highlights, and future promises". In: *Proceedings of the IEEE* (2021).

[181] David Leslie et al. "Does "AI" stand for augmenting inequality in the era of covid-19 healthcare?" In: *bmj* 372 (2021).

[182] Yaniv Bar et al. "Chest pathology detection using deep learning with non-medical training". In: *2015 IEEE 12th international symposium on biomedical imaging (ISBI)*. IEEE. 2015, pp. 294–297.

[183] Connor Shorten and Taghi M Khoshgoftaar. "A survey on image data augmentation for deep learning". In: *Journal of Big Data* 6.1 (2019), pp. 1–48.

[184] Ju Xu, Mengzhang Li, and Zhanxing Zhu. "Automatic data augmentation for 3D medical image segmentation". In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer. 2020, pp. 378–387.

[185] Taranjit Kaur and Tapan Kumar Gandhi. "Automated brain image classification based on VGG-16 and transfer learning". In: *2019 International Conference on Information Technology (ICIT)*. IEEE. 2019, pp. 94–98.

[186] Yujin Oh, Sangjoon Park, and Jong Chul Ye. "Deep learning covid-19 features on cxr using limited training data sets". In: *IEEE transactions on medical imaging* 39.8 (2020), pp. 2688–2700.

[187] Debesh Jha et al. "Doubleu-net: A deep convolutional neural network for medical image segmentation". In: *2020 IEEE 33rd International symposium on computer-based medical systems (CBMS)*. IEEE. 2020, pp. 558–564.

[188] Razvan Pascanu, Tomas Mikolov, and Yoshua Bengio. "On the difficulty of training recurrent neural networks". In: *International conference on machine learning*. PMLR. 2013, pp. 1310–1318.

[189] A Emin Orhan and Xaq Pitkow. "Skip connections eliminate singularities". In: *arXiv preprint arXiv:1701.09175* (2017).

[190] Shujian Yu and Jose C Principe. "Understanding autoencoders with information theoretic concepts". In: *Neural Networks* 117 (2019), pp. 104–123.

[191] Ravid Shwartz-Ziv and Naftali Tishby. "Opening the black box of deep neural networks via information". In: *arXiv preprint arXiv:1703.00810* (2017).

[192] Andrew M Saxe et al. "On the information bottleneck theory of deep learning". In: *Journal of Statistical Mechanics: Theory and Experiment* 2019.12 (2019), p. 124020.

[193] Dianna J Magliano, Edward J Boyko, et al. "IDF diabetes atlas". In: (2022).

[194] Lawrence A Lavery et al. "Preventing diabetic foot ulcer recurrence in high-risk patients: use of temperature monitoring as a self-assessment tool". In: *Diabetes care* 30.1 (2007), pp. 14–20.

[195] Subramaniam Bagavathiappan et al. "Infrared thermal imaging for detection of peripheral vascular disorders". In: *Journal of medical physics/Association of Medical Physicists of India* 34.1 (2009), p. 43.

[196] Enrique Villa, Natalia Arteaga-Marrero, and Juan Ruiz-Alzola. "Performance assessment of low-cost thermal cameras for medical applications". In: *Sensors* 20.5 (2020), p. 1321.

[197] Hayde Peregrina-Barreto et al. "Quantitative estimation of temperature variations in plantar angiosomes: a study case for diabetic foot". In: *Computational and mathematical methods in medicine* 2014 (2014).

[198] D Hernandez-Contreras et al. "Narrative review: Diabetic foot and infrared thermography". In: *Infrared Physics & Technology* 78 (2016), pp. 105–117.

[199] Chanjuan Liu et al. "Automatic detection of diabetic foot complications with infrared thermography by asymmetric analysis". In: *Journal of biomedical optics* 20.2 (2015), pp. 026003–026003.

[200] Kor H Hutting et al. "Infrared thermography for monitoring severity and treatment of diabetic foot infections". In: *Vascular Biology* 2.1 (2020), pp. 1–10.

[201] Abián Hernández et al. "Automatic segmentation based on deep learning techniques for diabetic foot monitoring through multimodal images". In: *International conference on image analysis and processing*. Springer. 2019, pp. 414–424.

[202] Natalia Arteaga-Marrero et al. "Segmentation approaches for diabetic foot disorders". In: *Sensors* 21.3 (2021), p. 934.

[203] Natalia Arteaga-Marrero et al. "Morphological Foot Model for Temperature Pattern Analysis Proposed for Diabetic Foot Disorders". In: *Applied Sciences* 11.16 (2021), p. 7396.

[204] Jaap J van Netten et al. "Diagnostic values for skin temperature assessment to detect diabetes-related foot complications". In: *Diabetes technology & therapeutics* 16.11 (2014), pp. 714–721.

[205] Lawrence A Lavery et al. "Home monitoring of foot skin temperatures to prevent ulceration". In: *Diabetes care* 27.11 (2004), pp. 2642–2647.

[206] Daniel Alejandro Hernandez-Contreras et al. "Plantar thermogram database for the study of diabetic foot complications". In: *IEEE Access* 7 (2019), pp. 161296–161307.

[207] Mohammed Elmogy et al. "Tissues classification for pressure ulcer images based on 3D convolutional neural network". In: *2018 25th IEEE International Conference on Image Processing (ICIP)*. IEEE. 2018, pp. 3139–3143.

[208] Gustavo Blanco et al. "A superpixel-driven deep learning approach for the analysis of dermatological wounds". In: *Computer methods and programs in biomedicine* 183 (2020), p. 105079.

[209] Moi Hoon Yap et al. "Deep learning in diabetic foot ulcers detection: a comprehensive evaluation". In: *Computers in Biology and Medicine* 135 (2021), p. 104596.

[210] Sujit Kumar Das, Pinki Roy, and Arnab Kumar Mishra. "Recognition of ischaemia and infection in diabetic foot ulcer: a deep convolutional neural network based approach". In: *International Journal of Imaging Systems and Technology* 32.1 (2022), pp. 192–208.

[211] Bill Cassidy et al. "The DFUC 2020 dataset: Analysis towards diabetic foot ulcer detection". In: *touchREVIEWS in Endocrinology* 17.1 (2021), p. 5.

[212] Moi Hoon Yap et al. "Analysis towards classification of infection and ischaemia of diabetic foot ulcers". In: *2021 IEEE EMBS International Conference on Biomedical and Health Informatics (BHI)*. IEEE. 2021, pp. 1–4.

[213] M Tan, R Pang, and QV Le. "EfficientDet: scalable and efficient object detection. arXiv". In: *arXiv preprint arXiv:1911.09070* 10 (2019).

[214] Laith Alzubaidi et al. "DFU_QUTNet: diabetic foot ulcer classification using novel deep convolutional neural network". In: *Multimedia Tools and Applications* 79.21 (2020), pp. 15655–15677.

[215] Laith Alzubaidi et al. "Robust application of new deep learning tools: an experimental study in medical imaging". In: *Multimedia Tools and Applications* 81.10 (2022), pp. 13289–13317.

[216] Israel Cruz-Vega et al. "Deep learning classification for diabetic foot thermograms". In: *Sensors* 20.6 (2020), p. 1762.

[217] Amith Khandakar et al. "A machine learning model for early detection of diabetic foot using thermogram images". In: *Computers in biology and medicine* 137 (2021), p. 104838.

[218] Andrés Anaya-Isaza and Martha Zequera-Diaz. "Fourier transform-based data augmentation in deep learning for diabetic foot thermograph classification". In: *Biocybernetics and Biomedical Engineering* 42.2 (2022), pp. 437–452.

[219] Rafael C Gonzalez, Richard E Woods, and Barry R Masters. *Digital image processing*. 2009.

[220] Kaiming He et al. "Deep residual learning for image recognition". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016, pp. 770–778.

[221] Stevo Bozinovski. "Reminder of the first paper on transfer learning in neural networks, 1976". In: *Informatica* 44.3 (2020).

[222] Laith Alzubaidi et al. "Towards a better understanding of transfer learning for medical imaging: a case study". In: *Applied Sciences* 10.13 (2020), p. 4523.

[223] Han Xiao, Kashif Rasul, and Roland Vollgraf. "Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms". In: *arXiv preprint arXiv:1708.07747* (2017).

[224] Sergey Ioffe and Christian Szegedy. "Batch normalization: Accelerating deep network training by reducing internal covariate shift". In: *International conference on machine learning*. pmlr. 2015, pp. 448–456.

[225] Zeyang Dou et al. “Band selection of hyperspectral images using attention-based autoencoders”. In: *IEEE Geoscience and Remote Sensing Letters* 18.1 (2020), pp. 147–151.

[226] Abian Hernandez-Guedes et al. “Feature Ranking by Variational Dropout for Classification Using Thermograms from Diabetic Foot Ulcers”. In: *Sensors* 23.2 (2023), p. 757.

[227] Natalia Arteaga-Marrero et al. “State-of-the-Art Features for Early-Stage Detection of Diabetic Foot Ulcers Based on Thermograms”. In: *Biomedicines* 11.12 (2023), p. 3209.

[228] RJNJ Bellman. “Dynamic programming princeton university press princeton”. In: *New Jersey Google Scholar* (1957), pp. 24–73.

[229] Isabelle Guyon and André Elisseeff. “An introduction to variable and feature selection”. In: *Journal of machine learning research* 3.Mar (2003), pp. 1157–1182.

[230] Michael E Tipping and Christopher M Bishop. “Mixtures of probabilistic principal component analyzers”. In: *Neural computation* 11.2 (1999), pp. 443–482.

[231] Visar Berisha et al. “Digital medicine and the curse of dimensionality”. In: *NPJ digital medicine* 4.1 (2021), p. 153.

[232] Beatriz Remeseiro and Veronica Bolon-Canedo. “A review of feature selection methods in medical applications”. In: *Computers in biology and medicine* 112 (2019), p. 103375.

[233] Early Treatment for Retinopathy of Prematurity Cooperative Group et al. “Revised indications for the treatment of retinopathy of prematurity: results of the early treatment for retinopathy of prematurity randomized trial”. In: *Archives of Ophthalmology (Chicago, Ill.: 1960)* 121.12 (2003), pp. 1684–1694.

[234] Verónica Bolón-Canedo et al. “Dealing with inter-expert variability in retinopathy of prematurity: a machine learning approach”. In: *Computer methods and programs in biomedicine* 122.1 (2015), pp. 1–15.

[235] Kenneth E Hild et al. “Feature extraction using information-theoretic learning”. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 28.9 (2006), pp. 1385–1392.

[236] Ryan J Urbanowicz et al. “Relief-based feature selection: Introduction and review”. In: *Journal of biomedical informatics* 85 (2018), pp. 189–203.

[237] Mario Beraha et al. “Feature selection via mutual information: New theoretical insights”. In: *2019 international joint conference on neural networks (IJCNN)*. IEEE. 2019, pp. 1–9.

[238] Isabelle Guyon et al. “Gene selection for cancer classification using support vector machines”. In: *Machine learning* 46 (2002), pp. 389–422.

[239] John H Holland. *Adaptation in natural and artificial systems: an introductory analysis with applications to biology, control, and artificial intelligence*. MIT press, 1992.

[240] Aleem Akhtar. “Evolution of Ant Colony Optimization Algorithm–A Brief Literature Review”. In: *arXiv preprint arXiv:1908.08007* (2019).

[241] Robert Tibshirani. “Regression shrinkage and selection via the lasso”. In: *Journal of the Royal Statistical Society Series B: Statistical Methodology* 58.1 (1996), pp. 267–288.

[242] Fadil Santosa and William W Symes. “Linear inversion of band-limited reflection seismograms”. In: *SIAM journal on scientific and statistical computing* 7.4 (1986), pp. 1307–1330.

[243] Hanchuan Peng, Fuhui Long, and Chris Ding. "Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy". In: *IEEE Transactions on pattern analysis and machine intelligence* 27.8 (2005), pp. 1226–1238.

[244] Makoto Yamada et al. "High-dimensional feature selection by feature-wise kernelized lasso". In: *Neural computation* 26.1 (2014), pp. 185–207.

[245] Jean Feng and Noah Simon. "Sparse-input neural networks for high-dimensional nonparametric regression and classification". In: *arXiv preprint arXiv:1711.07592* (2017).

[246] Ismael Lemhadri et al. "Lassonet: A neural network with feature sparsity". In: *The Journal of Machine Learning Research* 22.1 (2021), pp. 5633–5661.

[247] Balas Kausik Natarajan. "Sparse approximate solutions to linear systems". In: *SIAM journal on computing* 24.2 (1995), pp. 227–234.

[248] Christos Louizos, Max Welling, and Diederik P Kingma. "Learning sparse neural networks through L_0 regularization". In: *arXiv preprint arXiv:1712.01312* (2017).

[249] Chun-Hao Chang, Ladislav Rampasek, and Anna Goldenberg. "Dropout feature ranking for deep learning models". In: *arXiv preprint arXiv:1712.08645* (2017).

[250] Chris J Maddison, Andriy Mnih, and Yee Whye Teh. "The concrete distribution: A continuous relaxation of discrete random variables". In: *arXiv preprint arXiv:1611.00712* (2016).

[251] Yarin Gal, Jiri Hron, and Alex Kendall. "Concrete dropout". In: *Advances in neural information processing systems* 30 (2017).

[252] Gertraud Malsiner-Walli and Helga Wagner. "Comparing spike and slab priors for Bayesian variable selection". In: *arXiv preprint arXiv:1812.07259* (2018).

[253] Yutaro Yamada et al. "Feature selection using stochastic gates". In: *International Conference on Machine Learning*. PMLR. 2020, pp. 10648–10659.

[254] Durk P Kingma, Tim Salimans, and Max Welling. "Variational dropout and the local reparameterization trick". In: *Advances in neural information processing systems* 28 (2015).

[255] Jorge R Vergara and Pablo A Estévez. "A review of feature selection methods based on mutual information". In: *Neural computing and applications* 24 (2014), pp. 175–186.

[256] Dahua Lin and Xiaoou Tang. "Conditional infomax learning: An integrated framework for feature extraction and fusion". In: *Computer Vision–ECCV 2006: 9th European Conference on Computer Vision, Graz, Austria, May 7-13, 2006. Proceedings, Part I 9*. Springer. 2006, pp. 68–82.

[257] François Fleuret. "Fast binary feature selection with conditional mutual information." In: *Journal of Machine learning research* 5.9 (2004).

[258] Casper Kaae Sønderby et al. "How to train deep variational autoencoders and probabilistic ladder networks". In: *arXiv preprint arXiv:1602.02282* 3.2 (2016).

[259] Danfeng Hong et al. "Interpretable hyperspectral artificial intelligence: When nonconvex modeling meets hyperspectral remote sensing". In: *IEEE Geoscience and Remote Sensing Magazine* 9.2 (2021), pp. 52–87.

[260] Beatriz Martinez et al. “Most relevant spectral bands identification for brain cancer detection using hyperspectral imaging”. In: *Sensors* 19.24 (2019), p. 5481.

[261] Ilya Loshchilov and Frank Hutter. “Sgdr: Stochastic gradient descent with warm restarts”. In: *arXiv preprint arXiv:1608.03983* (2016).

[262] Alexander Quinn Nichol and Prafulla Dhariwal. “Improved denoising diffusion probabilistic models”. In: *International Conference on Machine Learning*. PMLR. 2021, pp. 8162–8171.

[263] Toshihiro Takamatsu et al. “Development of a visible to 1600 nm hyperspectral imaging rigid-scope system using supercontinuum light and an acousto-optic tunable filter”. In: *Optics Express* 32.9 (2024), pp. 16090–16102.

[264] Ibrahim Alkatout et al. “The development of laparoscopy—a historical overview”. In: *Frontiers in surgery* 8 (2021), p. 799442.

[265] Inderjeet Mani and I Zhang. “kNN approach to unbalanced data distributions: a case study involving information extraction”. In: *Proceedings of workshop on learning from imbalanced datasets*. Vol. 126. 1. ICML. 2003, pp. 1–7.

[266] Jiaojiao Li et al. “Hyperspectral image classification with imbalanced data based on orthogonal complement subspace projection”. In: *IEEE Transactions on Geoscience and Remote Sensing* 56.7 (2018), pp. 3838–3851.

[267] Kevin R Moon et al. “Visualizing structure and transitions in high-dimensional biological data”. In: *Nature biotechnology* 37.12 (2019), pp. 1482–1492.

[268] Robert H Wilson et al. “Review of short-wave infrared spectroscopy and imaging methods for biological tissue characterization”. In: *Journal of biomedical optics* 20.3 (2015), pp. 030901–030901.

[269] Qian Cao et al. “Multispectral imaging in the extended near-infrared window based on endogenous chromophores”. In: *Journal of biomedical optics* 18.10 (2013), pp. 101318–101318.

[270] Rami Nachabe et al. “Estimation of lipid and water concentrations in scattering media with diffuse optical spectroscopy from 900 to 1600 nm”. In: *Journal of biomedical optics* 15.3 (2010), pp. 037015–037015.

[271] Peiwen Chen, Matilde Cescon, and Paolo Bonaldo. “The role of collagens in peripheral nerve myelination and function”. In: *Molecular neurobiology* 52 (2015), pp. 216–225.

[272] Xiaofang Liu and Chun Yang. “A kernel spectral angle mapper algorithm for remote sensing image classification”. In: *2013 6th International Congress on Image and Signal Processing (CISP)*. Vol. 2. IEEE. 2013, pp. 814–818.

[273] Po-Ching Chen and Wei-Chiang Lin. “Spectral-profile-based algorithm for hemoglobin oxygen saturation determination from diffuse reflectance spectra”. In: *Biomedical optics express* 2.5 (2011), pp. 1082–1096.

[274] Behnood Rasti et al. “Noise reduction in hyperspectral imagery: Overview and application”. In: *Remote Sensing* 10.3 (2018), p. 482.

[275] José M Bioucas-Dias et al. “Hyperspectral unmixing overview: Geometrical, statistical, and sparse regression-based approaches”. In: *IEEE journal of selected topics in applied earth observations and remote sensing* 5.2 (2012), pp. 354–379.

[276] Ines A Cruz-Guerrero et al. "Classification of hyperspectral in vivo brain tissue based on linear unmixing". In: *Applied Sciences* 10.16 (2020), p. 5686.

[277] Nirmal Keshava and John F Mustard. "Spectral unmixing". In: *IEEE signal processing magazine* 19.1 (2002), pp. 44–57.

[278] Daniel U Campos-Delgado et al. "Nonlinear extended blind end-member and abundance extraction for hyperspectral images". In: *Signal Processing* 201 (2022), p. 108718.

[279] Fred A Kruse et al. "The spectral image processing system (SIPS)—interactive visualization and analysis of imaging spectrometer data". In: *Remote sensing of environment* 44.2-3 (1993), pp. 145–163.

[280] Freek Van der Meer. "The effectiveness of spectral similarity measures for the analysis of hyperspectral imagery". In: *International journal of applied earth observation and geoinformation* 8.1 (2006), pp. 3–17.

[281] Himar Fabelo et al. "Evaluating the use of Hyperspectral Imaging as Complementary Blood Sample Tests". In: *2021 XXXVI Conference on Design of Circuits and Integrated Systems (DCIS)*. IEEE. 2021, pp. 1–6.

[282] Alina Zare and KC Ho. "Endmember variability in hyperspectral analysis: Addressing spectral variability during spectral unmixing". In: *IEEE Signal Processing Magazine* 31.1 (2013), pp. 95–104.

[283] Ben Somers et al. "Endmember variability in spectral mixture analysis: A review". In: *Remote Sensing of Environment* 115.7 (2011), pp. 1603–1616.

[284] Lucas Drumetz et al. "Blind hyperspectral unmixing using an extended linear mixing model to address spectral variability". In: *IEEE Transactions on Image Processing* 25.8 (2016), pp. 3890–3905.

[285] Ricardo Augusto Borsoi et al. "Spectral variability in hyperspectral data unmixing: A comprehensive review". In: *IEEE geoscience and remote sensing magazine* 9.4 (2021), pp. 223–270.

[286] Tales Imbiriba, Ricardo Augusto Borsoi, and José Carlos Moreira Bermudez. "Generalized linear mixing model accounting for endmember variability". In: *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE. 2018, pp. 1862–1866.

[287] Nicolas Dobigeon et al. "Nonlinear unmixing of hyperspectral images: Models and algorithms". In: *IEEE Signal processing magazine* 31.1 (2013), pp. 82–94.

[288] Wenyi Fan et al. "Comparative study between a new nonlinear model and common linear model for analysing laboratory simulated-forest hyperspectral data". In: *International Journal of Remote Sensing* 30.11 (2009), pp. 2951–2962.

[289] Nareenart Raksuntorn and Qian Du. "Nonlinear spectral mixture analysis for hyperspectral imagery in an unknown environment". In: *IEEE Geoscience and Remote Sensing Letters* 7.4 (2010), pp. 836–840.

[290] Ben Somers et al. "Nonlinear hyperspectral mixture analysis for tree cover estimates in orchards". In: *Remote Sensing of Environment* 113.6 (2009), pp. 1183–1193.

[291] Elizabeth Baby George et al. “Assessment of spectral reduction techniques for endmember extraction in unmixing of hyperspectral images”. In: *Advances in Space Research* (2022).

[292] Antonio Plaza et al. “Spatial/spectral endmember extraction by multidimensional morphological operations”. In: *IEEE transactions on geoscience and remote sensing* 40.9 (2002), pp. 2025–2041.

[293] Derek M Rogge et al. “Integration of spatial–spectral information for the improved extraction of endmembers”. In: *Remote Sensing of Environment* 110.3 (2007), pp. 287–303.

[294] José M Bioucas-Dias and José MP Nascimento. “Hyperspectral subspace identification”. In: *IEEE Transactions on Geoscience and Remote Sensing* 46.8 (2008), pp. 2435–2445.

[295] Hanye Pu et al. “A fully constrained linear spectral unmixing algorithm based on distance geometry”. In: *IEEE transactions on geoscience and remote sensing* 52.2 (2013), pp. 1157–1176.

[296] Joseph W Boardman. “Geometric mixture analysis of imaging spectrometry data”. In: *Proceedings of IGARSS’94-1994 IEEE International Geoscience and Remote Sensing Symposium*. Vol. 4. IEEE. 1994, pp. 2369–2371.

[297] José MP Nascimento and José MB Dias. “Vertex component analysis: A fast algorithm to unmix hyperspectral data”. In: *IEEE transactions on Geoscience and Remote Sensing* 43.4 (2005), pp. 898–910.

[298] Michael E Winter. “N-FINDR: An algorithm for fast autonomous spectral end-member determination in hyperspectral data”. In: *Imaging Spectrometry V*. Vol. 3753. SPIE. 1999, pp. 266–275.

[299] Xin-Ru Feng et al. “Hyperspectral unmixing based on nonnegative matrix factorization: A comprehensive review”. In: *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing* (2022).

[300] Daniel C Heinz et al. “Fully constrained least squares linear spectral mixture analysis method for material quantification in hyperspectral imagery”. In: *IEEE transactions on geoscience and remote sensing* 39.3 (2001), pp. 529–545.

[301] Lidan Miao and Hairong Qi. “Endmember extraction from highly mixed data using minimum volume constrained nonnegative matrix factorization”. In: *IEEE Transactions on Geoscience and Remote Sensing* 45.3 (2007), pp. 765–777.

[302] José M Bioucas-Dias. “A variable splitting augmented Lagrangian approach to linear spectral unmixing”. In: *2009 First workshop on hyperspectral image and signal processing: Evolution in remote sensing*. IEEE. 2009, pp. 1–4.

[303] Jun Li and José M Bioucas-Dias. “Minimum volume simplex analysis: A fast algorithm to unmix hyperspectral data”. In: *IGARSS 2008-2008 IEEE International Geoscience and Remote Sensing Symposium*. Vol. 3. IEEE. 2008, pp. III–250.

[304] Aziz ul Rehman and Shahzad Ahmad Qureshi. “A review of the medical hyperspectral imaging systems and unmixing algorithms’ in biological tissues”. In: *Photodiagnosis and Photodynamic Therapy* 33 (2021), p. 102165.

[305] Andrea Baraldi et al. "Comparison of the multilayer perceptron with neuro-fuzzy techniques in the estimation of cover class mixture in remotely sensed data". In: *IEEE Transactions on Geoscience and Remote Sensing* 39.5 (2001), pp. 994–1005.

[306] Giorgio A Licciardi and Fabio Del Frate. "Pixel unmixing in hyperspectral data by means of neural networks". In: *IEEE transactions on Geoscience and remote sensing* 49.11 (2011), pp. 4163–4172.

[307] Behnood Rasti et al. "UnDIP: Hyperspectral unmixing using deep image prior". In: *IEEE Transactions on Geoscience and Remote Sensing* 60 (2021), pp. 1–15.

[308] Yuanchao Su et al. "Nonnegative sparse autoencoder for robust endmember extraction from remotely sensed hyperspectral images". In: *2017 IEEE International Geoscience and Remote Sensing Symposium (IGARSS)*. IEEE. 2017, pp. 205–208.

[309] Burkni Palsson et al. "Hyperspectral unmixing using a neural network autoencoder". In: *IEEE Access* 6 (2018), pp. 25646–25656.

[310] Danfeng Hong et al. "Endmember-guided unmixing network (EGU-Net): A general deep learning framework for self-supervised hyperspectral unmixing". In: *IEEE Transactions on Neural Networks and Learning Systems* 33.11 (2021), pp. 6518–6531.

[311] Preetam Ghosh et al. "Hyperspectral unmixing using transformer network". In: *IEEE Transactions on Geoscience and Remote Sensing* 60 (2022), pp. 1–16.

[312] Yue Shi et al. "Endmember extraction using minimum volume and information constraint nonnegative matrix factorization". In: *IEEE Geoscience and Remote Sensing Letters* 16.9 (2019), pp. 1427–1431.

[313] Zhu Han et al. "Multimodal hyperspectral unmixing: Insights from attention networks". In: *IEEE Transactions on Geoscience and Remote Sensing* 60 (2022), pp. 1–13.

[314] Chein-I Chang et al. "Linear spectral random mixture analysis for hyperspectral imagery". In: *IEEE transactions on geoscience and remote sensing* 40.2 (2002), pp. 375–392.

[315] Jessica D Bayliss, J Anthony Gualtieri, and Robert F Cromp. "Analyzing hyperspectral data with independent component analysis". In: *26th AIPR workshop: Exploiting new image sources and sensors*. Vol. 3240. SPIE. 1998, pp. 133–143.

[316] Shao-Shan Chiang, Chein-I Chang, and Irvin W Ginsberg. "Unsupervised hyperspectral image analysis using independent component analysis". In: *IGARSS 2000. IEEE 2000 International Geoscience and Remote Sensing Symposium. Taking the Pulse of the Planet: The Role of Remote Sensing in Managing the Environment. Proceedings (Cat. No. 00CH37120)*. Vol. 7. IEEE. 2000, pp. 3136–3138.

[317] José MP Nascimento and Jose MB Dias. "Does independent component analysis play a role in unmixing hyperspectral data?" In: *IEEE Transactions on Geoscience and Remote Sensing* 43.1 (2005), pp. 175–187.

[318] Hongming Li, Shujian Yu, and José C Príncipe. "Deep Deterministic Independent Component Analysis for Hyperspectral Unmixing". In: *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE. 2022, pp. 3878–3882.

[319] Feiyun Zhu. “Hyperspectral unmixing: ground truth labeling, datasets, benchmark performances and survey”. In: *arXiv preprint arXiv:1708.05125* (2017).

[320] Magnus Magnusson et al. “Creating RGB images from hyperspectral images using a color matching function”. In: *IGARSS 2020-2020 IEEE International Geoscience and Remote Sensing Symposium*. IEEE. 2020, pp. 2045–2048.

[321] Bruce Hapke. “Bidirectional reflectance spectroscopy: 1. Theory”. In: *Journal of Geophysical Research: Solid Earth* 86.B4 (1981), pp. 3039–3054.

[322] J-B Féret et al. “PROSPECT-D: Towards modeling leaf optical properties through a complete lifecycle”. In: *Remote Sensing of Environment* 193 (2017), pp. 204–215.

[323] Jingfeng Xiao and Aaron Moody. “A comparison of methods for estimating fractional green vegetation cover within a desert-to-upland transition zone in central New Mexico, USA”. In: *Remote sensing of environment* 98.2-3 (2005), pp. 237–250.

[324] Himar Fabelo et al. “Novel Methodology for Alzheimer’s Disease Biomarker Identification in Plasma using Hyperspectral Microscopy”. In: *2020 XXXV Conference on Design of Circuits and Integrated Systems (DCIS)*. IEEE. 2020, pp. 1–6.

[325] Simon Henrot, Jocelyn Chanussot, and Christian Jutten. “Dynamical spectral unmixing of multitemporal hyperspectral images”. In: *IEEE Transactions on Image Processing* 25.7 (2016), pp. 3219–3232.

[326] Guillaume Tochon et al. “From local to global unmixing of hyperspectral images to reveal spectral variability”. In: *2016 8th Workshop on Hyperspectral Image and Signal Processing: Evolution in Remote Sensing (WHISPERS)*. IEEE. 2016, pp. 1–5.

[327] Feiyun Zhu et al. “Spectral unmixing via data-guided sparsity”. In: *IEEE Transactions on Image Processing* 23.12 (2014), pp. 5412–5427.

[328] Dar A Roberts et al. “Mapping chaparral in the Santa Monica Mountains using multiple endmember spectral mixture models”. In: *Remote sensing of environment* 65.3 (1998), pp. 267–279.

[329] Lucas Drumetz et al. “Hyperspectral image unmixing with endmember bundles and group sparsity inducing mixed norms”. In: *IEEE Transactions on Image Processing* 28.7 (2019), pp. 3435–3450.

[330] José M Bioucas-Dias and Mário AT Figueiredo. “Alternating direction algorithms for constrained sparse regression: Application to hyperspectral unmixing”. In: *2010 2nd Workshop on Hyperspectral Image and Signal Processing: Evolution in Remote Sensing*. IEEE. 2010, pp. 1–4.

[331] Fanuel Mehari et al. “Investigation of the differentiation of ex vivo nerve and fat tissues using laser-induced breakdown spectroscopy (LIBS): Prospects for tissue-specific laser surgery”. In: *Journal of biophotonics* 9.10 (2016), pp. 1021–1032.

[332] Douglas M Hawkins. *Identification of outliers*. Vol. 11. Springer, 1980.

[333] Durgesh Samariya et al. “Detection and explanation of anomalies in healthcare data”. In: *Health Information Science and Systems* 11.1 (2023), p. 20.

[334] Sonal Kothari et al. "Pathology imaging informatics for quantitative analysis of whole-slide images". In: *Journal of the American Medical Informatics Association* 20.6 (2013), pp. 1099–1108.

[335] Raghavendra Chalapathy and Sanjay Chawla. "Deep learning for anomaly detection: A survey". In: *arXiv preprint arXiv:1901.03407* (2019).

[336] Guansong Pang et al. "Deep learning for anomaly detection: A review". In: *ACM computing surveys (CSUR)* 54.2 (2021), pp. 1–38.

[337] Srikanth Thudumu et al. "A comprehensive survey of anomaly detection techniques for high dimensional big data". In: *Journal of Big Data* 7 (2020), pp. 1–30.

[338] Muhammad Adeel Azam et al. "A review on multimodal medical image fusion: Compendious analysis of medical modalities, multimodal databases, fusion techniques and quality metrics". In: *Computers in biology and medicine* 144 (2022), p. 105253.

[339] Ralph Foorthuis. "On the nature and types of anomalies: a review of deviations in data". In: *International journal of data science and analytics* 12.4 (2021), pp. 297–331.

[340] Varun Chandola, Arindam Banerjee, and Vipin Kumar. "Anomaly detection: A survey". In: *ACM computing surveys (CSUR)* 41.3 (2009), pp. 1–58.

[341] Francis Ysidro Edgeworth. "On discordant observations". In: *The london, edinburgh, and dublin philosophical magazine and journal of science* 23.143 (1887), pp. 364–375.

[342] Frank E Grubbs. "Procedures for detecting outlying observations in samples". In: *Technometrics* 11.1 (1969), pp. 1–21.

[343] Cecilia Surace and K Worden. "A novelty detection method to diagnose damage in structures: an application to an offshore platform". In: *ISOPE International Ocean and Polar Engineering Conference*. ISOPE. 1998, ISOPE–I.

[344] Bovas Abraham and George EP Box. "Bayesian analysis of some outlier problems in time series". In: *Biometrika* 66.2 (1979), pp. 229–236.

[345] Deepak Agarwal. "Detecting anomalies in cross-classified streams: a bayesian approach". In: *Knowledge and information systems* 11.1 (2007), pp. 29–44.

[346] Dipankar Dasgupta and Fernando Nino. "A comparison of negative and positive selection algorithms in novel pattern detection". In: *Smc 2000 conference proceedings. 2000 ieee international conference on systems, man and cybernetics.'cybernetics evolving to systems, humans, organizations, and their complex interactions'(cat. no. 0.* Vol. 1. IEEE. 2000, pp. 125–130.

[347] Ali Bou Nassif et al. "Machine learning for anomaly detection: A systematic review". In: *Ieee Access* 9 (2021), pp. 78658–78700.

[348] Martin Ester et al. "A density-based algorithm for discovering clusters in large spatial databases with noise". In: *kdd*. Vol. 96. 34. 1996, pp. 226–231.

[349] Bernhard Schölkopf et al. "Estimating the support of a high-dimensional distribution". In: *Neural computation* 13.7 (2001), pp. 1443–1471.

[350] David MJ Tax and Robert PW Duin. "Support vector data description". In: *Machine learning* 54 (2004), pp. 45–66.

[351] Fei Tony Liu, Kai Ming Ting, and Zhi-Hua Zhou. “Isolation forest”. In: *2008 eighth ieee international conference on data mining*. IEEE. 2008, pp. 413–422.

[352] Simon Hawkins et al. “Outlier detection using replicator neural networks”. In: *International Conference on Data Warehousing and Knowledge Discovery*. Springer. 2002, pp. 170–180.

[353] Robert Hecht-Nielsen. “Replicator neural networks for universal optimal source coding”. In: *Science* 269.5232 (1995), pp. 1860–1863.

[354] Laura Beggel, Michael Pfeiffer, and Bernd Bischl. “Robust anomaly detection in images using adversarial autoencoders”. In: *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*. Springer. 2019, pp. 206–222.

[355] Yuanyuan Wei et al. “LSTM-autoencoder-based anomaly detection for indoor air quality time-series data”. In: *IEEE Sensors Journal* 23.4 (2023), pp. 3787–3800.

[356] Mahmoud Said Elsayed et al. “Network anomaly detection using LSTM based autoencoder”. In: *Proceedings of the 16th ACM Symposium on QoS and Security for Wireless and Mobile Networks*. 2020, pp. 37–45.

[357] Pengfei Liu et al. “Arrhythmia classification of LSTM autoencoder based on time series anomaly detection”. In: *Biomedical Signal Processing and Control* 71 (2022), p. 103228.

[358] Chong Zhou and Randy C Paffenroth. “Anomaly detection with robust deep autoencoders”. In: *Proceedings of the 23rd ACM SIGKDD international conference on knowledge discovery and data mining*. 2017, pp. 665–674.

[359] Emmanuel J Candès et al. “Robust principal component analysis?” In: *Journal of the ACM (JACM)* 58.3 (2011), pp. 1–37.

[360] Jinghui Chen et al. “Outlier detection with autoencoder ensembles”. In: *Proceedings of the 2017 SIAM international conference on data mining*. SIAM. 2017, pp. 90–98.

[361] Lukas Ruff et al. “Deep one-class classification”. In: *International conference on machine learning*. PMLR. 2018, pp. 4393–4402.

[362] Lukas Ruff et al. “Deep semi-supervised anomaly detection”. In: *arXiv preprint arXiv:1906.02694* (2019).

[363] Mayu Sakurada and Takehisa Yairi. “Anomaly detection using autoencoders with nonlinear dimensionality reduction”. In: *Proceedings of the MLSDA 2014 2nd workshop on machine learning for sensory data analysis*. 2014, pp. 4–11.

[364] Pascal Vincent et al. “Extracting and composing robust features with denoising autoencoders”. In: *Proceedings of the 25th international conference on Machine learning*. 2008, pp. 1096–1103.

[365] Guillaume Alain and Yoshua Bengio. “What regularized auto-encoders learn from the data-generating distribution”. In: *The Journal of Machine Learning Research* 15.1 (2014), pp. 3563–3593.

[366] Jinwon An and Sungzoon Cho. “Variational autoencoder based anomaly detection using reconstruction probability”. In: *Special lecture on IE* 2.1 (2015), pp. 1–18.

[367] Alireza Makhzani et al. “Adversarial autoencoders”. In: *arXiv preprint arXiv:1511.05644* (2015).

[368] Valentin Leveau and Alexis Joly. "Adversarial autoencoders for novelty detection". PhD thesis. Inria-Sophia Antipolis, 2017.

[369] Shuangfei Zhai et al. "Deep structured energy based models for anomaly detection". In: *International conference on machine learning*. PMLR. 2016, pp. 1100–1109.

[370] Longyuan Li et al. "Anomaly detection of time series with smoothness-inducing sequential variational auto-encoder". In: *IEEE transactions on neural networks and learning systems* 32.3 (2020), pp. 1177–1191.

[371] Julia Wolleb et al. "Diffusion models for medical anomaly detection". In: *International Conference on Medical image computing and computer-assisted intervention*. Springer. 2022, pp. 35–45.

[372] Erik Norlander and Alexandros Sopasakis. "Latent space conditioning for improved classification and anomaly detection". In: *arXiv preprint arXiv:1911.10599* (2019).

[373] Szymon Sacher, Laura Battaglia, and Stephen Hansen. "Hamiltonian Monte Carlo for regression with high-dimensional categorical data". In: *arXiv preprint arXiv:2107.08112* (2021).

[374] Diederik P Kingma, Max Welling, et al. "An introduction to variational autoencoders". In: *Foundations and Trends® in Machine Learning* 12.4 (2019), pp. 307–392.

[375] Xavier Glorot and Yoshua Bengio. "Understanding the difficulty of training deep feedforward neural networks". In: *Proceedings of the thirteenth international conference on artificial intelligence and statistics*. JMLR Workshop and Conference Proceedings. 2010, pp. 249–256.

[376] Jiancheng Yang et al. "Medmnist v2-a large-scale lightweight benchmark for 2d and 3d biomedical image classification". In: *Scientific Data* 10.1 (2023), p. 41.

[377] Daniel S Kermany et al. "Identifying medical diagnoses and treatable diseases by image-based deep learning". In: *cell* 172.5 (2018), pp. 1122–1131.

[378] Angela McLuckie. *Respiratory disease and its management*. Springer Science & Business Media, 2009.

[379] F. Pedregosa et al. "Scikit-learn: Machine Learning in Python". In: *Journal of Machine Learning Research* 12 (2011), pp. 2825–2830.

[380] Bjoern H Menze et al. "The multimodal brain tumor image segmentation benchmark (BRATS)". In: *IEEE transactions on medical imaging* 34.10 (2014), pp. 1993–2024.

[381] Masike Malatji and Alaa Tolah. "Artificial intelligence (AI) cybersecurity dimensions: a comprehensive framework for understanding adversarial and offensive AI". In: *AI and Ethics* (2024), pp. 1–28.

[382] Firas Almasri. "Exploring the impact of artificial intelligence in teaching and learning of science: A systematic review of empirical research". In: *Research in Science Education* 54.5 (2024), pp. 977–997.

[383] D Hernandez-Contreras et al. "A quantitative index for classification of plantar thermal changes in the diabetic foot". In: *Infrared Physics & Technology* 81 (2017), pp. 242–249.

[384] Amith Khandakar et al. "A Novel Machine Learning Approach for Severity Classification of Diabetic Foot Complications Using Thermogram Images". In: *Sensors* 22.11 (2022), p. 4249.

[385] Pi-Chang Sun et al. “Relationship of skin temperature to sympathetic dysfunction in diabetic at-risk feet”. In: *Diabetes research and clinical practice* 73.1 (2006), pp. 41–46.

[386] James Bergstra and Yoshua Bengio. “Random search for hyper-parameter optimization.” In: *Journal of machine learning research* 13.2 (2012).

ULPGC
Universidad de
Las Palmas de
Gran Canaria