IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 19 September 2022, accepted 16 November 2022, date of publication 28 November 2022,
date of current version 1 December 2022.

Digital Object Identifier 10.1109/ACCESS.2022.3225107

APPLIED RESEARCH

Performance Evaluation of Deep Learning Models for Image Classification Over Small Datasets: Diabetic Foot Case Study

ABIAN HERNANDEZ-GUEDES[1,2], IDAFEN SANTANA-PEREZ[1], NATALIA ARTEAGA-MARRERO[3], HIMAR FABELO[2,4], GUSTAVO M. CALLICO[2], (Senior Member, IEEE), AND JUAN RUIZ-ALZOLA[1,3]

[1]Research Institute in Biomedical and Health Sciences (IUIBS), University of Las Palmas de Gran Canaria, 35016 Las Palmas de Gran Canaria, Spain
[2]Research Institute for Applied Microelectronics (IUMA), University of Las Palmas de Gran Canaria, 35017 Las Palmas de Gran Canaria, Spain
[3]IACTEC Medical Technology Group, Instituto de Astrofísica de Canarias (IAC), 38205 San Cristóbal de La Laguna, Spain
[4]Fundación Canaria Instituto de Investigación Sanitaria de Canarias (FIISC), 35019 Las Palmas de Gran Canaria, Spain

Corresponding author: Abian Hernandez-Guedes (abian.hernandez@ulpgc.es)

This work was supported in part by the Spanish Government and European Union (FEDER funds) as Part of Support Program in the Context of TALENT-HExPERIA (HypErsPEctRal Imaging for Artificial Intelligence Applications) Project under Contract PID2020-116417RB-C42. The work of Abian Hernandez was supported by the "Agencia Canaria de Investigacion, Innovacion y Sociedad de la Información (ACIISI)" through Pre-Doctoral Grant by the "Consejería de Economía, Conocimiento y Empleo," which is Partly-Financed by the European Social Fund (FSE) [POC 2014-2020, Eje 3 Tema Prioritario 74 (85%)]. The work of Himar Fabelo was supported by MCIN/AEI/10.13039/501100011033 funded by the European Union "NextGenerationEU/PRTR" under Grant FJC2020-043474-I.

ABSTRACT Data scarcity is a common and challenging issue when working with Artificial Intelligence solutions, especially those including Deep Learning (DL) models for tasks such as image classification. This is particularly relevant in healthcare scenarios, in which data collection requires a long-lasting process, involving specific control protocols. The performance of DL models is usually quantified by different classification metrics, which may provide biased results, due to the lack of sufficient data. In this paper, an innovative approach is proposed to evaluate the performance of DL models when labeled data is scarce. This approach, which aims to detect the poor performance provided by DL models, in spite of traditional assessing metrics indicating otherwise, is based on information theoretic concepts and motivated by the Information Bottleneck framework. This methodology has been evaluated by implementing several experimental configurations to classify samples from a plantar thermogram dataset, focused on early stage detection of diabetic foot ulcers, as a case study. The proposed network architectures exhibited high results in terms of classification metrics. However, as our approach shows, only two of those models are indeed consistent to generalize the data properly. In conclusion, a new methodology was introduced and tested to identify promising DL models for image classification over small datasets without relying exclusively on the widely employed classification metrics. Example code and supplementary material using a state-of-the-art DL model are available at https://github.com/mt4sd/PerformanceEvaluationScarceDataset.

INDEX TERMS Deep learning, information theory, information bottleneck, diabetes, thermal imaging.

I. INTRODUCTION

Artificial intelligence is on trend for multiple medical applications, such as segmentation, localization, classification, raw data analysis, or risk assessment for chronic wounds [1], [2], [3], [4]. Currently, the main approach in this area is the usage of Deep Learning (DL) models. These models have the capacity of manipulating large amounts of data to solve problems such as, for instance, detecting diseases from medical images. The performance of these models are comparable to

The associate editor coordinating the review of this manuscript and approving it for publication was Wenming Cao.

those obtained by healthcare professionals [5]. As a result, DL models provide an easily adaptable output with high accuracy, while reducing human bias, associated costs, and the burden of time-consuming tasks.

In spite of the advantages provided by DL models, its implementation and use in health-care scenarios have not been completely achieved. More studies are required to consider the integration of these algorithms in the health-care setting [3], [5]. The main challenge in this domain is often the lack of sufficient labeled data [6], known as the *data challenge*. This is caused by the difficulty to perform a systematic collection of data to create large and well-curated datasets for training DL models. In general, DL models tend to have a high number of parameters and may *overfit* the training dataset, which is common bias if the training dataset is not large enough. As a consequence, the model will perform well on the training dataset and poorly on new data. There are techniques to mitigate this problem, such as *transfer learning* [7] or *data augmentation* [8], [9]. Furthermore, these problems are magnified by the current trend to deeper neural networks [10], [11], [12], where the *vanishing gradient* problem [13] is highly pervasive. Albeit *skip-connections* have been proved to work out this limitation and provide other benefits during the training process [14].

When working on classification problems, in general, the performance of the model is measured by metrics, such as the *accuracy*, using a *test set*. However, due to data challenge, the number of samples in the test set is probably insufficient, and thus these metrics are not robust enough to properly measure the performance of the model. In addition, *interpretability* is especially complex in DL models, making demanding to understand the outputs generated by the model. For this reason, it is often complicated to trust DL model's predictions, particularly in the medical domain, where clinical decision-making relies heavily on evidence interpretation [6].

In this study, we demonstrate the effectiveness of using information theoretic concepts [15] to improve the interpretability of DL models when working with small sets of labelled data. This allows us to also identify overfitting in DL models, which could not be assessed with traditional classification metrics, due to the reduced number of samples in the test set. To the best of our knowledge, this kind of methodology has not been previously reported for a small dataset.

In order to evaluate this methodology, several experimental configurations have been carried out to classify samples from a plantar thermogram dataset for Diabetic Foot Ulcer (DFU) detection [16], [17], [18] as case study. This type of studies aim at predicting the location of a possible future wound, by analyzing the temperature pattern of the entire plantar aspects of both feet. Abnormal patterns may indicate a foot disorder, such as peripheral arterial disease, neuropathy, and infection among others [19], [20]. The main goal of our approach is to identify the most suitable model for the classification task among those implemented. However, interesting observations revealed some additional contributions which are summarized as follows:

- The analysis of neural networks from the framework of information theory allows us to identify promising models in such a way that it is not necessary to rely exclusively on classification metrics.
- This analysis is affordable with a scarce dataset, providing a means to identify the different features presented in the state-of-the-art, which has been analyzed with popular datasets.
- The use of skip-connections indicates that some layers may be irrelevant, and the information is exclusively transmitted through these skip-connections. In order to evaluate such behavior, a visualization tool is proposed to estimate the similarity between filters in a convolutional layer.
- In addition to being able to identify cases of overfitting, the underfitting is also noticeable based on this analysis.
- Data normalization performed to improve temperature patterns, which is acceptable for our application, looks promising and allows the unification of independent datasets.

This paper is structured as follows. A state-of-the-art review is provided in Section II. A description of the methods used for the analysis, the DL architectures and the dataset are presented in Section III. Section IV presents the experimental configuration and process. The results from the different experiments are reported in Section V. Finally, the conclusions are drawn in Section VII.

II. RELATED WORK

As previously mentioned, the early stage detection of DFU [16], [17] constitutes a challenging medical analysis scenario for the classification task. Although extensive literature can be found regarding the application of DL for wound classification [21], and particularly for diabetic ulcer identification [22], [23], the use of thermal imaging for DFU detection is an emerging area of research. The main challenge associated to this task is related to the lack of datasets containing enough curated information to train DL models. This is partially due to the lack of standardized acquisition protocols to generate high-quality data.

Currently, one of the largest DFU detection oriented datasets is the Diabetic Foot Ulcers Grand Challenge (DFUC 2020) [24]. This database is focused on locating ulcers that are already visible and contains 4000 visible images, showing close-ups of the foot, which are equally split for training and testing (2000/2000). This dataset has been widely used and tested using various state-of-the-art models [25]. The model created in [26] reported the best results (an F1-Score of 74.3%). Furthermore, other public dataset can be found, containing 754 visible images of healthy and diabetic ulcer skin from different patients [27]. This dataset has been comprehensively evaluated providing an F1-Score over 97% at best [28]. However, the differing imaging

modality and the scenario captured, prevent the use of this type of dataset on our intended application.

Regarding Diabetic Foot Thermograms, the INAOE (Instituto Nacional de Astrofísica, Óptica y Electrónica) thermogram database [16], released on December 2019, contains infrared images from 122 diabetic and 45 non-diabetic subjects. This database has been widely used for the classification task. Machine learning and Deep Neural Networks (DNNs) were studied applying a first step of segmentation, from which a vector of features was extracted, and then used as input [29]. The reported accuracy was close to 100% when using more complex models previously trained with another dataset. In addition, the image enhancement effect was reported for the detection of the diabetic foot using several state-of-the-art Convolutional Neural Networks (CNNs) [30], in which an F1-Score of 95% was achieved with MobileNetV2. At the same time, a feature extraction from the temperature map was carried out for classification using ML models. In this case, an F1-Score of 97% was reported as the best result when using AdaBoost and 10 features. Furthermore, three state-of-the-art DL architectures were studied to classify subjects with diabetes [31]. However, the authors acknowledged the issue associated to these complex models, which require large amount of data to train the thousands of parameters of the model, since the INAOE dataset is not large enough to train these models. For this reason, authors proposed an augmentation technique based on Fourier transform, achieving values above 95%, and even a perfect score of 100% with ResNetV2.

The high dependence of these models on the amount of data complicates their evaluation and data augmentation is the most commonly applied technique. However, the study of DNNs from a theoretical framework is a suitable alternative, validating empirical results with theoretical concepts. DNNs were previously expressed as information theoretic concept, considering a trade-off between compression and prediction, based on the Information Bottleneck (IB) method [32]. Thus, DNNs find a maximally compressed mapping of the input variable, preserving as much as possible the information on the output variable. In this way, Schwartz-Zi et al. [15], motivated by the IB framework, demonstrated the effectiveness of using visualization tools for a better understating of the training dynamics, learning processes and internal representations in DL.

III. MATERIALS AND METHODS

In this section we expose the different architectures explored in our analysis, as well as how *Mutual Information* (MI) and *saliency maps* were used in our approach to work with small datasets. We also introduce the thermal image databases we have used and discuss how they have been improved and combined.

A. 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 [33], VGG [34]) to make a representation and reduce the amount of information, generating a latent space Z with lower dimensionality than the input space X. 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. 1(a) illustrates an example of such architecture. Since the final goal is to classify images, the training objective is defined as follows:

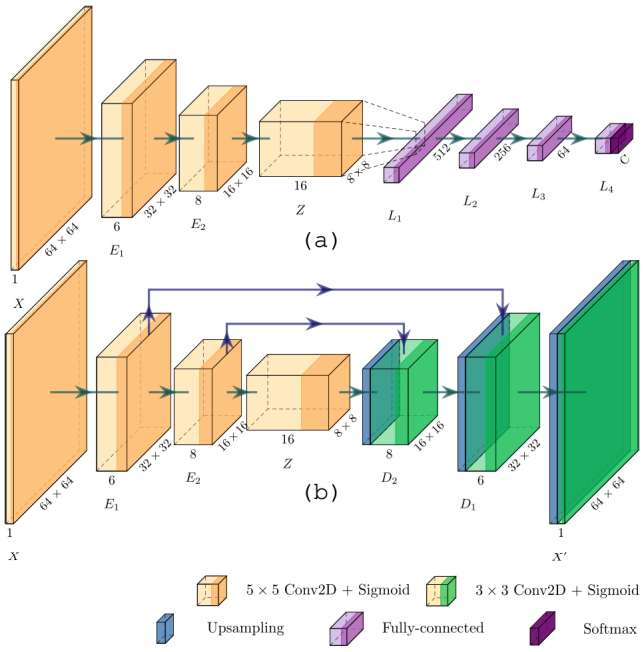


FIGURE 1. Proposed network architectures in this work for DFU classification. (a) Architecture proposed for classification where the convolutional layers are provided by the pre-trained AE. (b) Convolutional AE proposed where the skip connections are optional depending on the experiment.

$$\underset{\theta}{\operatorname{argmin}}\, \mathcal{L}(\hat{f}(;\theta)), \tag{1}$$

where θ is the parameter of $\hat{f}(;\theta)$ model and $\mathcal{L}$ is the Cross Entropy (CE) loss function:

$$CE(\hat{y}) = -\sum_{c} \lambda_c \mathbb{1}(y = c) \log(p(\hat{y})), \tag{2}$$

where $\hat{y} = \hat{f}(x;\theta)$, $\mathbb{1}(.)$ is the indicator function and $p(\hat{y})$ denotes the softmax probability of sample x. The ground truth label corresponds to y and c denotes the class category whose weight is indicated by λ_c.

When working with small datasets, transfer learning improves the classification performance of the model by using a previously trained model, tuning it to fit the classification task with the samples being studied [35]. The initial state θ_0 is generated from a training process with another dataset, which is usually larger and more complete.

In this study, an AutoEncoder (AE) architecture [36], depicted in Fig. 1(b), has been used to apply transfer learning from a pre-trained AE to the first layers of our classification architecture Fig. 1(a). AEs are characterized by three main differential components: the *encode path*, the *bottleneck*, which is the compressed latent space in the AE, and the *decode path*. The encode and decode paths correspond to a series of layers, $\{T_{E_1}, \ldots, T_{E_L}\}$ and $\{T_{D_L}, \ldots, T_{D_1}\}$, respectively, where L is the number of layers. An *encoder* is composed by the encode path following by the bottleneck, i.e, it corresponds to $T_E \in \{T_{E_1}, \ldots, T_{E_L}, T_Z\}$ and the *decoder* is composed by the decode path. Fig. 1(b) illustrates an example of a convolutional AE, based on U-Net architecture [37], which was used in our experiments.

B. INFORMATION PLANE ANALYSIS

In order to study the evolution of the models we are proposing, *Information Plane* (IP) [15], [32] has been used. IP is a visualization tool used to analyze how the estimation of MI [38] of a layer T, from a DNN with the input X and target Y, changes with the training epoch t [15], [39]. This type of analysis will increase the interpretability of the model by using information theory.

Regarding MI estimation, noted as I(.;.), it is often necessary to estimate the *Probability Mass Function* (PMF) by applying, for instance, a binning method [40]. However, the estimation of PMF in high-dimensionality data, which is common in DL image classification problems, is a computationally demanding task. Furthermore, the PMF estimation for a scarce dataset is not robust enough. Giraldo et al. [41] proposed a framework for data entropy estimation using infinitely divisible kernels and the axiomatic characterization of Renyi's α-order entropy, without assuming that the probabilities of events were estimated. Wickstrøm et al. [42] proposed to use the kernel-based MI estimator for the IP estimation. This kernel-based estimator is mathematically well-defined and computationally efficient, being a good choice for DNNs, where the output layer tends to have high-dimensionality. For this reason, we will use this approach in our system.

In this study, we are interested in using the Shannon's entropy definition and, accordingly, the limit $\alpha \to 1$ in the kernel-based estimator [41] was used to approximate the generalized Renyi's entropy to the Shannon's entropy [42], noted as $H(.)$. Finally, for the kernel-based IP estimation, a library (IPDL[1]) was developed whose workflow is integrated to run in PyTorch [43].

1) INFORMATION PLANE EVOLUTION

The evolution of the IP estimation during the training process contains two phases [15]. In the first phase, *fitting phase*, the layers increase the information on Y (i.e. $I(T;Y)$ increases). During the second phase, *compression phase*, the layers reduce the information on X, (i.e. $I(X;T)$ decreases). The compression phase is linked to generalization, where irrelevant information is compressed to prevent overfitting [15]. Nevertheless, the link between compression and generalization is still under discussion [44].

[1]https://github.com/mt4sd/IPDL

TABLE 1. Dataset summary.

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

2) DATA PROCESSING INEQUALITY

Due to the architecture of a DNN, where the output of a layer T_i depends on the output of layer T_{i-1}, a Markov Chain is formed [15], [42], [45]. This information path should satisfy the following *Data Processing Inequality* (DPI):

$$I(X;T_1) \geq I(X;T_2) \geq \ldots \geq I(X;T_L),$$

where L is the number of layers in the DNN.

C. SALIENCY MAP VISUALIZATION

There are different approaches to compute importance scores for generating a feature-importance map (*saliency map*). The reason for visualizing a saliency map for a specific image is to try to gain some understanding of what features our model detects, and it is widely used for Convolutional Neural Networks (CNN). In this work, DeepLIFT algorithm was employed [46]. This algorithm assigns importance scores, or attributions, by looking at the differences of the output with respect to the reference output in terms of differences between inputs and their reference inputs. This means that, given Δt as the difference between the output of a neuron x_i for a given input with its reference output, it can assign feature contribution scores $C_{\Delta x_i\ \Delta t}$ to the differences of the activations of neurons Δx_i:

$$\sum_{i=1}^{n} C_{\Delta x_i\ \Delta t} = \Delta t,$$

where n is the number of neurons in the intermediate layer or set of layers that are necessary and sufficient to compute t. Note that $C_{\Delta x_i\ \Delta t}$ can be non-zero even when $\partial t/\partial x_i$ is zero which may occur in integrated gradients methods [47].

In addition, DeepLIFT manages to compute these contribution scores by specifying some rules, which are discussed in detail in [46]. In this work, Rescale rule, which applies to nonlinear transformations such as ReLU or Sigmoid, was employed since the implementation used from Captum supports this rule [48].

D. INFRARED THERMAL IMAGE DATASET

With the purpose of testing our approach in a binary classification task, 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

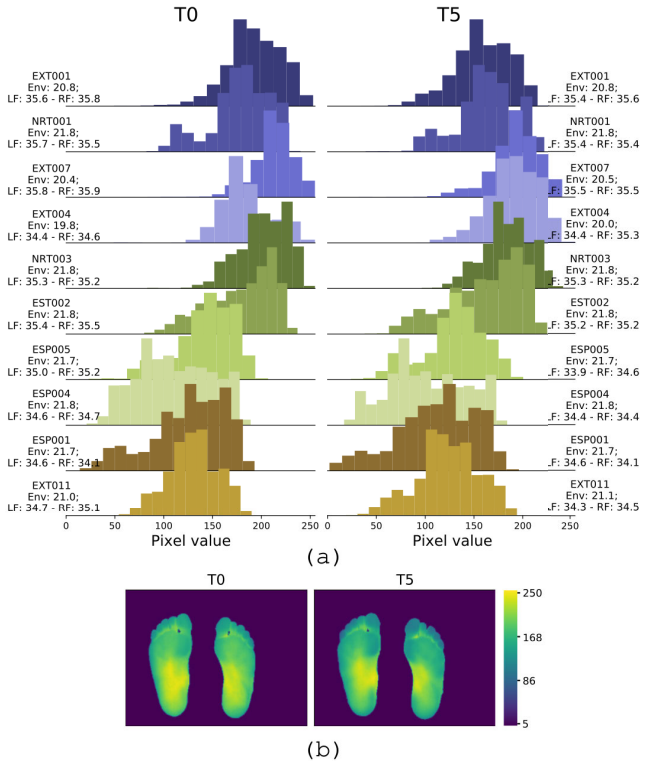


FIGURE 2. (a) Histograms of a pair of images (T0 and T5) from the same subject extracted from the IACTEC dataset. The temperature of each foot, left (LF) and right (RF) foot, and environment (Env) is detailed on the sides in degrees Celsius. (b) Thermal maps of T0 and T5 from such subject.

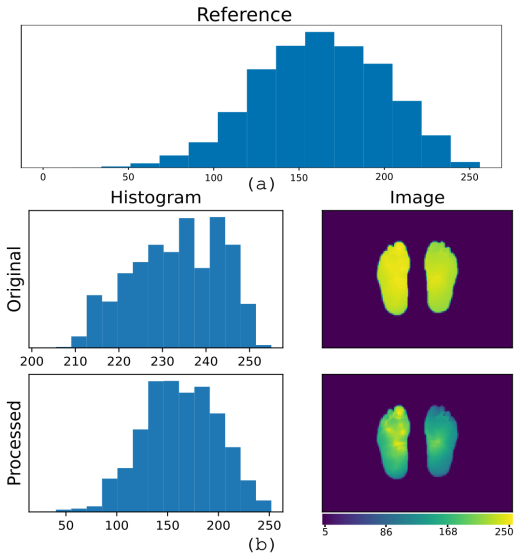


FIGURE 3. (a) Reference histogram. (b) Example of the histogram matching processing for a single subject.

from diabetic and non-diabetic subjects, as summarized in Table 1. The INAOE dataset was employed to support the analysis carried out to evaluate the performance of different types of architectures. The dataset was originally intended to study how the temperature is distributed in the plantar region from diabetic and non-diabetic subjects, and how those differences can be measured. The dataset is composed by 167 volunteers, 105 female and 62 male, with a mean age of 27.76 ± 8.09 in the control group and 55.98 ± 10.57 from diabetic group. For the acquisition, the authors used two different infrared cameras (FLIR E60 and FLIR E6). As stated by the authors, the dataset is slightly unbalanced towards diabetic cases that almost tripled those from the control group.

In order to balance the number of samples per class we integrated a second dataset, generated by IACTEC,[2] the technology center associated to the Astrophysical Research Institute from the Canary Islands (Instituto de Astrofísica de Canarias, IAC).[3] This dataset [17] 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 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.

[2]https://www.iac.es/es/observatorios-de-canarias/iactec
[3]https://www.iac.es/en

1) DATA PREPROCESSING

The samples from INAOE dataset were saved on CSV format and feet from the same person were separated into different files. Thus, images were preprocessed to unify both feet into the same images. Additionally, the same spatial resolution used in the IACTEC dataset was applied. Finally, the resulting thermal images were normalized and saved as 8-bit PNG. Normalization was performed as follows:

$$x_i' = (x_i * 255)/x_{max},$$

where x is the pixel value, the subscript i represents the pixel index and x_{max} is the max value in the image.

2) 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 [49]. 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 on the images for the purpose of our analysis.

TABLE 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	-

In this way, the IACTEC dataset was established as reference, since it offers a well-known acquisition protocol [17]. Fig. 2(a) 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 (Fig. 2(b)), 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. 2(a), 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. Fig. 3(a) illustrates the reference histogram, while Fig. 3(b) 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. 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 exactly 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.

IV. PROPOSED EXPERIMENTAL CONFIGURATIONS

In order to evaluate our approach for classifying DFU thermal images as diabetic vs non-diabetic, and how it can be applied to different scenarios, several experimental configurations were defined based on the architecture depicted in Fig. 1(a). These configurations can be divided into two main categories: using the pretrained encoder, identified by 'P', and not using it (NP). Table 2 shows the summary of the proposed configurations that are discussed in the following sections. In addition, a description of the fine-tuning process employed to generate the pretrained encoder is detailed below.

A. AUTOENCODERS FOR FINE-TUNING

For the experiments where transfer learning was applied, an AE was trained. Subsequently, the encode path layers were used as T_E in the proposed model to classify the DFU dataset (see Section III-A). The main advantage of AEs is that a labelled dataset is not required. However, in this work, the main reason to use AEs was the evaluation of skip-connections technique and its effect on the compressed representation of the input, which has been studied in Section V-B.

Considering this approach, AEs were trained to reconstruct an input X from a compressed representation Z (i.e., the compressed latent space from the bottleneck T_Z). Thus, applying image reconstruction is straightforward, as labeled samples are not necessary. Given an output $X' = \hat{f}(X;\theta)$, representing the reconstructed image, the model is evaluated by comparing X' with the original one X, using the Mean Square Error (MSE). Thus, the loss function, $\mathcal{L}$ in (1), is replaced by:

$$MSE(\hat{f}(X;\theta), X) = \frac{1}{N}\sum_{x\in X}(x - \hat{f}(x;\theta))^2, \tag{3}$$

where N is the number of samples in the dataset.

This training process has been carried out in two steps: a first training using a dataset with a large number of 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 [50]. However, since the INAOE dataset [16] is the only public thermogram dataset for DFU, currently, it is not possible to obtain another dataset in the same domain. Therefore, Fashion-MNIST (FMNIST) [51] was selected as the reference dataset for pretraining. It consists on a large dataset of grayscale images with a black background and a normalized histogram, being a dataset with similar features to our preprocessed dataset. 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 have demonstrated in our experiments to be robust for obtaining coherent feature extraction filters in Z.

B. EXPERIMENT DESCRIPTIONS

In this section, the different experiments depicted in Fig. 4 are described in details, including each network's architecture, as well as their most relevant hyperparameters.

1) EXPERIMENT 1, PCAE

A Pretrained Convolutional AE without skip-connections (*PCAE*) was used to generate the encoder for DFU classification. The architecture of this convolutional AE is illustrated

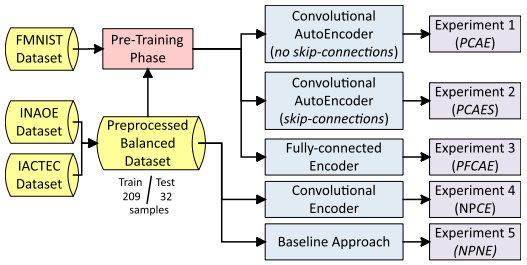


FIGURE 4. The different configurations for the experiments, where the pretraining phases were done using AEs.

in Fig. 1(b), excluding the skip-connections. The encoder contains the convolutional layers, with convolutional kernel size ($\mathcal{K}$) of 5×5 with stride and padding of 2. The decoder contains upsample layers, using 2D nearest neighbor interpolation, followed by convolutional layers with $\mathcal{K}$ 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.

2) EXPERIMENT 2, PCAES

A Pretrained Convolutional AE, similar to the previous case but including skip-connections (*PCAES*) was employed. This experiment uses the architecture illustrated in Fig. 1(a,b), applying skip-connections.

3) EXPERIMENT 3, PFCAE

A Pretrained AE of fully-connected layers (*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 1) and, in this case, $Z \in \mathbb{R}^{256}$.

4) EXPERIMENT 4, NPCE

This classification model architecture is similar to the ones obtained in PCAE and PCAES experiments. However, the classification model uses a Non-Pretrained Convolutional Encoder (*NPCE*).

5) EXPERIMENT 5, NPNE

This is the baseline approach, in which there is not a latent space generated from an encoder, being the original image the input of the classifier. This experiment will be referred to as *NPNE*.

V. EXPERIMENTAL RESULTS

The results presented in this section have been obtained using a batch of 128 samples for training and 32 samples for testing. The test set is balanced, taking 16 samples from the control group and the rest from diabetic group. ADAM optimizer [52] was used as optimizer for the DNN training. The initial learning rate (lr) was set to $5e^{-4}$ and the parameters to control exponential decay rates for the moment estimation, β_1 and β_2, were set to 0.9 and 0.999 respectively. The learning rate

TABLE 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

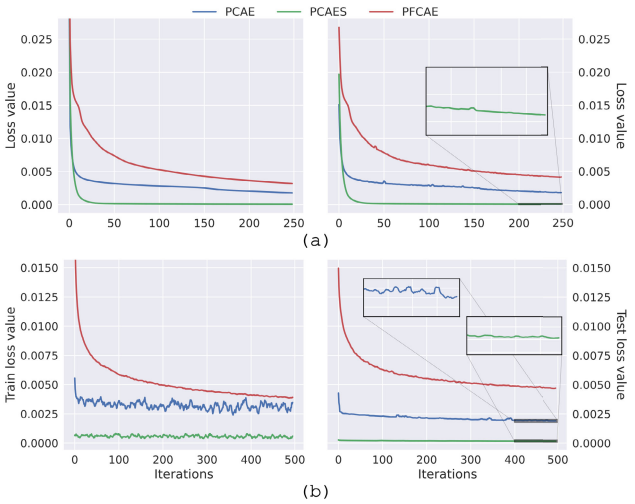


FIGURE 5. Loss value in training (left) and testing (right): (a) FMNIST dataset and (b) DFU dataset.

decays by κ every iteration t:

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

where κ was set to 0.999.

In order to evaluate the performance of the classification models, *sensitivity*, *specificity*, *precision*, and *accuracy* were estimated. Using these metrics, the classification results of the experimental configurations are summarized in Table 3. According to the estimations of the implemented metrics, all models seem promising considering the task at hand. A comprehensive analysis, described below, has been carried out to characterize the different models and truly identify the most promising ones.

A. DFU CLASSIFICATION ANALYSIS

The classification task was analyzed to check whether the models generated in the proposed configurations are as promising as suggested by the classification metrics shown in Table 3. In the experiments where a pretrained AE, which is detailed in Section IV-A, was used, the encode path and bottleneck were used for initializing the encoder in the classification model. Hence, the layers $T_E \in \{E_1, E_2, Z\}$ have been initialized using the AE configuration, see Fig. 1(b).

In those experiments that use a pretrained encoder (i.e., PCAE, PCAES and PFCAE), the AEs exhibit good performance, taking into account the evolution of the loss value depicted in Fig. 5. The MSE (3) loss value, which represents

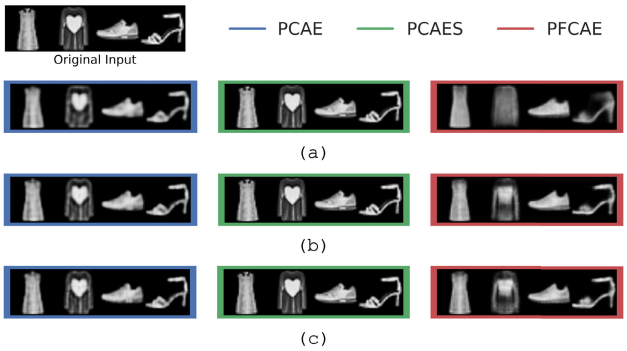


FIGURE 6. Qualitative results in different iterations during FMNIST training: (a) 24 iterations, (b) 170 iterations and (c) 250 iterations.

the error between desired output X and the reconstruction X', tends to decrease in all configurations. This behavior is present in both cases, training and testing, suggesting that overfitting does not occur in the AEs. As can be seen in Fig. 6, the final qualitative result of X' is satisfactory for the different configurations, being worse in the AE based on fully-connected layers (PFCAE) as expected. Taking into account these results, models that used the pretrained encoder can be expected to have good performance.

Subsequently, the classification model has to be evaluated and, as previously mentioned, this analysis was supported by the IP estimation with the main limitation that our test set was sparse. Regarding the kernel-based MI estimation, the *Radial Basis Function* (RBF) kernel was applied. In these experiments, the *kernel width* (σ) selection was carried out by maximizing the *kernel alignment loss* between the non-normalized Gram matrix of a given layer K_σ and the label matrix K_y, $\mathcal{A}(K_\sigma, K_y)$, as proposed by [42]. Thus, they choose the optimal σ as:

$$\sigma^* = \underset{\sigma}{\operatorname{argmax}}\, \mathcal{A}(K_\sigma, K_y). \tag{4}$$

Equation (4) 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 done in [42]. To stabilize the σ values across mini batches, an exponential moving average has been used.

Regarding the configurations, the loss value in the iteration t is illustrated in Fig. 7 (left side) where, excluding NPNE, convergence to an optimal solution is achieved based on CE loss value. In the NPNE case, the simplest case where the encoder was discarded, it is clear that the model is overfitting, decreasing the training loss value and increasing the test loss value per iteration (Fig. 7(e)).

The IP trajectories shown in Fig. 7 (right side) were used just in the classifier layers $T_{Cl} \in \{L_1, L_2, L_3, L_4\}$, discarding T_E. In the experiments where a pretrained AE was used, the input X for computing MI in IP trajectories is Z. In the NPCE and NPNE experiments, the input is the original X. 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

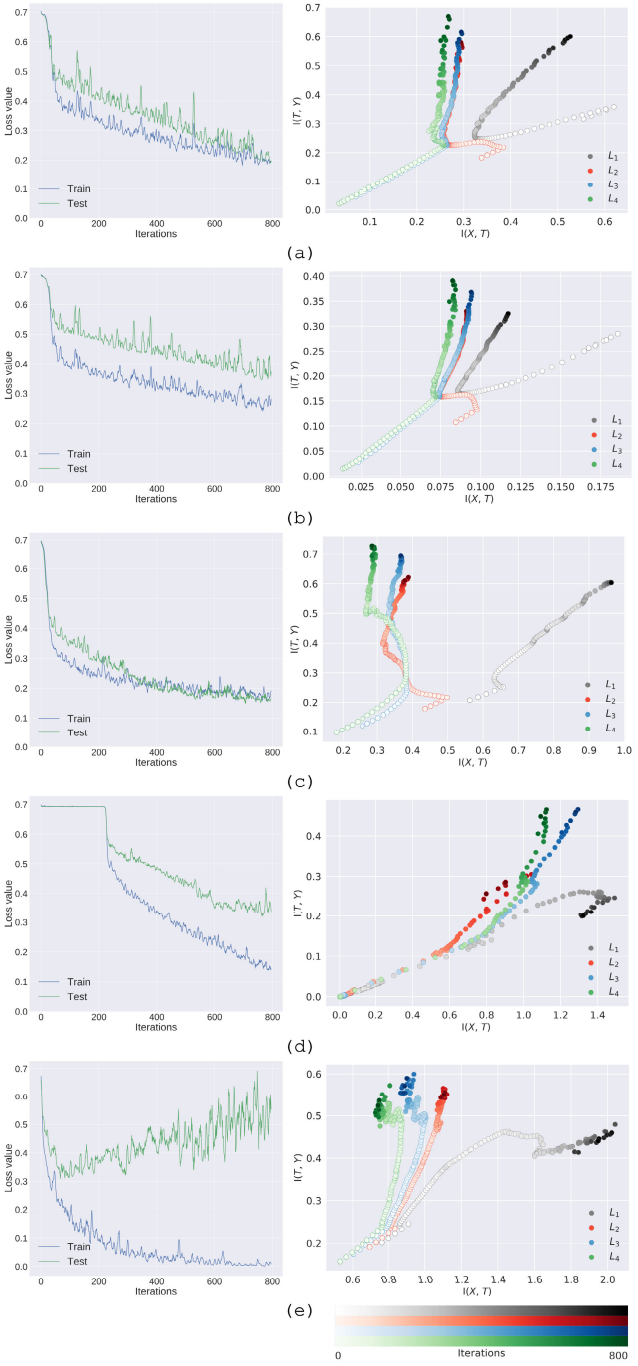


FIGURE 7. Loss error (left) and IP estimation (right) in the different experiments: (a) PCAE, (b) PCAES, (c) PFCAE, (d) NPCE and (e) NPNE.

theoretical maximum value in $I(T; Y)$ is given by $log_2(C) = log_2(2) = 1$.

Analyzing the non-pretrained experiments (NPCE and NPNE), Fig. 7(d,e), it can be observed that NPCE has too many iterations where the training CE (2) loss value was not able to converge, but it does afterward. The significant difference between training and testing losses that is depicted in Fig. 7(d,e) left, which is gradually increasing, indicates that the model is facing overfitting and perhaps underfitting. It is

crucial and results in poor progress and less generalization in labeled data. However, this problem is not so noticeable from Fig. 7(d) left. However, the MI estimation, depicted in Fig. 7(d) right, shows a violation in DPI (see Section III-B2), where $I(X, L_2) < I(X, L_3)$. This is clearly observable in later iterations after the compression phase. This violation could be related to the overfitting in the model during the compression phase as Wickstrom et al. suggested [42]. Finally, the $I(T; Y)$ estimation is far from the theoretical maximum value, indicating that the models are not performing properly.

Regarding NPNE, there is a clear overfitting as observed in Fig. 7(e) left, where the test CE loss value increases while the training CE loss value constantly decreases. The IP trajectories show that the estimation has an adjustment process. However, during training, such estimation gets stuck in a closed range, i.e, the estimation fluctuates constantly without showing an increase or decrease pattern. Considering this evidence, this might be a sign that the model is also underfitting at this moment, the number of parameters is not enough to characterize the data and accurately capture relationships between the input and target. At the same time, there is a DPI violation in Fig. 7(e), where $I(L_3; Y) > I(L_4; Y)$. Nonetheless, the range is so close that it can be due to the estimation of the metrics. Therefore, it can be concluded that non-pretrained models are not as promising as indicated by the performance metrics presented in Table 3.

From the pretrained approaches using transfer learning, the different models show a decreasing test CE loss value, as shown in Fig. 7(a,b,c), achieving lower values than non-pretrained approaches. Additionally, IP trajectories of PCAE, PCAES and PFCAE (Fig. 7(a,b,c)) show a constant fitting phase in most layers, discarding L_1 in PCAE and PCAES. In such cases, there is a decreasing trend in $I(X; T)$ in the earliest iterations, followed by a fitting phase. DPI violations were not observed, albeit the IP trajectories of L_2 and L_3 are close in PCAE and PCAES, overlapping each other. This overlap might be interpreted as both layers being similar, indicating that the model could be further reduced. The IP trajectories of PCAE and PCAES are similar, having both similar pattern. However, PCAES has the same problem that the non-pretrained approach: the $I(T; Y)$ is far from the theoretical maximum value. On the other hand, PCAE and PCAE have the closest values to the maximum theoretical value of $I(T; Y)$. As a conclusion, transfer learning works even when the reference dataset does not belong to the same domain as the target dataset, the DFU dataset.

B. SKIP-CONNECTION EFFECTS IN COMPRESSED REPRESENTATION

Following the results of the previous section, it has been possible to conclude that PCAES exhibited the worst performance among the models where transfer learning was applied. In this section, a comparison between PCAE and PCAES is carried out in order to understand why skip-connections have had such a negative impact on the obtained compressed latent space Z. In both cases, Z corresponds to an output of convolutional filters. Both AEs were trained identically in such a way that it can be considered a fair comparison between both experiments, with skip-connections being the only differentiating factor.

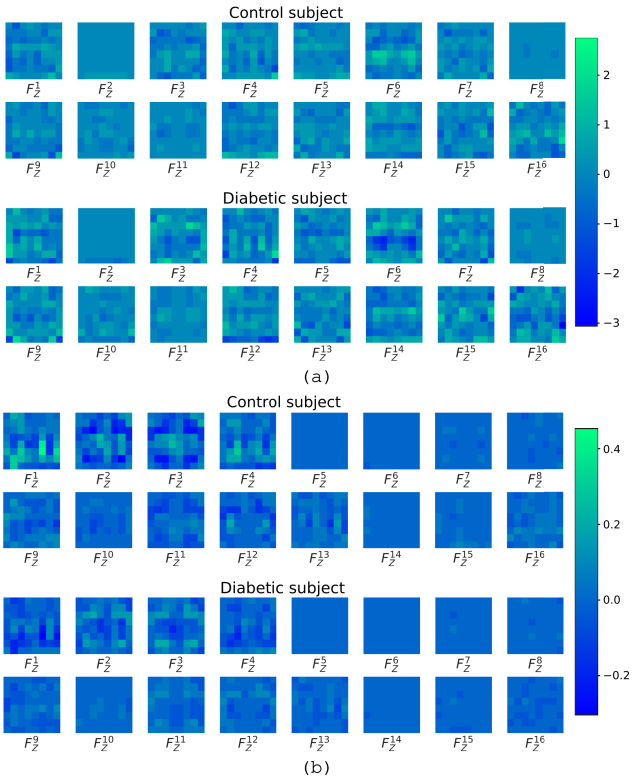


FIGURE 8. Saliency maps with features contribution in F_Z^i filters for DFU classification by DeepLIFT. The figure (a) corresponds to PCAE and (b) belong to PCAES.

Taking advantage of the fact that these models are able to accurately classify the samples, DeepLIFT was applied to identify which features from Z are being taken into account by the classifier, i.e., generating a saliency map in Z. This saliency map computes an importance score using a reference image. The selection of the reference image is critical, and the result will depend on this parameter, as it defines what is of interest in the input. For the DFU dataset, an all-zeros input was used as reference, representing the black background. This is represented in the saliency map of Fig. 8(a) and Fig. 8(b) for PCAE and PCAES configurations, respectively. In order to facilitate the interpretation of Fig. 8 a binning-clustering process was applied for 15 equidistant bins.

These results show that the feature contribution in both cases, control and diabetic group, is quite similar, obtaining saliency maps with similar patterns and conjugated values, since hot values (green color) in control tend to be cold values (blue color) in diabetic subjects. In PCAE, Fig. 8(a), most filters have spatial irregular patterns in the feature

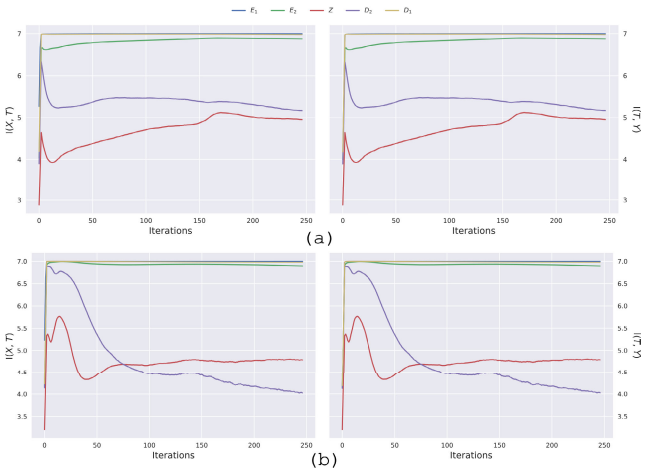


FIGURE 9. IP estimation in the FMNIST pretraining step: (a) PCAE, (b) PCAES.

contribution estimation. However, F_Z^2 and F_Z^8 show a limited activation, having a low feature contribution. Regarding PCAES, it can be observed in Fig. 8(b) that there are only a few relevant filters, F_Z^{1-4}. However, the scale of the feature contribution by DeepLIFT, depicted in the colorbar in Fig. 8, differs significantly and is reduced by a factor of 10 in PCAES.

In addition, the kernel-based MI estimation was applied to estimate the IP during the FMNIST reconstruction training using AEs (Fig. 9) and to obtain an interpretation of the different filters in Z, F_Z^i, estimating the MI between them (Fig. 10). As in the previous section, RBF kernel was applied for kernel-based MI estimation. The σ selection was carried out using Silverman's rule [53] that depends on an empirically determined constant, γ, which has been set $\gamma = 2$ in this work.

As expected, the IP trajectories (Fig. 9), generated applying the kernel-based MI estimation, in the different AEs, shows that $I(X,T) \approx I(T,Y)$. This is due to the fact that the desired output is roughly similar to the input. Those symmetrical trajectories are highly presented because of the high-quality reconstruction in few iterations. In order to verify that these estimations are correct, the DPI principle should be satisfied on the encoder and decoder layers [45]:

$$I(X,E_1) \geq \ldots \geq I(X,E_L) \geq I(X,Z),$$
$$I(Z,Y) \leq I(D_L,Y) \leq \ldots \leq I(D_1,Y).$$

Observing Fig. 9, a violation of the DPI principle ocurred in PCAES, as $I(Z,Y) > I(D_2,Y)$. However, the DPI principle is based on the assumption that a DNN can be interpreted as a Markov Chain model, but skip-connections link the encode and decode path (see Fig. 1(b)), 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.

Following our hypothesis, layers with redundant information should contain a high MI value and, at the same time, the filters with low entropy estimation do not contain relevant information of X. In order to validate this hypothesis, a visualization tool is proposed for evaluating the similarity between filters in a convolutional layer. The results are depicted in Fig. 10 where cells with intense color illustrate that its MI estimation is close to entropy estimation on each filter (the diagonal of the matrix). In any case, the Z used as the optimal representation of X contains minimal redundant information as well as low entropy. Due to the specific features of the DFU dataset, where the background is presented as a uniform black color and images are similar, it is understandable that, in a low-resolution image, the entropy is low because likely values are undervalued.

Observing PCAE results, Fig. 10(a), there are two specific filters, F_Z^2 and F_Z^8, in which entropy values are quite lower than in other filters. This is probably because the results of both filters are images with uniform values in a close range (i.e., images with reduced information). The saliency maps (Fig. 8(a)) reinforce the conclusions gathered by studying Fig. 10(a). Thus, it can be concluded that such filters could be removed without drastically affecting the classifier.

Regarding PCAES, Fig. 10(b) shows that most filters from Z contain a low entropy, which means that most activations are present in limited regions giving as a result a homogeneous output. Taking this into account, it can be concluded that PCAES contains homogeneous filters that offer very limited information. Therefore, it constitutes a poor representation from input X, which is supported by that observed in the IP estimation. This would explain the problems in the classification observed in Fig. 7(b). Moreover, the most complex features correspond to the ones with higher values in MI estimation, $F_Z^{8,10}$, as shown in Fig. 10(b).

VI. DISCUSSION

The lack of sufficient labeled data has several limitations. There is a trade-off between the number of samples used for training and those used to validate the model, and thus, it is difficult to assess whether the model is overfitting. In this work, this problem is presented since the use of thermal imaging for early stage detection of DFU is an emerging area of research, and there is a lack of data for the use of DNNs. As a consequence, the different experiments proposed in this work show some striking metrics, as in Table 3, where some perfect scores can be observed. These results are difficult to justify. On the one hand, the test set, which is composed of only 32 samples, might be a non-representative subset to evaluate the models. On the other hand, the DNN models could be overfitted.

For the aforementioned reasons, DNNs were analyzed via the theoretical framework of the IB principle, which was previously studied and validated using popular datasets with large number of samples for testing the models. This analysis

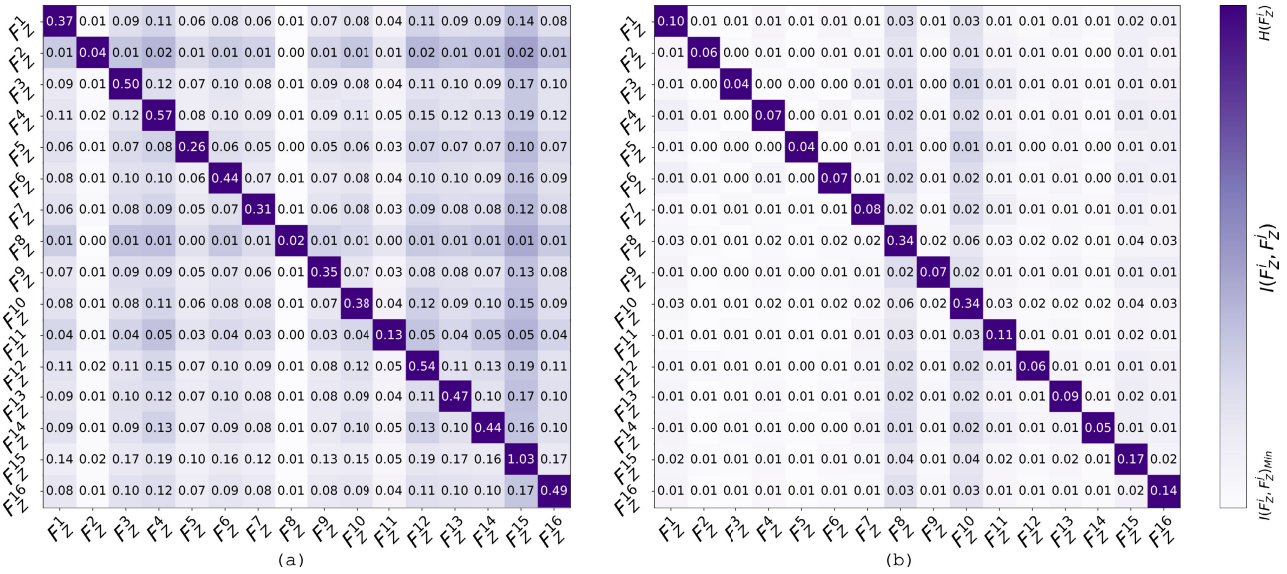


FIGURE 10. MI estimation between different filters in (a) PCAE and (b) PCAES. The trace of the heat-map represents $I(F_Z^i; F_Z^i) = H(F_Z^i)$. The colorbar applies to both charts.

was implemented using the IP, which has been estimated by the scarce test set and a kernel-based MI estimation, to validate the different models proposed. Transfer learning was applied to some of these models and those were expected to perform better.

As a result, theoretical characteristics were identified, already discussed in the state-of-the-art, in those experiments where transfer learning was applied in a way that validated the analysis for small datasets. As expected, transfer learning, even using a dataset with different target but certain features in common, improved the performance of the models. However, in one of these (PCAES), some limitations were detected which have cast doubt on its performance. Summarizing, it has been observed that skip-connections worsened the compressed space of X, in concordance with the IB theoretical framework. In addition, this conclusion was further supported by the use of saliency maps generated in Z by DeepLIFT and a MI estimation between the filters, where it was observed that many latent space filters do not contribute to the classification of the samples.

In those experiments where transfer learning was not used, the results were erratic and showed behaviors that cannot be validated following the IB theoretical framework, such as the DPI violation presented in both experiments, NPCE and NPNE. Furthermore, it was identified that the NPNE model was not able to characterize the data, due to the small number of parameters of this model in comparison with the others. This could indicate an underfitting problem, as the model was not complex enough to accurately capture relationships between the input and the target.

VII. CONCLUSION

The great potential of supervised models is found in using datasets with a large number of samples, so that the model can generalize. Nonetheless, this is not always possible, especially in the medical field where sample collection depends on time-consuming protocols. In this work, the use of the theoretical framework of the IB principle is proposed and tested for evaluating models, implemented using different architectures, in which the training dataset is rather small in terms of samples. For those cases, even when traditional classification metrics show great performance, our analysis clarifies whether those metric results are due to overfitting.

We conducted several experiments, designed with different DNN architectures, including the usage of transfer learning in three of them. As results, the classification evaluation metrics were promising in all experiments. When analyzing those experiments, using mainly IP analysis, we could conclude that just two out of the five experiments (i.e. PCAE and PFCAE) showed a consistent performance. This allows us also to conclude that the results obtained using just classification metrics on the other three experiments were not reliable.

Analyzing the results for our experiments, we can observe that they all contain several characteristics, already discussed in the state-of-the-art using a large dataset, that support the validity of our analysis. The analysis based on kernel-based MI has shown the different IP phases of DNN, specially in Z, as discussed above. In order to conclude which models have a better performance, the estimation $I(T; Y)$ has been compared with the maximum theoretical value, which is a straightforward method to estimate in classification problems for a balanced test set.

Finally, the effect of skip-connections was studied since it constitutes the differentiating factor in PCAES. For this analysis, an IP estimation in the pretrained AE was used for concluding that the skip-connections is a violation of

the Markov property which is fundamental for the IB principle. In addition, MI estimation between the convolutional filters in Z was carried out to estimate which filters provide appropriate information, supported by the saliency maps obtained from DeepLIFT. In conclusion, the skip-connections in PCAES resulted in Z, being a worse representation of compressed X. This approach might be interesting to obtain more efficient CNN architectures when the dataset available is scarce by, for example, applying regularization based on the information obtained by this analysis.

As for future work, we plan to extend the current evaluation to include new and different datasets, which fit the scope of being scarce. Finally, we would like to evaluate how the PCAE and PFCAE models work with larger DFU datasets, provided that such data would be available in the future.

ACKNOWLEDGMENT

The original codes developed in the current study are available from the corresponding author on reasonable request.

REFERENCES

[1] M. Elmogy, B. García-Zapirain, C. Burns, A. Elmaghraby, and A. Ei-Baz, "Tissues classification for pressure ulcer images based on 3D convolutional neural network," in *Proc. 25th IEEE Int. Conf. Image Process. (ICIP)*, Oct. 2018, pp. 3139–3143.

[2] M. Goyal, N. D. Reeves, S. Rajbhandari, and M. H. Yap, "Robust methods for real-time diabetic foot ulcer detection and localization on mobile devices," *IEEE J. Biomed. Health Informat.*, vol. 23, no. 4, pp. 1730–1741, Jul. 2019.

[3] J. Tulloch, R. Zamani, and M. Akrami, "Machine learning in the prevention, diagnosis and management of diabetic foot ulcers: A systematic review," *IEEE Access*, vol. 8, pp. 198977–199000, 2020.

[4] L. Wang, P. C. Pedersen, E. Agu, D. M. Strong, and B. Tulu, "Area determination of diabetic foot ulcer images using a cascaded two-stage SVM-based classification," *IEEE Trans. Biomed. Eng.*, vol. 64, no. 9, pp. 2098–2109, Sep. 2017.

[5] X. Liu, L. Faes, A. U. Kale, S. K. Wagner, D. J. Fu, A. Bruynseels, T. Mahendiran, G. Moraes, M. Shamdas, C. Kern, J. R. Ledsam, M. K. Schmid, K. Balaskas, E. J. Topol, L. M. Bachmann, P. A. Keane, and A. K. Denniston, "A comparison of deep learning performance against health-care professionals in detecting diseases from medical imaging: A systematic review and meta-analysis," *Lancet Digit. Health*, vol. 1, no. 6, pp. e271–e297, Oct. 2019.

[6] S. K. Zhou, H. Greenspan, C. Davatzikos, J. S. Duncan, B. Van Ginneken, A. Madabhushi, J. L. Prince, D. Rueckert, and R. M. Summers, "A review of deep learning in medical imaging: Imaging traits, technology trends, case studies with progress highlights, and future promises," *Proc. IEEE*, vol. 109, no. 5, pp. 820–838, May 2021.

[7] Y. Bar, I. Diamant, L. Wolf, S. Lieberman, E. Konen, and H. Greenspan, "Chest pathology detection using deep learning with non-medical training," in *Proc. IEEE 12th Int. Symp. Biomed. Imag. (ISBI)*, Apr. 2015, pp. 294–297.

[8] C. Shorten and T. M. Khoshgoftaar, "A survey on image data augmentation for deep learning," *J. Big Data*, vol. 6, no. 1, pp. 1–48, Dec. 2019.

[9] J. Xu, M. Li, and Z. Zhu, "Automatic data augmentation for 3D medical image segmentation," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.* Cham, Switzerland: Springer, 2020, pp. 378–387.

[10] T. Kaur and T. K. Gandhi, "Automated brain image classification based on VGG-16 and transfer learning," in *Proc. Int. Conf. Inf. Technol. (ICIT)*, Dec. 2019, pp. 94–98.

[11] Y. Oh, S. Park, and J. C. Ye, "Deep learning COVID-19 features on CXR using limited training data sets," *IEEE Trans. Med. Imag.*, vol. 39, no. 8, pp. 2688–2700, Aug. 2020.

[12] D. Jha, M. A. Riegler, D. Johansen, P. Halvorsen, and H. D. Johansen, "DoubleU-Net: A deep convolutional neural network for medical image segmentation," in *Proc. IEEE 33rd Int. Symp. Comput.-Based Med. Syst. (CBMS)*, Jul. 2020, pp. 558–564.

[13] R. Pascanu, T. Mikolov, and Y. Bengio, "On the difficulty of training recurrent neural networks," in *Proc. Int. Conf. Mach. Learn.*, 2013, pp. 1310–1318.

[14] A. E. Orhan and X. Pitkow, "Skip connections eliminate singularities," 2017, *arXiv:1701.09175*.

[15] R. Shwartz-Ziv and N. Tishby, "Opening the black box of deep neural networks via information," 2017, *arXiv:1703.00810*.

[16] D. A. Hernandez-Contreras, H. Peregrina-Barreto, J. D. J. Rangel-Magdaleno, and F. J. Renero-Carrillo, "Plantar thermogram database for the study of diabetic foot complications," *IEEE Access*, vol. 7, pp. 161296–161307, 2019.

[17] N. Arteaga-Marrero, A. Hernández, E. Villa, S. González-Pérez, C. Luque, and J. Ruiz-Alzola, "Segmentation approaches for diabetic foot disorders," *Sensors*, vol. 21, no. 3, p. 934, Jan. 2021.

[18] A. Hernández, N. Arteaga-Marrero, E. Villa, H. Fabelo, G. M. Callicó, and J. Ruiz-Alzola, "Automatic segmentation based on deep learning techniques for diabetic foot monitoring through multimodal images," in *Proc. Int. Conf. Image Anal. Process.* Cham, Switzerland: Springer, 2019, pp. 414–424.

[19] C. Liu, J. J. van Netten, J. G. van Baal, S. A. Bus, and F. van der Heijden, "Automatic detection of diabetic foot complications with infrared thermography by asymmetric analysis," *J. Biomed. Opt.*, vol. 20, no. 2, Feb. 2015, Art. no. 026003.

[20] K. Roback, "An overview of temperature monitoring devices for early detection of diabetic foot disorders," *Expert Rev. Med. Devices*, vol. 7, no. 5, pp. 711–718, Sep. 2010.

[21] G. Blanco, A. J. M. Traina, C. Traina, Jr., P. M. Azevedo-Marques, A. E. S. Jorge, D. de Oliveira, and M. V. N. Bedo, "A superpixel-driven deep learning approach for the analysis of dermatological wounds," *Comput. Methods Programs Biomed.*, vol. 183, Jan. 2020, Art. no. 105079.

[22] M. H. Yap, R. Hachiuma, A. Alavi, R. Brüngel, B. Cassidy, M. Goyal, H. Zhu, J. Rückert, M. Olshansky, X. Huang, and H. Saito, "Deep learning in diabetic foot ulcers detection: A comprehensive evaluation," *Comput. Biol. Med.*, vol. 135, Aug. 2021, Art. no. 104596.

[23] M. Goyal, N. D. Reeves, S. Rajbhandari, N. Ahmad, C. Wang, and M. H. Yap, "Recognition of ischaemia and infection in diabetic foot ulcers: Dataset and techniques," *Comput. Biol. Med.*, vol. 117, Feb. 2020, Art. no. 103616.

[24] B. Cassidy, N. D. Reeves, J. M. Pappachan, D. Gillespie, C. O'Shea, S. Rajbhandari, A. G. Maiya, E. Frank, A. J. Boulton, D. G. Armstrong, and B. Najafi, "The DFUC 2020 dataset: Analysis towards diabetic foot ulcer detection," *touchREVIEWS Endocrinol.*, vol. 17, no. 1, p. 5, 2021.

[25] M. H. Yap, B. Cassidy, J. M. Pappachan, C. O'Shea, D. Gillespie, and N. D. Reeves, "Analysis towards classification of infection and ischaemia of diabetic foot ulcers," in *Proc. IEEE EMBS Int. Conf. Biomed. Health Informat. (BHI)*, Jul. 2021, pp. 1–4.

[26] M. Tan, R. Pang, and Q. V. Le, "EfficientDet: Scalable and efficient object detection," 2019, *arXiv:1911.09070*.

[27] L. Alzubaidi, M. A. Fadhel, S. R. Oleiwi, O. Al-Shamma, and J. Zhang, "DFU_QUTNet: Diabetic foot ulcer classification using novel deep convolutional neural network," *Multimedia Tools Appl.*, vol. 79, nos. 21–22, pp. 15655–15677, Jun. 2020.

[28] L. Alzubaidi, M. A. Fadhel, O. Al-Shamma, J. Zhang, J. Santamaría, and Y. Duan, "Robust application of new deep learning tools: An experimental study in medical imaging," *Multimedia Tools Appl.*, vol. 81, no. 10, pp. 13289–13317, Apr. 2022.

[29] I. Cruz-Vega, D. Hernandez-Contreras, H. Peregrina-Barreto, J. D. J. Rangel-Magdaleno, and J. M. Ramirez-Cortes, "Deep learning classification for diabetic foot thermograms," *Sensors*, vol. 20, no. 6, p. 1762, Mar. 2020.

[30] A. Khandakar, M. E. H. Chowdhury, M. B. I. Reaz, S. H. M. Ali, M. A. Hasan, S. Kiranyaz, T. Rahman, R. Alfkey, A. Ashrif A. Bakar, and R. A. Malik, "A machine learning model for early detection of diabetic foot using thermogram images," 2021, *arXiv:2106.14207*.

[31] A. Anaya-Isaza and M. Zequera-Diaz, "Fourier transform-based data augmentation in deep learning for diabetic foot thermograph classification," *Biocybernetics Biomed. Eng.*, vol. 42, no. 2, pp. 437–452, Apr. 2022.

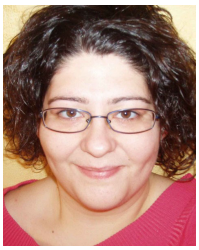
[32] N. Tishby and N. Zaslavsky, "Deep learning and the information bottleneck principle," in *Proc. IEEE Inf. Theory Workshop (ITW)*, Apr. 2015, pp. 1–5.

[33] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.

[34] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2014, *arXiv:1409.1556*.
[35] S. Bozinovski, "Reminder of the first paper on transfer learning in neural networks, 1976," *Informatica*, vol. 44, no. 3, pp. 1–12, Sep. 2020.
[36] G. E. Hinton and R. R. Salakhutdinov, "Reducing the dimensionality of data with neural networks," *Science*, vol. 313, no. 5786, pp. 504–507, 2006.
[37] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.* Cham, Switzerland: Springer, 2015, pp. 234–241.
[38] J. C. Principe, *Information Theoretic Learning: Renyi's Entropy and Kernel Perspectives*. Cham, Switzerland: Springer, 2010.
[39] B. C. Geiger, "On information plane analyses of neural network classifiers—A review," *IEEE Trans. Neural Netw. Learn. Syst.*, early access, Jun. 30, 2021, doi: 10.1109/TNNLS.2021.3089037.
[40] K. Pearson, "Contributions to the mathematical theory of evolution," *Philos. Trans. Roy. Soc. London*, vol. 185, pp. 71–110, Jan. 1894.
[41] L. G. S. Giraldo, M. Rao, and J. C. Principe, "Measures of entropy from data using infinitely divisible kernels," *IEEE Trans. Inf. Theory*, vol. 61, no. 1, pp. 535–548, Jan. 2015.
[42] K. Wickstrøm, S. Løkse, M. Kampffmeyer, S. Yu, J. Principe, and R. Jenssen, "Information plane analysis of deep neural networks via matrix-based Renyi's entropy and tensor kernels," 2019, *arXiv:1909.11396*.
[43] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, and A. Desmaison, "PyTorch: An imperative style, high-performance deep learning library," in *Proc. Adv. Neural Inf. Process. Syst.*, H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, Eds. Red Hook, NY, USA: Curran Associates, 2019, pp. 8024–8035.
[44] A. M. Saxe, Y. Bansal, J. Dapello, M. Advani, A. Kolchinsky, B. D. Tracey, and D. D. Cox, "On the information bottleneck theory of deep learning," *J. Stat. Mech., Theory Exp.*, vol. 2019, no. 12, Dec. 2019, Art. no. 124020.
[45] S. Yu and J. C. Príncipe, "Understanding autoencoders with information theoretic concepts," *Neural Netw.*, vol. 117, pp. 104–123, Sep. 2019.
[46] A. Shrikumar, P. Greenside, and A. Kundaje, "Learning important features through propagating activation differences," in *Proc. Int. Conf. Mach. Learn.*, 2017, pp. 3145–3153.
[47] M. Sundararajan, A. Taly, and Q. Yan, "Axiomatic attribution for deep networks," in *Proc. Int. Conf. Mach. Learn.*, 2017, pp. 3319–3328.
[48] N. Kokhlikyan, V. Miglani, M. Martin, E. Wang, B. Alsallakh, J. Reynolds, A. Melnikov, N. Kliushkina, C. Araya, S. Yan, and O. Reblitz-Richardson, "Captum: A unified and generic model interpretability library for PyTorch," 2020, *arXiv:2009.07896*.
[49] R. C. Gonzalez, R. E. Woods, and B. R. Masters, *Digital Image Processing*. Upper Saddle River, NJ, USA: Prentice-Hall, 2009.
[50] L. Alzubaidi, M. A. Fadhel, O. Al-Shamma, J. Zhang, J. Santamaría, Y. Duan, and S. R. Oleiwi, "Towards a better understanding of transfer learning for medical imaging: A case study," *Appl. Sci.*, vol. 10, no. 13, p. 4523, Aug. 2020.
[51] H. Xiao, K. Rasul, and R. Vollgraf, "Fashion-MNIST: A novel image dataset for benchmarking machine learning algorithms," 2017, *arXiv:1708.07747*.
[52] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2014, *arXiv:1412.6980*.
[53] B. W. Silverman, *Density Estimation for Statistics and Data Analysis*. Evanston, IL, USA: Routledge, 2018.

ABIAN HERNANDEZ-GUEDES received the bachelor's degree in computer science and the M.Sc. degree in telecommunication technologies from the ULPGC, Spain, in 2015 and 2017, respectively. He is currently pursuing the Ph.D. degree in telecommunication technologies and computer engineering, focusing his studies on image processing. He was a Computer Scientist. He has participated in various projects related to medical image processing (HELICoiD and ITHaCA) and a project related to the development of sustainable medical technology (MACbioIDi). He has participated in different National Alliance for Medical Image Computing (NA-MIC) events, presenting and collaborating on several projects to develop multimodal image processing algorithms, especially segmentation, registration, and classification processes. In 2020, he obtained a Predoctoral Research Grant from the Canary Islands Government. In 2022, he performed a research stay with the Department of Mechanical Engineering, Tokyo University of Science, Japan, headed by Prof. Takemura, collaborating in the use of hyperspectral imaging analysis and feature selection methods based on deep learning approaches. His research interests include medical image segmentation and deep learning algorithms for medical applications.

IDAFEN SANTANA-PEREZ received the degree in computer science from the Universidad de Las Palmas de Gran Canaria, in 2008, and the M.Sc. degree in complex software systems and the Ph.D. degree in artificial intelligence from the Universidad Politécnica de Madrid, in 2010 and 2016, respectively. He is currently working as a Developer, a Lecturer, and a Research Fellow at the DSC Department, Las Palmas de Gran Canaria University, focusing on sensor data for light pollution, medical image processing, and natural language processing. He was formerly a Research Fellow at the Ontology Engineering Group, Politécnica de Madrid University, where he worked on topics related to open science and linked data in general. From 2011 to 2015, he held a FPU Scholarship at the Artificial Intelligence Department of the Computer Science (2012). He was awarded for the second best academic record for his degree. During his research career, he has worked with several international research groups, including a research stay at the Information Science Institute, University of Southern California.



NATALIA ARTEAGA-MARRERO received the M.Sc. degree in applied physics from the University of La Laguna, Spain, in 2001, and the Licentiate and Ph.D. degrees in engineering from Lund University, Sweden, in 2007 and 2010, respectively. She is currently a Researcher at the Instituto de Astrofísica de Canarias (IACTEC, Medical Technology Group) focused on biomedical image acquisition and processing. She enrolled the Marie Curie Research Training Network (CELLION) to develop new tools for low-dose radiation research at LTH, Lund University. She became a Postdoctoral Researcher in a project dedicated to molecular imaging of cancer at the University of Bergen, Norway, in 2011. Subsequently, she joined a project focused on multi-parametric imaging for radiotherapy as a Postdoctoral Researcher at Umeå University, Sweden, in 2015. She has expertise and interest in cross-disciplinary projects related to cancer research covering the fields of molecular imaging (MRI, US, PET), including diagnosis, monitoring and prognosis, and particle irradiation (therapy).

HIMAR FABELO received the master's degree in telecommunication engineering and the Ph.D. degree in telecommunication technologies from the University of Las Palmas de Gran Canaria, Las Palmas de Gran Canaria, Spain, in 2014, and 2019, respectively. Since then, he has conducting his research activity in the Integrated System Design Division, Institute for Applied Microelectronics, University of Las Palmas de Gran Canaria, in the field of electronic and bioengineering. In 2015, he started to work as a Coordination Assistant and a Researcher in the HELICoiD (618080) European project, co-funded by the European Commission. In 2018, he performed a research stay with the Department of Bioengineering, Erik Jonsson School of Engineering and Computer Science, The University of Texas at Dallas, collaborating with Prof. Baowei Fei in the use of medical hyperspectral imaging analysis using deep learning. In 2021, he worked as a Project Manager and a Postdoctoral Researcher in the WARIFA (101017385) European project. In 2022, he started to work at the Fundación Canaria Instituto de Investigación Sanitaria de Canarias (FIISC) with a Juan De La Cierva grant. He has published more than 40 journal articles, 40 international conference papers, two book chapters, and holds one international patent (WO2018095516A1). His research interests include the use of machine learning and deep learning techniques applied to hyperspectral images to discriminate between healthy and tumor tissue in real-time during surgery and also for early cancer detection and screening.

GUSTAVO M. CALLICO (Senior Member, IEEE) received the M.S. degree (Hons.) in telecommunication engineering and the Ph.D. and European Doctorate degrees (Hons.) from the ULPGC, in 1995 and 2003, respectively. He is currently a Full Professor with ULPGC. From 1996 to 1997, he worked on a research grant from the National Educational Ministry. In 1997, he was hired by the ULPGC as an Electronics Lecturer. In 1994, he joined the Institute for Applied Microelectronics (IUMA). From 2000 to 2001, he worked at the Philips Research Laboratories (NatLab), Eindhoven, The Netherlands, as a Visiting Scientist, where he developed his Ph.D. thesis. He currently conducts his research activities in the Integrated Systems Design Division, IUMA. He has more than 200 publications in national and international journals, conferences, and book chapters. He has participated in 19 research projects funded by the European Community, the Spanish Government, and international private industries. Since 2015, he has been responsible for a scientific-technological equipment project entitled ''Hyperspectral image acquisition system of high spatial and spectral definition,'' granted by the General Directorate of Research and Management of the National Research and Development Plan, funded through the General Directorate of Scientific Infrastructure. He has been the Coordinator of the European project HELICoiD [Future and Emerging Technologies (FET)] under the Seventh Framework Program. He was a Visiting Professor at the University of Pavia, Italy, in October 2015 and March 2019, where he has been belonging to the Council of Doctors, since 2015. His current research interests include hyperspectral imaging for real-time cancer detection, real-time super-resolution algorithms, synthesis-based design for systems on chips and circuits for multimedia processing, and video coding standards, especially for H.264 and scalable video coding. He was an Associate Editor of the IEEE TRANSACTIONS ON CONSUMER ELECTRONICS, from 2009 to 2022, for which he was also a Senior Associate Editor. He has been an Associate Editor of IEEE ACCESS, since 2016.

JUAN RUIZ-ALZOLA received the degree in telecommunication engineering and the Ph.D. degree from the Polytechnic University of Madrid, in 1992 and 1998, respectively. He is currently an Imaging Technologies Full Professor in the knowledge area of signal processing and communications at the ULPGC, where he is a member of the Department of Signal and Communications, School of Telecommunication Engineering and Electronics. He is also an Affiliate Researcher at the Canary Islands Institute of Astrophysics (IAC), leading the medical technology program aimed at leveraging astrophysics technology for medical applications, particularly microwave, and infrared passive sensors for early detection of subcutaneous ulcerations. He is also the Director of the Advanced Technologies Service for the Smart Specialization of the Canary Islands, ULPGC Science and Technology Park Foundation, and the Medical and Audiovisual Technology Group, ULPGC Institute for Biomedical and Health Research. He was a Postdoctoral Visiting Research Fellow with the Surgical Planning Laboratory (SPL), Brigham and Women's Hospital, and Harvard Medical School (1999–2000), supported by a grant from the Spanish Government. Furthermore, he extended his collaboration as a Visiting Associate Professor at such institutions, until he was appointed as the Research and Technology Director of the Canary Islands Institute of Technology (2004–2007). He has directed several European, national, and regional research and innovation projects. He is the coauthor for over 100 research papers.