Attention-based Skin Cancer Classification Through Hyperspectral Imaging

Marco La Salvia
Department of Electrical, Computer and Biomedical Engineering
University of Pavia
Pavia, Italy
marco.lasalvia01@universitadipavia.it

Emanuele Torti
Department of Electrical, Computer and Biomedical Engineering
University of Pavia
Pavia, Italy
emanuele.torti@unipv.it

Marco Gazzoni
Department of Electrical, Computer and Biomedical Engineering
University of Pavia
Pavia, Italy
marco.lasalvia01@universitadipavia.it

Elisa Marenzi
Department of Electrical, Computer and Biomedical Engineering
University of Pavia
Pavia, Italy
elisa.marenzi@unipv.it

Raquel Leon
Institute for Applied Microelectronics (IUMA)
Universidad de Las Palmas de Gran Canaria
Las Palmas de Gran Canaria, Spain
slmartin@iuma.ulpgc.es

Samuel Ortega
Institute for Applied Microelectronics (IUMA)
Universidad de Las Palmas de Gran Canaria
Las Palmas de Gran Canaria, Spain
sortega@iuma.ulpgc.es

Himar Fabelo
Institute for Applied Microelectronics (IUMA)
Universidad de Las Palmas de Gran Canaria
Las Palmas de Gran Canaria, Spain
hfabelo@iuma.ulpgc.es

Gustavo M. Callico
Institute for Applied Microelectronics (IUMA)
Universidad de Las Palmas de Gran Canaria
Las Palmas de Gran Canaria, Spain
gustavo@iuma.ulpgc.es

Francesco Leporati
Department of Electrical, Computer and Biomedical Engineering
University of Pavia
Pavia, Italy
francesco.leporati@unipv.it

***Abstract*—In recent years, hyperspectral imaging has been employed in several medical applications, targeting automatic diagnosis of different diseases. These images showed good performance in identifying different types of cancers. Among the methods used for classification, machine learning and deep learning techniques emerged as the most suitable algorithms to handle these data. In this paper, we propose a novel hyperspectral image classification architecture exploiting Vision Transformers. We validated the method on a real hyperspectral dataset containing 76 skin cancer images. Obtained results clearly highlight that the Vision Transforms are a suitable architecture for this task. Measured results outperform the state-of-the-art both in terms of false negative rates and of processing times. Finally, the attention mechanism is evaluated for the first time on medical hyperspectral images.**

Keywords—Vision Transformers, medical hyperspectral imaging, skin cancer, deep learning.

I. Introduction

Hyperspectral imaging (HSI) acquires information about a scene in the spatial and in the spectral domains. Thus, the data shapes a cube where a spectral vector is associated with each pixel. This spectral vector contains the fraction of incident electromagnetic radiation reflected upon a surface at a specific wavelength. Each material features a unique variation of reflectance compared to the wavelengths. This variation is called spectral signature and enables precise discrimination of different materials [1], including tissues [2]. A classification system targets the recognition of the material contained in each pixel. In the literature, several approaches have been proposed to classify hyperspectral images (HS), including Machine Learning (ML) [3]–[5] and Deep Learning (DL) [6]–[8] techniques.

Among the scientific fields exploiting HS classification, medicine emerged as one of the most promising. Two main research tracks arised from the literature. The first is the development of guidance tools to help the surgeon during surgical procedures [9], the latter concerns the diagnostic support [10].

Early diagnosis is of utmost importance in treating cancer. One of the most common cancer forms is skin cancer categorizable as non-melanoma skin cancer (NMSC) and melanoma. NMSC is extremely common, being the 5th most common form of cancer worldwide in 2018 [11]. Moreover, an extreme progression of melanocytes causes Pigmented Skin Lesions (PSLs) classifiable as malignant or benign. Finally, atypical moles or dysplastic nevi, are benign PSLs, associated with an increased risk of evolving to melanoma [12].

In the literature, HSI for skin cancer detection has been investigated in [10], where K-Means clustering and Support Vector Machine (SVM) classification were employed to discriminate between benign and malignant PSLs. Thus, the literature focuses on ML techniques for medical HS images classification. However, in recent years, DL networks emerged as the ideal solution for end-to-end classification tasks [13]. On the other hand, DL algorithms are mainly applied to HSI related to remote sensing applications. Thus, to the best of the

2771-2508/22/$31.00 ©2022 IEEE
DOI 10.1109/DSD57027.2022.00122

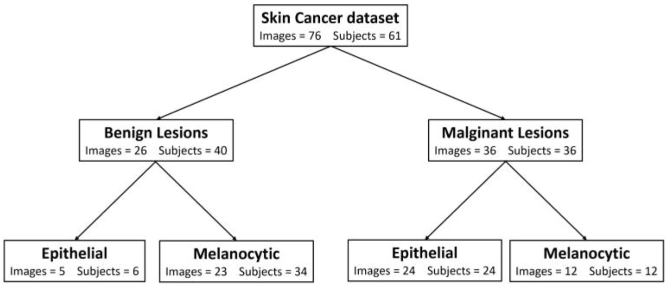
Skin Cancer dataset
Images = 76 Subjects = 61
Benign Lesions
Images = 26 Subjects = 40
Malginant Lesions
Images = 36 Subjects = 36
Epithelial
Images = 5 Subjects = 6
Melanocytic
Images = 23 Subjects = 34
Epithelial
Images = 24 Subjects = 24
Melanocytic
Images = 12 Subjects = 12

Figure 1 – Classification taxonomy and images distribution.

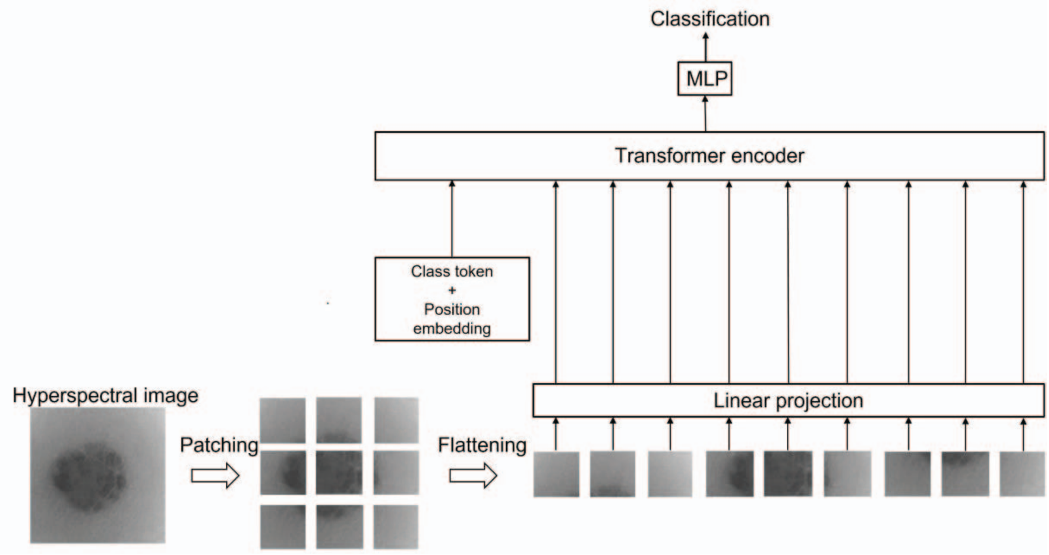

Figure 2 – Architecture of a Vision Transformer.

authors' knowledge, DL architectures are not adopted for HS medical images classification till now.

Among the different DL methods, Vision Transformers (ViT) have recently appeared in literature [14]. These networks rely on the self-attention mechanism, at first designed for Natural Language Processing (NLP) applications and possess a very high number of parameters.

In this paper, we propose a ViT-based classifier targeting HS skin cancer images. The paper is organized as follows. Section II introduces the HS dataset and the adopted attention-based architecture. Moreover, it details the performance evaluation methodology and the adopted metrics. Section III describes the obtained results and compares them with the state-of-the-art. Section IV concludes the paper and suggests possible future research lines.

II. Materials and Methods

A. Skin cancer Hyperspectral Dataset

The HS dataset of skin cancer in-vivo samples includes 76 images obtained from 61 subjects. 46 images contain malignant lesions, while 30 represent benign skin cancer. These images were acquired by the system described in [15], which is based on a snapshot HS camera (Cubert UHD 185, Cubert GmbH, Ulm, Germany) coupled to a Cinegon 1.9/10 (Schneider Optics Inc., Hauppauge, NY, USA) lens with an F-number of 1.9 and a focal length of 10.4 nm. The illumination system (Dolan-Jenner, Boxborough, MA, USA) employs a 150 W QTH (Quartz-Tungsten Halogen) lamp coupled to an optic fiber ring light guide to obtain cold light emission.

Each acquired image contains 125 spectral bands covering the visual and near-infrared (VNIR) spectral range from 450 to 950 nm, having a spatial resolution of 50 × 50 pixels (pixel size of 240 × 240 μm). All the images have been labelled using the tool described in [16]. The labelling procedure consists of four different classes, namely Benign Epithelial (BE), Benign Melanocytic (BM), Malignant Epithelial (ME) and Malignant Melanocytic (MM). Figure 1 shows the taxonomy of the classification.

A further data processing aimed to standardize the spectral signature of each pixel. This calibration exploits two reference images acquired before the dataset collection. Namely, a white reference image (WI) was acquired, captured from a white reference tile able to reflect 99% of the incident light, and a dark reference image (DI), recorded when the light was turned off and the camera shutter was closed. Eq. (1) provides the calibrated image (CI) is obtained from the raw HS image (RI):

$$CI = \frac{RI - DI}{WI - DI} \qquad (1)$$

Then, we removed the first four and the last five bands due to the low spectral response of the HS sensor. The spectral noise has been further reduced by applying a smoothing filter based on a moving average algorithm with a window of 5. Finally, each spectral signature was normalized in the range [0,1] with the min-max procedure. Thus, the pre-processed image contains 116 spectral bands for each pixel.

It is worth noticing that this dataset includes 76 images, which are not sufficient to train any DL model. Thus, we increased the dataset size by applying data augmentation. This procedure includes geometric transformation, filtering, random centre cropping, colour transformations and pixel substitution. Random pixels of tumours of the same category undergo bilinear interpolation or are directly exchanged. This procedure also applies to the skin pixels. Furthermore, the training set was enlarged by introducing salt-and-pepper white noise. The augmentation procedure was carried out iteratively. One of the data augmentation techniques was applied to the training set. Then, we created a new data cluster by unifying the original images and the augmented images and applied a second technique to the new group. Finally, this procedure was recursively applied to broaden the training set exponentially. We did not apply such augmentation techniques to neither the validation nor the test sets to prevent the results to be biased.

B. Vision Transformers for Hyperspectral Imaging

Vision transformers (ViT) are DL architectures based on the self-attention mechanism [14]. Figure 2 shows the structure of a ViT. Typically, a ViT receives as input a 1-D array. Thus, N-D data, such as HS images are transformed into 1-D arrays by dividing the original image into patches of the same dimension. This partitioning is performed through a convolution operation. Let $X \in \mathbb{R}^{H \times W \times C}$ be a HS image with a spatial dimension of $H \times W$ and C spectral channels. Each patch is denoted with $X_p \in \mathbb{R}^{H \times P \times P \times C}$ where $P \times P$ is the resolution of a single patch. Thus, the number of patches, of which an image consists, is $N = HW/P^2$. The ViT uses a Q-D array of latent variables to project the patches in a new space. Then, a class token is associated to each patch, together with an array containing information about the relative position of each patch with respect to the original image (position embedding). These data represent the input to the Transformer encoder, based on three main components: Multi-head Self Attention (MSA), Multi-Layer Perceptron (MLP) and normalization. The components are connected as shown in Figure 3.

In the self-attention mechanism, each input vector is projected to generate three vectors: Key (K), Query (Q) and Value (V). For each input vector, the attention map is computed according to eq. 2:

$$Attention(K, Q, V) = softmax\left(\frac{QK^T}{\sqrt{d}}\right)V \qquad (2)$$

where $\sqrt{d}$ is a normalization term.

The Multi-head attention mechanism is based on eq. (2), but the main difference is that the Q, K and V vectors are linearly projected into a suitable space and then, in parallel, the attention mechanism is applied to the new vectors. Then, the attention values are concatenated to obtain the output. Typically, L MSL layers are concatenated to produce the input for the MLP that generates the final classification.

HS images feature a higher number of channels than standard RGB images. Thus, the number of multiplications performed by the MSA layer is very high. This issue can be solved by introducing convolution operations before applying the patching procedures. In particular, the solution proposed in this paper is to use three convolutional layers, each one featuring a 2-D convolutional layer, a normalization and a ReLU activation function. Each convolutional layer is based on 3 × 3 filters. The number of filters is 58, 29 and 14 for the first, the second and the last convolutional layer, respectively.

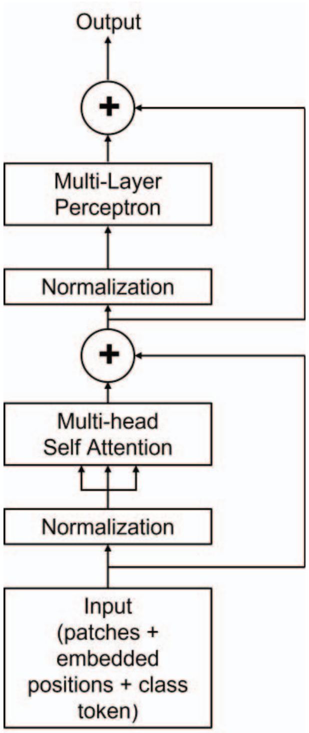
Output
Multi-Layer Perceptron
Normalization
Multi-head Self Attention
Normalization
Input (patches + embedded positions + class token)

Figure 3 – The architecture of a transformer encoder.

These layers reduced the channels from 116 to 14. Therefore, the size of the image given as input to the ViT is 50 × 50 × 14.

The proposed strategy reduces both the computational complexity and the memory occupancy of the ViT architecture compared to giving as input the original HS image.

C. Performance Metrics

The ViT was trained to classify the lesions into four categories, namely malignant melanocytic, benign melanocytic, malignant epithelial and benign epithelial.

Since the number of samples included in the original dataset is limited, K- fold cross-validation is adopted. This statistical method produces metrics estimations offering a lower bias than other techniques. This method features a single parameter called *k*, which refers to the number of groups in which the data sample is split. Mainly the cross-validation technique is used in applied machine learning to estimate the performance of a model on unseen data, which was not used during the model training.

The original HS dataset comprising 76 images is split into k groups. Next, each unique group was selected as test data and the model was trained on the remaining groups. Thus, data contained in the groups used for training were augmented, as described in Section II.A. The model was fit on the training set and was evaluated on the test set, retaining the prediction evaluated at each iteration and discarding the model. In this work, k is set equal to 10. Therefore, the model is trained k times, and the estimations are stored for each test set. Hence, the performance metrics were assessed on the aggregated group of predictions.

We evaluated the classification performance in terms of accuracy and specificity, defined according to eqs. (3) and (4):

$$accuracy = \frac{TP+TN}{TP+FP+TN+FN} \quad (3)$$

$$specificity = \frac{TN}{TN+FP} \quad (4)$$

where TN , TP , FN and FP indicate the true negative, the true positive, the false negative and the false positive.

Another metric adopted for evaluating the classification is the False Negative Rate per class (FNRc), defined by eq. (5):

$$FNRc = \frac{FN_i}{P} \quad (5)$$

where FN_i is the false negative rate of the $i-th$ class and P is the total number of positive predicted samples.

III. Experimental Results

The ViT architecture described in Section II.B has been implemented in MATLAB 2020a, by writing custom scripts exploiting the Deep Learning Toolbox. The code runs on a PC equipped with an Intel i9 9900C CPU processor working at 3.5 GHz and featuring 128 GB of DDR4 RAM memory. The PC is also equipped with two NVIDIA RTX 2080 GPUs, each one featuring 2944 cores working at 1.8 GHz and with 8 GB of DDR6 RAM memory.

The network was trained and fine-tuned with a small-size dataset. Then, the performance is evaluated employing a K-fold cross-validation methodology considering K set at 10. Moreover, the taxonomy proposed in Figure 1 was adopted as a trade-off between being medically relevant, complete, and well-suited for DL classifiers. Indeed, the considered tree-structure categorization allows to treat patients according to the highest healthcare standards and provides the best achievable classification [17]. Notice that training data were augmented adopting the techniques described in Section II.

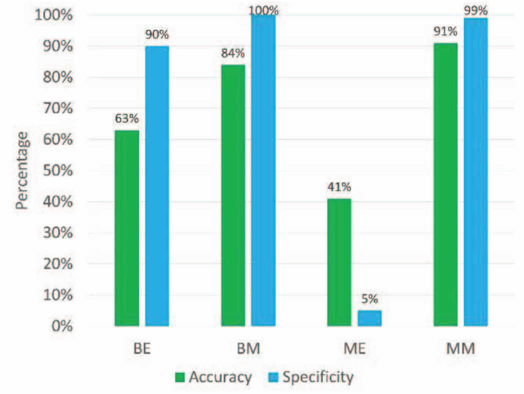

Figure 4- Performance of the proposed ViT. Accuracy and Specificity are reported as percentages. BE, BM, ME and MM represent Benign Epithelia, Benign Melanocytic, Malignant Epithelial and Malignant Melanocytic, respectively.

Figure 4 shows the performance computed on the aggregated predictions of the K-fold cross validation technique. This chart clearly shows that the proposed architecture is capable of classifying the benign and malignant melanocytic lesions with high accuracy and specificity. On the other hand, considering the BE class, the network features a high specificity, but the accuracy is around 60 %. Therefore, the low number of images contained in the original database labelled as BE is not enough to efficiently train the proposed network.

To the best of the authors' knowledge, HS imaging for skin cancer detection was investigated in [10], [15]. Both the works rely on the same processing chain exploiting K-Means clustering and SVM classification. It is important to highlight that the classification taxonomy adopted in these works is not the same considered in the proposed research. Thus, a direct comparison is not fair and can be carried out only in terms of FNRc. The work in [10] computed the FNRc for 18 images, obtaining values up to 60%. In the proposed work, the metric is computed adopting the K-fold cross validation, being more robust and reliable than the values reported in the state-of-the-art. The ViT obtained FNRc values ranging from 6% to 30%. Therefore, the proposed attention-based network represents an interesting and promising solution for the skin cancer detection in HS images.

We characterized the performance of the network also in terms of processing time measured considering the classification of 100 images and computing the mean processing time and its standard deviation. The mean processing time is equal to 65.2 ms, with a standard deviation of 7.5 ms. The system proposed in [15] and parallelized in [10] takes variable processing times, ranging from 350.0 ms to 2.06 s. Hence, the proposed work outperforms the state-of-the-art in terms of processing speed. Moreover, the variability of the processing time featured by this work is significantly lower than the state-of-the-art. Indeed, the ViT architecture has a fixed number of layers that perform a fixed number of operations. On the other hand, the processing chain proposed in literature includes the K-Means clustering, which iterates the operations based on the clustering error. Thus, the number of iterations performed by this method is not deterministic and strictly depends on the initial values of cluster centroids.

IV. Conclusion

In this paper, we proposed a novel attention-based network to classify skin cancer exploiting HS images. The proposed network is designed and validated using a real HS dataset, adopting the K-fold cross validation technique to produce robust results.

Since the original dataset featured only 76 image, we applied data augmentations to the real data. Performed augmentations included geometrical transformations, filtering, random centre cropping, colour transformations, pixel substitution and random addition of gaussian white noise.

The model was trained augmenting at runtime the training set and then performing the tests only on the real imaging, considering a number of folds equals to 10.

The obtained results clearly highlight that the attention-based mechanism is an interesting and promising solution for medical HS images classification, since the false negative rate is half compared to the state-of-the-art.

Moreover, the classification times are significantly lower than the best solutions proposed in the literature. Finally, the proposed network adopts a fixed number of layers while the number of operations is deterministic, making the measured processing time more stable than the results reported in the literature.

Future research will focus on improving the proposed network, evaluating different configuration of the layers.

Acknowledgment

This work has been 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. Additionally, this work was completed while Raquel Leon was beneficiary of a 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%)).

References

[1] J. M. Meyer, R. F. Kokaly, and E. Holley, “Hyperspectral remote sensing of white mica: A review of imaging and point-based spectrometer studies for mineral resources, with spectrometer design considerations,” *Remote Sensing of Environment*, vol. 275, p. 113000, Jun. 2022, doi: 10.1016/J.RSE.2022.113000.

[2] A. ul Rehman and S. A. Qureshi, “A review of the medical hyperspectral imaging systems and unmixing algorithms’ in biological tissues,” *Photodiagnosis and Photodynamic Therapy*, vol. 33, Mar. 2021, doi: 10.1016/J.PDPDT.2020.102165.

[3] D. Saha and A. Manickavasagan, “Machine learning techniques for analysis of hyperspectral images to determine quality of food products: A review,” *Current Research in Food Science*, vol. 4, pp. 28–44, Jan. 2021, doi: 10.1016/J.CRFS.2021.01.002.

[4] G. P. Petropoulos, K. Arvanitis, and N. Sigrimis, “Hyperion hyperspectral imagery analysis combined with machine learning classifiers for land use/cover mapping,” *Expert Systems with Applications*, vol. 39, no. 3, pp. 3800–3809, Feb. 2012, doi: 10.1016/J.ESWA.2011.09.083.

[5] R. Lazcano *et al.*, “Parallel Implementations Assessment of a Spatial-Spectral Classifier for Hyperspectral Clinical Applications,” *IEEE Access*, vol. 7, pp. 152316–152333, 2019, doi: 10.1109/ACCESS.2019.2938708.

[6] X. Yang, Y. Ye, X. Li, R. Y. K. Lau, X. Zhang, and X. Huang, “Hyperspectral image classification with deep learning models,” *IEEE Transactions on Geoscience and Remote Sensing*, vol. 56, no. 9, pp. 5408–5423, Sep. 2018, doi: 10.1109/TGRS.2018.2815613.

[7] S. Li, W. Song, L. Fang, Y. Chen, P. Ghamisi, and J. A. Benediktsson, “Deep learning for hyperspectral image classification: An overview,” *IEEE Transactions on Geoscience and Remote Sensing*, vol. 57, no. 9, pp. 6690–6709, Sep. 2019, doi: 10.1109/TGRS.2019.2907932.

[8] A. Signoroni, M. Savardi, A. Baronio, and S. Benini, “Deep Learning Meets Hyperspectral Image Analysis: A Multidisciplinary Review,” *Journal of Imaging 2019, Vol. 5, Page 52*, vol. 5, no. 5, p. 52, May 2019, doi: 10.3390/JIMAGING5050052.

[9] G. Florimbi *et al.*, “Towards Real-Time Computing of Intraoperative Hyperspectral Imaging for Brain Cancer Detection Using Multi-GPU Platforms,” *IEEE Access*, vol. 8, pp. 8485–8501, 2020, doi: 10.1109/ACCESS.2020.2963939.

[10] E. Torti *et al.*, “Parallel Classification Pipelines for Skin Cancer Detection Exploiting Hyperspectral Imaging on Hybrid Systems,” *Electronics 2020, Vol. 9, Page 1503*, vol. 9, no. 9, p. 1503, Sep. 2020, doi: 10.3390/ELECTRONICS9091503.

[11] F. Bray, J. Ferlay, I. Soerjomataram, R. L. Siegel, L. A. Torre, and A. Jemal, “Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries,” *CA: A Cancer Journal for Clinicians*, vol. 68, no. 6, pp. 394–424, Nov. 2018, doi: 10.3322/CAAC.21492.

[12] A. Perkins and R. L. Duffy, “Atypical moles: diagnosis and management.,” *Am Fam Physician*, vol. 91, no. 11, pp. 762–7, Jun. 2015.

[13] X. Hu *et al.*, “Hyperspectral Anomaly Detection Using Deep Learning: A Review,” *Remote Sensing*, vol. 14, no. 9, p. 1973, Apr. 2022, doi: 10.3390/RS14091973.

[14] A. Kolesnikov *et al.*, “An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale,” 2021.

[15] H. Fabelo *et al.*, “Dermatologic Hyperspectral Imaging System for Skin Cancer Diagnosis Assistance,” *2019 34th Conference on Design of Circuits and Integrated Systems, DCIS 2019*, Nov. 2019, doi: 10.1109/DCIS201949030.2019.8959869.

[16] H. Fabelo *et al.*, “In-Vivo Hyperspectral Human Brain Image Database for Brain Cancer Detection,” *IEEE Access*, vol. 7, pp. 39098–39116, 2019, doi: 10.1109/ACCESS.2019.2904788.

[17] Y. Fujisawa *et al.*, “Deep-learning-based, computer-aided classifier developed with a small dataset of clinical images surpasses board-certified dermatologists in skin tumour diagnosis,” *Br J Dermatol*, vol. 180, no. 2, pp. 373–381, Feb. 2019, doi: 10.1111/BJD.16924.