LiwTERM: A Lightweight Transformer-based Model for Dermatological Multimodal Lesion Detection

Luis A. Souza Jr., André G. C. Pacheco,
Gabriel G. de Angelo, Thiago Oliveira-Santos
Federal University of Espírito Santo
Graduate Program of Informatics
Vitória, Brazil
Email: {la.souza, apacheco,todsantos}@inf.ufes.br

Christoph Palm
OTH Regensburg
Regensburg Medical Image Computing (ReMIC)
Regensburg, Germany
Email: christoph.palm@oth-regensburg.de

João P. Papa
São Paulo State Univesity
Department of Computing
São Paulo, Brazil
Email: joao.papa@unesp.br

***Abstract*—Skin cancer is the most common type of cancer in the world, accounting for approximately 30% of all diagnosed tumors. Early diagnosis reduces mortality rates and prevents disfiguring effects in different body regions. In recent years, machine learning techniques, particularly deep learning, have shown promising results in this task, presenting studies that have demonstrated that combining a patient's clinical information with images of the lesion is crucial for improving the classification of skin lesions. Despite that, meaningful use of clinical information with multiple images is mandatory, requiring further investigation. Thus, this project aims to contribute to developing multimodal machine learning-based models to cope with the skin lesion classification task employing a lightweight transformer model. As a main hypothesis, models can take multiple images from different sources as input, along with clinical information from the patient's history, leading to a more reliable diagnosis. Our model deals with the not-trivial task of combining images and clinical information (from anamneses) concerning the skin lesions in a lightweight transformer architecture that does not demand high computation resources but still presents competitive classification results.**

***Index Terms*—Deep learning, Skin Lesion Detection, Transformers, Lightweight Architectures.**

I. Introduction

Skin cancer is the most common dysplasia in the world. The World's Health Organization (WHO) estimates that skin cancer accounts for approximately 30% of all types of cancer diagnosed worldwide [1]. The National Cancer Institute [2] estimates that for the period 2023-2025 there will be 220 thousand new cases of skin cancer, a number that makes it the most common type of cancer in the country, with 31.2% of all types [2]. Even being the most common one, the skin cancer's mortality rate is low, around 1% if early diagnosed [2]. Late diagnosis is the main factor contributing to the rise in this mortality rate. However, even if it does not lead to death, the tumor can leave significant mutilations on the skin – for example, removing part of the nose – if the lesion does not receive appropriate prognosis and treatment.

For clinical diagnosis of skin cancer, dermatologists perform a visual examination of the potential lesion and consider the patient's medical history, relying on their expertise to make an accurate assessment. This process is challenging and needs specialized training and experience in dermoscopy. Kittler et al. [3] and Sinz et al. [4] have shown that dermoscopy greatly enhances diagnostic accuracy, although its effectiveness is heavily influenced by the dermatologist's level of experience. Furthermore, the high workload of professionals and human factors such as fatigue, stress, and emotional issues can momentarily harm diagnostic capacity, especially when tracking early-staged anomalies. The situation becomes even more serious in peripheral and rural regions due to limited access to experts and specialized equipment. Thus, taking into account the high incidence rate of skin cancer and the lack of required resources, especially in rural areas [5] and emerging countries [6], the development of Computer Aided Diagnosis (CAD) systems for skin cancer detection becomes a highly desirable technology to increase effectiveness and speed up clinical diagnosis in healthcare systems.

In recent years, the use of CAD systems to assist in skin lesion analysis has been intensely investigated [7]–[9]. The most successfully employed strategy, in terms of performance, has been the use of machine learning with deep learning [10]–[12]. Despite the promising results in the area, several challenges need to be overcome to enable the implementation of such technology in a safer and more satisfactory manner. Among the challenges, the following stand out: uncertainty in the data, biases, datasets with a low number of samples, low generalization of models, and low explainability of predictions [12]. To diminish some of these problems, it was proposed to use images and clinical data from anamnesis to classify skin cancer [13]–[15].

Most works cited earlier deal with skin lesion identification using deep architectures, which are usually costly. A few works proposed lightweight architectures in the context of neoplasia identification. Hou et al. [16] introduced an approach for early neoplasia identification in Barrett's esophagus-diagnosed samples using attentive hierarchical aggregation and self-distillation. Their work employs a SE-ResNet50 as the back-

bone, a variation of the well-known ResNet50 with squeeze-and-excitation modules. The authors reported promising results concerning their method's efficiency.

In the skin cancer detection context, Tuncer et al. [17] proposed a lightweight model based on Convolution Neural Networks (CNN) architectures to classify dermatoscopical skin lesion images between benign and malignant. Something similar has been proposed and conducted by Li et al. [18], where a lightweight CNN-based model is proposed to deal with the classification of 8 different skin lesions based on dermatoscopic databases. As one can observe, most current investigations focused on using images in their lightweight approaches, not combining their descriptions with any clinical information and mostly focusing on CNN-based variations.

It is well-known within the ML field that combining features extracted from images with other features obtained from different sources describes a common problem, i.e., the image is the main source of information and the extra data – hereby defined as metadata – provides supplementary information about the problem. But the question remains: how to provide such a combination? Kharazmi et al. [19] proposed a feature fusion system based on concatenation and Sparse Autoencoder (SAE) to detect Basal Cell Carcinoma in skin tissues, and Sierra et al. [20] also used concatenation to combine image features, extracted using two CNN architectures, with textual metadata to predict gender. Recently, Pacheco and Krohling [15] conducted a similar work to predict six different skin lesions by imposing transformations to image features calculated within CNN models based on the metadata.

Hence, this work proposes LiwTERM, a lightweight neural architecture that combines features learned by (i) a Vision Transformer (ViT) [21] and (ii) a language-processing Tokenizer into a shallow and fully connected model to distinguish among six different skin lesions from clinical images. We report competitive results with high efficiency and low computational cost. To the best of our knowledge, this is the first time transformers have been employed to combine images and text for skin cancer description and generalization.

The remainder of this paper is organized as follows. Sections II and III introduce the proposed approach and the methodology, respectively. Section IV presents the experiments, and Section V discusses the outcomes. Last but not least, Section VI states conclusions and future works.

II. Proposed Approach - LiwTERM

This work proposes a lightweight model that combines features from pre-trained ViTs and text-based tokenizers without incurring a high computational cost. ViT models, a generalization of Transformers for image-based tasks such as identification, description, and classification, are robust tools for solving image-driven problems. The shortcomings of such models concern the fine-tuning process for some specific tasks, which is harmed by the high training computational requirements. Conversely, the inference process of pre-trained ViT models does not demand the same computational cost, providing a powerful image representation.

With the current advances in Generative Pre-trained Transformers (GPTs), the text-processing field has shown accountable progress in the past few years, with models that can properly encode-decode text and represent intrinsic features such as context and word positioning - very difficult tasks to be accomplished. As the first Natural Language Processing (NLP) step [22], the tokenization process copes with the decomposition of a sentence to be further consumed as tokens. Tokens are the basic description units of text in the NLP field, accompanied by positioning and delimiters to compose the word description [22], bridging raw text and context for language models (LMs) [23]. Some current LM, also known as a sequence-to-sequence models, use the tokenization process in their generalization and are built with transformer encoder-decoder designs, with bidirectional encoders (BERT-like [24]) and autoregressive (GPT-like) decoders, aiming to receive sentences and give as output, also sentences. At the halfway of the processing, text features, or embeddings, can be obtained to represent the original input sentence as a high-encoded information.

The fine-tuning process for transformer-based models in general, including the ones that deal with text or images (ViT specifically), demands a huge amount of samples to perform well. However, considering the current scenario for this kind of architecture, already pre-trained checkpoints can provide powerful feature inference for a wide range of contexts, benefiting the description for other simpler model training. This enables simpler setups to perform training and inference processes using the feature generalization from transformers and using a reduced amount of available samples for such a task, something recurrent in the medical field.

Also, using transformer-based architectures to process both image and text may lead to the representation of two different representations to a third similar (if not equal) domain. The ViT representation proposes an image serialization that considers patch positioning, tokenization, and linear projections, which can be easily compared to the tokenization process performed by several NLP approaches. Due to this similarity in representation, the combination of image and text representations in a common environment can be designed, enabling the use of both feature descriptions to complement each other.

Hence, we propose LiwTERM, a hybrid-and-shallow transformer-based model in which image and text features feed a fine-tuned encoder to distinguish between six skin lesions and take advantage of both representation methods. The embeddings of ViT and LM models are combined by full connection transformations from the second last layer of each architecture, feeding an encoder composed of four fully connected layers accompanied by ReLU, batch normalization, and dropout transformations, hereby named shallow lightweight model (SLM). Finally, a SoftMax head defines the classes for the final classification task. Notice that the fine-tuning is performed only from the fully connected layer of the ViT and tokenizer features, configuring a shallow training

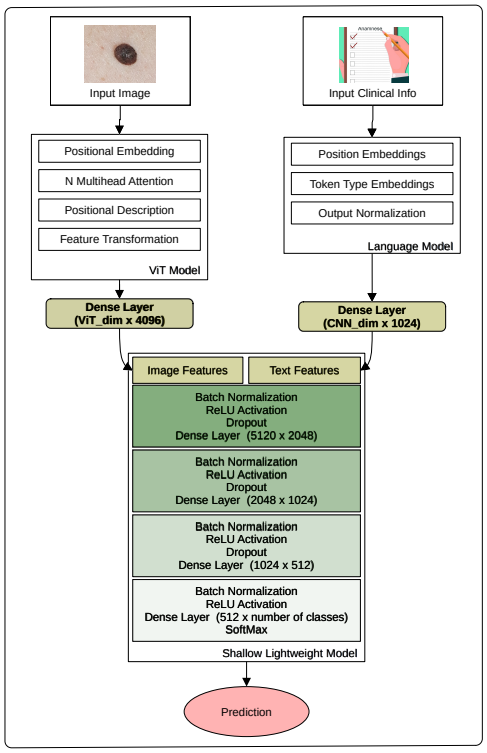


Fig. 1: LiwTERM pipeline: the proposed model has two sections: (i) the feature extraction (with no color) and (ii) the shallow lightweight model section (with colors). The colored section concerns the trainable part of the proposed method; fed from the deep and complex ViT and tokenizer architectures, this section is in charge of learning the proper weights to provide the classification of skin lesions based on images and clinical information.

phase. Figure 1 illustrates the LiwTERM pipeline [1].

III. Methodology

A. Datasets

The proposed method has been evaluated over two public datasets named (i) PAD-UFES-20 [25] and (ii) ISIC 2019 [26]. The PAD-UFES-20 dataset is composed of clinical skin lesion images collected in the Espírito Santo State, in Brazil, along with clinical information of each patient and lesion. In total, 2.298 clinical images were collected from smartphone devices, 21 patient clinical features – such as age, gender, anatomical region, cancer history, skin prototype, family background, among others – and six different skin lesions, Basal Cell Carcinoma (BCC), Squamous Cell Carcinoma (SCC), Actinic Keratosis (ACK), Seborrheic Keratosis (SEK), Melanoma (MEL), and Nevus (NEV). The ISIC 2019 is a skin lesion dataset which comprises 25,331 public dermoscopy images, three clinical features collected from anamneses, e.g., age, gender, and anatomical region, and eight skin lesions, Melanoma (MEL), Melanocytic Nevus (NEV), Basal Cell Carcinoma (BCC), Actinic Keratosis (AK), Benign Keratosis (BKL), Dermatofibroma (DF), Vascular Lesion (VASC), and Squamous Cell Carcinoma (SCC). Figures 2 and 3 illustrates examples of PAD-UFES-20 and ISIC 2019 dataset classes, respectively.

[1]The LiwTERM code repository is publicly available at https://github.com/luisouza/liwterm/.

B. Evaluation Measures

We employed three well-known quantitative measures to evaluate the proposed approach: Sensitivity (S), Specificity (P), and Balanced Accuracy (BACC). The experimental setup also comprises a statistical evaluation using Wilcoxon's signed-rank test [27] with a significance level of 5%.

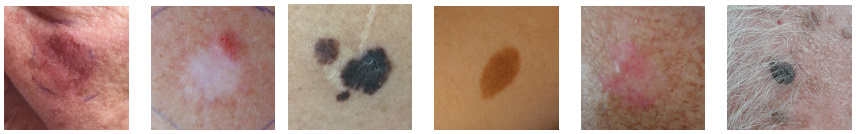
C. Experimental Delineation

Three experimental approaches evaluate the robustness of LiwTERM: (i) a 5-fold cross-validation using only the ViT bottleneck for the feature inference, (ii) a 5-fold cross-validation using only the text-tokenization bottleneck for the feature inference, and (iii) a 5-fold cross-validation employing the entire method, with feature generalization from both images and clinical information. Additionally, the baseline results for each designed approach, i.e., (iv) the classification of cancerous skin tissue only based on the pre-trained ViT architecture, and (v) the skin lesion classification based only on LM with the correspondent checkpoint, both designs avoiding the shallow lightweight training portion, were conducted for the sake of comparison. All the experimental designs were trained over 65 epochs with a batch size of 24. All five experimental folds were constructed based on a class-stratification fashion, balancing the amount of samples of each skin lesion class for each fold.

As LiwTERM is based on pre-trained ViT and LM, for the feature calculation step, we had their weights frozen, keeping the configuration of the pre-trained states ("google/vit-large-patch16-224" and "facebook/bart-base," respectively, and both from HuggingFace). The shallow lightweight model weights (Figure 1 - color) have been started from scratch, with a scheduling learning rate from $1e^{-3}$ to $1e^{-6}$. All parameters were selected empirically based on multiple experiments.

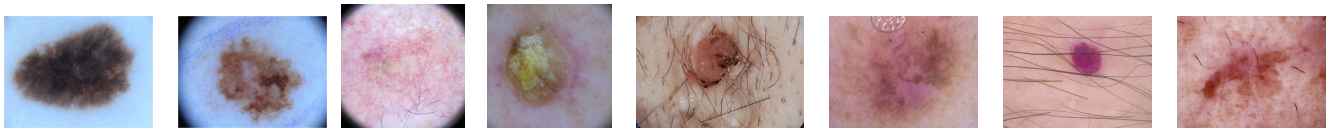
D. Implementation Details

The experiments employed a computer with 16 GB RAM and an NVIDIA RTX®3070 Graphics card of 8 GB VRAM. The implementation used the Pytorch framework. It is essential to highlight that the proposed model is designed to cope with the drawback of high computational costs imposed by transformers. As one can observe, a simple computer configuration has been set for the experimental step, illustrating that our model does not require high-end computer configurations. The proposed model was assessed over GPU and CPU sets, with the same performance outcomes (despite the longer training time).



(a) ACK (b) BCC (c) MEL (d) NEV (e) SCC (f) SEK

Fig. 2: PAD-UFES-20 dataset samples.



(a) NEV (b) MEL (c) BCC (d) AK (e) BKL (f) DF (g) VASC (h) SCC

Fig. 3: ISIC 2019 dataset samples.

TABLE I: LiwTERM classification results using 5-fold validation protocol over the PAD-UFES-20 and ISIC 2019 datasets. Bold lines mean the overall best-obtained outcome.

LiwTERM's Backbone	Dataset	Composition	S	P	BACC
ViT	PAD-UFES-20	ViT+SLM	0.51 ± 0.02	0.68 ± 0.02	0.63 ± 0.02
		baseline	0.44 ± 0.03	0.48 ± 0.06	0.46 ± 0.04
	ISIC 2019	ViT+SLM	0.47 ± 0.06	0.55 ± 0.08	0.52 ± 0.02
		baseline	0.42 ± 0.05	0.46 ± 0.07	0.44 ± 0.06
LM	PAD-UFES-20	LLM+SLM	0.61 ± 0.02	0.68 ± 0.03	0.65 ± 0.03
		baseline	0.47 ± 0.05	0.56 ± 0.07	0.51 ± 0.06
	ISIC 2019	LLM+SLM	0.57 ± 0.04	0.59 ± 0.03	0.57 ± 0.05
		baseline	0.43 ± 0.07	0.50 ± 0.07	0.46 ± 0.06
ViT + LM	PAD-UFES-20	ViT+LLM+SLM	0.69 ± 0.02	0.76 ± 0.02	0.74 ± 0.01
	ISIC 2019	ViT+LLM+SLM	0.66 ± 0.02	0.77 ± 0.03	0.73 ± 0.03

IV. Experimental Results

LiwTERM focuses on three main aspects: the correct classification of six skin lesions, the computational cost required for the fine-tuning process (already presented in the last section), and the time consumption for the model's training process. Table I presents the model classification results on the PAD-UFES-20 and ISIC 2019 datasets for all the evaluated approaches, highlighting the impact of each selected backbone generalization for the feature-extraction composition performed for the features. For comparison purposes, Table I also presents the baseline results of the evaluated backbones proposed for LiwTERM model. Figures 4a and 4b illustrate the overall confusion matrices of complete LiwTERM model over PAD-UFES-20 and ISIC 2019 datasets, respectively.

V. Discussion

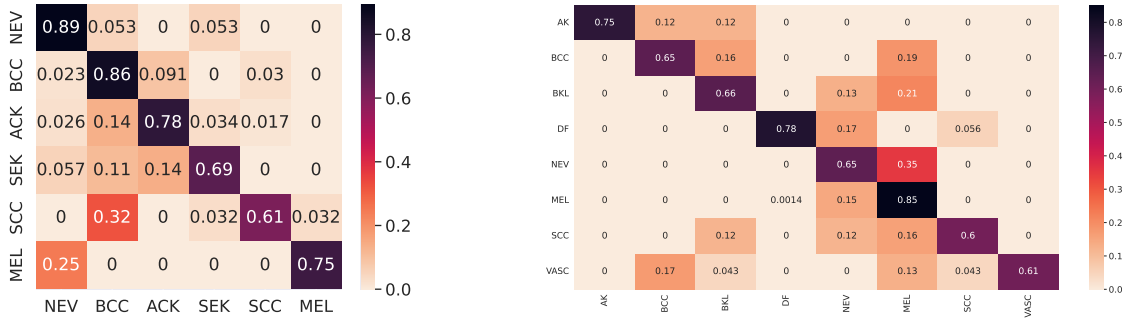
A. LiwTERM's Backbone Analysis

As one can observe in Table I, the obtained results clearly highlight the efficiency of our method, where the feature-encoded information provided by ViT and LM models could complement each other and enhance the correct prediction of skin lesions (Fig. 4). Using only the pre-trained ViT or LM models for predicting the skin lesions is not enough (baselines), so the fine-tuning must be conducted to make such a classification feasible. The introduction of the trainable shallow lightweight portion to the model could enhance the prediction results, leading to the best ones when both ViT and LM models work together for the feature generalization of LiwTERM. Additionally, no statistical similarity was found between the best results, i.e., the ones employing ViT and LM features statistically outperformed all the other experimental designs, including the baselines.

A benchmark evaluation of LiwTERM approaches was conducted, focusing on memory usage. The baseline models, using only ViT and LLM backbones, have approximately 60M and 0.79M parameters, consuming around 200MB and 4MB of memory, respectively. Adding the SLM portion increased these to about 63M parameters and 260MB for ViT, and 1.45M parameters and 15MB for LLM, showing that the additional layers slightly impacted the original feature extraction. The final LiwTERM model, which integrates both ViT and LLM with the SLM module, has approximately 66M parameters and uses 300MB of memory, with accuracy improvements detailed in Table I.

We also compared the proposed approaches in terms of training and inference time. For PAD-UFES-20, training times were approximately 4.06h for ViT+SLM, 3.89h for baseline ViT, 1.87h for LLM+SLM, 1.59h for baseline LLM, and 5.15h for the full LiwTERM model. For ISIC19, the times were 6.10h for ViT+SLM, 5.75h for baseline ViT, 2.17h for LLM+SLM, 2.00h for baseline LLM, and 6.34h for LiwTERM. Inference times for all approaches are quite small, ranging between $10^{-4}sec$ and $10^{-6}sec$ per sample. The SLM module adds minimal overhead to ViT and LLM models in terms of parameters, memory, or training/inference times, making it feasible to use even with CPU configurations – though GPUs are preferable for reducing training time.

Table II presents a comparison between our method and the ones reported by Pacheco and Krohling [15], which employed the same evaluation protocol as described in this work. As one can observe, our method presents competitive performance to the CNN-based approaches. The Concatenation, Metablock, and MetaNet methods process the textual information using one-hot encoding [15]. Although it is a simple and efficient approach, it fails, for example, to handle out-of-vocabulary words



(a) LiwTERM (ViT + LM) - PAD-UFES-20

(b) LiwTERM (ViT + LM) - ISIC 2019

Fig. 4: Overall confusion matrices (original labels vs. predicted labels) for (a) complete LiwTERM design on PAD-UFES-20 dataset, and (b) complete LiwTERM design on ISIC 2019 dataset.

TABLE II: Comparison of LiwTERM with state-of-the-art works. Statistical similarity concerning PAD-UFES-20 dataset is presented in bold, while statistical similarity found for ISIC19 dataset among the models is underlined.

Dataset	Model	Design	BACC
PAD-UFES-20	**LiwTERM**	Lightweight Transformer-based	**0.74 ± 0.01**
	[15]	CNNs	0.65 ± 0.02
	[15]	**CNNs + Concatenation**	**0.76 ± 0.01**
	[15]	**CNNs + MetaBlock**	**0.77 ± 0.02**
	[15]	CNNs + MetaNet	0.75 ± 0.03
ISIC 2019	LIWTERM	Lightweight Transformer-based	0.73 ± 0.03
	[15]	CNNs	0.75 ± 0.04
	[15]	CNNs + Concatenation	0.77 ± 0.02
	[15]	CNNs + MetaBlock	0.77 ± 0.01
	[15]	CNNs + MetaNet	0.76 ± 0.01

or/and missing words, which are common issues in medical anamnesis. Our method, on the other hand, uses a transformer-based architecture to process the textual information, which is much more robust to these issues. We also carried out a statistical analysis to compare the methods, and the results show that our method is statistically equivalent to the CNN-based approaches in terms of balanced accuracy (as one can observe in Table II). It is important to note that these similar results are achieved using the same amount of data, which is relevant for a transformer-based architecture, as it is known to be data-hungrier than CNN-based architectures.

B. LiwTERM Strengths and Limitations

We propose a lightweight training model that combines two pre-trained deep backbones (ViT and LM architectures) with a shallow, trainable neural block called LiwTERM. Our approach addresses the challenge of limited resources for training deep architectures. Unlike traditional methods, LiwTERM requires less computational power since it only trains the final embedding calculations. Additionally, by focusing on the feature description layers and the shallow lightweight block, LiwTERM delivers competitive results even with a reduced amount of training data.

Using two backbones significantly enhances the dimensionality of skin lesion descriptions by combining features from different domains (images and text). Data availability is a key challenge in medical applications, especially with sensitive data requiring legal permissions. LiwTERM leverages deep representations without needing to fine-tune bottleneck models, yet still achieves competitive accuracy. The model can process images alone, clinical data alone, both together, or a combination of an image and partial anamnesis, making it adaptable to scenarios where complete clinical information is not available.

Our method has limitations, notably the reliance on pre-trained ViT and LM models. While we advocate for the lightweight approach, we acknowledge the need for prior computational resources to create the checkpoints used in LiwTERM feature generalization. Our goal is to leverage existing resources to optimize lightweight performance.

Also, we also recognize a limitation in the inference process. Although LiwTERM reduces training requirements when compared to the baseline methods, it still depends on ViT and LM feature extraction for final skin lesion predictions. This means that while our model offers clear training advantages, it does not lessen the computational load during testing. However, as previously noted, training costs are significantly higher than those of a single inference, which still justifies using LiwTERM.

VI. Conclusions and Future Works

In this work, we proposed LiwTERM, a lightweight training model that combines two pre-trained deep backbones (ViT and LM architectures) with a shallow, trainable neural block. Our approach addresses the challenge of limited resources for training deep architectures, requiring less computational power by focusing only on the final embedding calculations. This allows LiwTERM to deliver competitive results compared to other methods in the literature. LiwTERM also offers advantages over other multimodal approaches, providing a transformer-based solution that is suitable for low-resource scenarios and can be trained on a CPU while achieving results

comparable to state-of-the-art methods for skin lesion classification. Our model introduces a new approach to combining image and text features, using ViT for images and LLMs for text, with classification handled by a shallow lightweight neural network. Additionally, LiwTERM improves feature availability by functioning even when some data, such as incomplete anamnesis, is missing, making it adaptable to real-world clinical scenarios. In future work, we plan to incorporate other neural architectures as baselines and explore different methods for integrating image and text features.

ACKNOWLEDGMENTS

The authors thank the Espírito Santo Research Foundation (FAPES), 2022-NGKM5, 2021-GL60J; São Paulo Research Foundation (FAPESP); the Alexander von Humboldt Foundation; the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES), Finance Code 001; the National Council for Scientific and Technological Development (CNPq); the Brazilian Ministry of Health (MoH); and Brazilian National Program of Genomics and Precision Health (Genomas Brasil).

REFERENCES

[1] OMS, "Radiation: Ultraviolet (UV) radiation and skin cancer," World Health Organization (WHO), 2017, available at: http://www.who.int/uv/faq/skincancer/en/index1.html. Last access: 06/05/2023.

[2] INCA, "Incidência do câncer no Brasil," Instituto Nacional do Câncer (INCA), 2022, available at: https://www.inca.gov.br/sites/ufu.sti.inca.local/files//media/document//estimativa-2023.pdf. Last access: 06/05/2023.

[3] H. Kittler, H. Pehamberger, K. Wolff, and M. Binder, "Diagnostic accuracy of dermoscopy," *The Lancet Oncology*, vol. 3, no. 3, pp. 159–165, 2002. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1470204502006794

[4] C. Sinz, P. Tschandl, C. Rosendahl, B. N. Akay, G. Argenziano, A. Blum, R. P. Braun, H. Cabo, J.-Y. Gourhant, J. Kreusch, A. Lallas, J. Lapins, A. A. Marghoob, S. W. Menzies, J. Paoli, H. S. Rabinovitz, C. Rinner, A. Scope, H. P. Soyer, L. Thomas, I. Zalaudek, and H. Kittler, "Accuracy of dermatoscopy for the diagnosis of nonpigmented cancers of the skin," *Journal of the American Academy of Dermatology*, vol. 5444, no. 6, pp. A1–A50, 2017. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0190-9622(17)32151-5

[5] H. Feng, J. Berk-Krauss, P. W. Feng, and J. A. Stein, "Comparison of dermatologist density between urban and rural counties in the united states," *JAMA Dermatology*, vol. 154, pp. 1265—-1271, 2018. [Online]. Available: https://api.semanticscholar.org/CorpusID:52174956

[6] R. M. Scheffler, J. X. Liu, Y. Kinfu, and M. R. D. Poz, "Forecasting the global shortage of physicians: an economic- and needs-based approach." *Bulletin of the World Health Organization*, vol. 867, pp. 516–523B, 2008. [Online]. Available: https://api.semanticscholar.org/CorpusID:11908516

[7] A. Green, N. Martin, J. Pfitzner, M. O'Rourke, and N. Knight, "Computer image analysis in the diagnosis of melanoma," *Journal of the American Academy of Dermatology*, vol. 31, no. 6, pp. 958–964, 1994. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0190962294702640

[8] G. Argenziano, G. Fabbrocini, P. Carli, V. De Giorgi, E. Sammarco, and M. Delfino, "Epiluminescence Microscopy for the Diagnosis of Doubtful Melanocytic Skin Lesions: Comparison of the ABCD Rule of Dermatoscopy and a New 7-Point Checklist Based on Pattern Analysis," *Archives of Dermatology*, vol. 134, no. 12, pp. 1563–1570, 12 1998. [Online]. Available: https://doi.org/10.1001/archderm.134.12.1563

[9] A. Masood and A. Al-Jumaily, "Computer aided diagnostic support system for skin cancer: A review of techniques and algorithms," *International journal of biomedical imaging*, vol. 2013, p. 323268, 01 2013.

[10] A. Esteva, B. Kuprel, R. A. Novoa, J. Ko, S. M. Swetter, H. M. Blau, and S. Thrun, "Dermatologist-level classification of skin cancer with deep neural networks," *Nature*, vol. 542, pp. 115–, 2017.

[11] A. G. C. Pacheco, A. Ali, and T. Trappenberg, "Skin cancer detection based on deep learning and entropy to detect outlier samples," *CoRR*, vol. abs/1909.04525, 2019. [Online]. Available: http://arxiv.org/abs/1909.04525

[12] A. G. C. Pacheco and R. A. Krohling, "Recent advances in deep learning applied to skin cancer detection," 2019.

[13] A. G. Pacheco and R. A. Krohling, "The impact of patient clinical information on automated skin cancer detection," *Computers in Biology and Medicine*, vol. 116, p. 103545, 2020. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0010482519304019

[14] A. G. C. Pacheco, T. Trappenberg, and R. A. Krohling, "Learning dynamic weights for an ensemble of deep models applied to medical imaging classification," in *2020 International Joint Conference on Neural Networks (IJCNN)*, 2020, pp. 1–8.

[15] A. G. C. Pacheco and R. A. Krohling, "An attention-based mechanism to combine images and metadata in deep learning models applied to skin cancer classification," *IEEE Journal of Biomedical and Health Informatics*, vol. 25, no. 9, pp. 3554–3563, 2021.

[16] W. Hou, L. Wang, S. Cai, Z. Lin, R. Yu, and J. Qin, "Early neoplasia identification in barrett's esophagus via attentive hierarchical aggregation and self-distillation," *Medical Image Analysis*, vol. 72, p. 102092, 2021.

[17] T. Tuncer, P. D. Barua, I. Tuncer, S. Dogan, and U. R. Acharya, "A lightweight deep convolutional neural network model for skin cancer image classification," *Applied Soft Computing*, p. 111794, 2024. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1568494624005684

[18] Y. Li, H. Mao, and Z. Wang, "A lightweight skin cancer detection model based on convolutional neural network," in *CAIBDA 2022; 2nd International Conference on Artificial Intelligence, Big Data and Algorithms*, 2022, pp. 1–7.

[19] P. Kharazmi, S. Kalia, H. Lui, Z. J. Wang, and T. K. Lee, "A feature fusion system for basal cell carcinoma detection through data-driven feature learning and patient profile," *Skin Research and Technology*, vol. 24, no. 2, pp. 256–264, 2018. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1111/srt.12422

[20] S. Sierra and F. A. González, "Combining textual and visual representations for multimodal author profiling: Notebook for PAN at CLEF 2018," in *Working Notes of CLEF 2018 - Conference and Labs of the Evaluation Forum, Avignon, France, September 10-14, 2018*, ser. CEUR Workshop Proceedings, L. Cappellato, N. Ferro, J. Nie, and L. Soulier, Eds., vol. 2125. CEUR-WS.org, 2018. [Online]. Available: https://ceur-ws.org/Vol-2125/paper_219.pdf

[21] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16x16 words: Transformers for image recognition at scale," *CoRR*, vol. abs/2010.11929, 2020. [Online]. Available: https://arxiv.org/abs/2010.11929

[22] J. J. Webster and C. Kit, "Tokenization as the initial phase in nlp," in *Proceedings of the 14th Conference on Computational Linguistics - Volume 4*, ser. COLING '92. USA: Association for Computational Linguistics, 1992, p. 1106–1110. [Online]. Available: https://doi.org/10.3115/992424.992434

[23] C. W. Schmidt, V. Reddy, H. Zhang, A. Alameddine, O. Uzan, Y. Pinter, and C. Tanner, "Tokenization is more than compression," 2024.

[24] J. Devlin, M. Chang, K. Lee, and K. Toutanova, "BERT: pre-training of deep bidirectional transformers for language understanding," *CoRR*, vol. abs/1810.04805, 2018. [Online]. Available: http://arxiv.org/abs/1810.04805

[25] A. G. Pacheco, G. R. Lima, A. S. Salomão, B. Krohling, I. P. Biral, G. G. de Angelo, F. C. Alves Jr, J. G. Esgario, A. C. Simora, P. B. Castro, F. B. Rodrigues, P. H. Frasson, R. A. Krohling, H. Knidel, M. C. Santos, R. B. do Espírito Santo, T. L. Macedo, T. R. Canuto, and L. F. de Barros, "Pad-ufes-20: A skin lesion dataset composed of patient data and clinical images collected from smartphones," *Data in Brief*, vol. 32, p. 106221, 2020. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S235234092031115X

[26] ISIC, "Skin lesion analysis towards melanoma detection," International Skin Imaging Collaboration, 2019, last accessed: 10 March 2020. [Online]. Available: https://www.isic-archive.com

[27] F. Wilcoxon, "Individual Comparisons by Ranking Methods," *Biometrics Bulletin*, vol. 1, no. 6, pp. 80–83, 1945.