Journal of the Brazilian Computer Society, 2026, 32:1, doi: 10.5753/jbcs.2026.5871

LiwTERM-r: A Revised Lightweight Transformer-based Model for Multimodal Skin Lesion Detection Robust to Incomplete Input

Luis Antonio de Souza Júnior [**Universidade Federal do Espírito Santo** | **Graduate Program of Informatics** | ***la.souza@inf.ufes.br***]
André Georghton Cardoso Pacheco [**Universidade Federal do Espírito Santo** | **Graduate Program of Informatics** | ***apacheco@inf.ufes.br***]
Thiago Oliveira dos Santos [**Universidade Federal do Espírito Santo** | **Graduate Program of Informatics** | ***todsantos@inf.ufes.br***]
Wyctor Fogos da Rocha [**Universidade Federal do Espírito Santo** | **Graduate Program of Informatics** | ***wyctor.rocha@edu.ufes.br***]
Pedro Henrique Bouzon [**Universidade Federal do Espírito Santo** | **Graduate Program of Informatics** | ***pedro.bouzon@edu.ufes.br***]
Christoph Palm [**Ostbayerische Technische Hochschule Regensburg** | **Regensburg Medical Image Computing** | ***christoph.palm@oth-regensburg.de***]
João Paulo Papa [**Universidade Estadual Paulista** | **Department of Computing** | ***joao.papa@unesp.br***]

Department of Informatics, Universidade Federal do Espírito Santo, Av. Fernando Ferrari, 514 - Goiabeiras, Vitória, ES, 29075-910, Brazil.

Received: 09 April 2025 • **Accepted:** 05 September 2025 • **Published:** 16 March 2026

Abstract. As the most common type of cancer in the world, skin cancer accounts for approximately 30% of all diagnosed tumor-based lesions. Early diagnosis can reduce mortality and prevent disfiguring in different skin regions. With the application of machine learning techniques in recent years, especially deep learning, promising results in this task could be achieved, presenting studies demonstrating that the combination of patients' clinical anamneses and images of the injured lesion is essential for improving the correct classification of skin lesions. Despite that, meaningful use of anamneses with multiple collected images of the same skin lesion is mandatory, requiring further investigation. Thus, this project aims to contribute to developing multimodal machine learning-based models to solve the skin lesion classification problem by employing a lightweight transformer model that is robust to missing clinical information input. As a main hypothesis, models can be fed by multiple images from different sources as input along with clinical anamneses from the patient's historical evaluations, leading to a more factual and trustworthy diagnosis. Our model deals with the not-trivial task of combining images and clinical information concerning the skin lesions in a lightweight transformer architecture that does not demand high computation resources or even all the information from the anamneses but still presents competitive classification results.

Keywords: Deep learning, Skin Lesion Detection, Transformers, Lightweight Architectures.

1 Introduction

As stated by the World Health Organization (WHO), skin cancer is the most common dysplasia in the world, accounting for approximately 30% of all types of cancer diagnosed worldwide (OMS [2017]). With an expectation of fast and progressive growth for the next years, the Brazilian National Cancer Institute (INCA [2022]) projects that for the period 2024-2025, 220,000 new cases of skin cancer will be diagnosed, making it the most common cancer in Brazil, with approximately 31.2% of all cancer types (INCA [2022]]). However, even with the highest incidence, skin cancer's mortality rate is low, approximately 1% if ensured prognosis (INCA [2022]). Hence, late diagnosis is the main reason for the rise in this mortality rate. Taking in to account the low lethality, the tumor can imply mutilations on the skin if the lesion does not receive appropriate early diagnosis and correct treatment.

Experts in dermatology perform visual examinations of the potential injury and consider the patient's medical history in the clinical diagnosis of skin cancer, relying on theoretical and expertise-based insights to make an accurate evaluation. Clearly, such a process is challenging and requires specialized training and experience in dermoscopy. Kittler *et al*. [2002] and Sinz *et al*. [2017] have shown that the dermoscopy tool greatly improves the correct diagnostic, although its effectiveness is heavily dependent on the specialist's level of expertise. Moreover, the burden related to the high workloads the professionals have to deal with and human factors such as fatigue, stress, and emotional issues can harm their diagnostic capability, especially when tracking early-stage injuries. The situation becomes even worse in peripheral and rural regions that present a limited number of available experts and specialized equipment. Thus, considering the high incidence of skin cancer and the bottleneck that lack of required resources describes, especially in rural areas (Feng *et al*. [2018]) and emerging countries (Scheffler *et al*. [2008]), the development of Computer Aided Diagnosis (CAD) systems to aim at skin cancer detection becomes a highly desirable trend, with poten-

tial to increase effectiveness and enhance precision in clinical diagnosis to be incorporated in healthcare systems.

The use of CAD systems to assist in skin lesion analysis has been intensely investigated in recent years (Green *et al*. [1994]; Argenziano *et al*. [1998]; Masood and Al-Jumaily [2013]), with the most successfully and significantly used strategy, in terms of performance, being the machine learning, and more precisely, deep learning (Esteva *et al*. [2017]; Pacheco *et al*. [2019]; Pacheco and Krohling [2019]). Even showing promising results, several challenges need to be addressed to enable the implementation of such technologies more safely and reliably. Among the challenges, the uncertainty and problems in the data, biases, datasets with a small or not representative number of samples, low models' generalization, and low capability of prediction explanation and understanding can be highlighted (Pacheco and Krohling [2019]). To handle some of these problems, images and clinical data from anamnesis have been used to classify skin cancer (Pacheco and Krohling [2020]; Pacheco *et al*. [2020b]; Pacheco and Krohling [2021]; Souza Jr. *et al*. [2024]).

The aforementioned works deal with identifying skin lesions by employing deep architectures, which are generally costly. A few works proposed lightweight architectures in the context of identifying neoplastic tissues. Hou *et al*. [2021] proposed an approach for early identification of cancer in Barrett's esophagus-diagnosed samples using attentive hierarchical aggregation and self-distillation. Their work used a SE-ResNet50 as the model's backbone, a variation of the well-known ResNet50 with squeeze-and-excitation modules, reporting promising results and a complex but efficient model.

In light of the skin cancer detection context, Tuncer *et al*. [2024] introduced a lightweight Convolution Neural Networks (CNN) architecture to classify dermatoscopical skin images between benign and malignant, aligned to work proposed and conducted by Li *et al*. [2022], where a lightweight CNN-based model to deal with the classification of 8 different skin lesions based on dermatoscopic databases have been developed. As one can observe, most current investigations focused on using images in their approaches, not making use of combining any other type of description, such as clinical information. Still, their architectures are mostly focusing on CNN-based variations.

Within the ML field, combining features extracted from images with other features obtained from different sources describes a common problem, i.e., the image as the main source of information and the extra data – hereby defined as metadata – as supplementary information about the issue. However, the question still remains: How can we combine such different information from the same problem? Kharazmi *et al*. [2018] proposed a feature fusion system based on concatenation and Sparse Autoencoder (SAE) to detect Basal Cell Carcinoma in skin tissue samples, while Sierra and González [2018] used concatenation transformations to fuse image features, extracted using two CNN architectures, with text-based information in the prediction of gender. More recently, Pacheco and Krohling [2021] conducted a similar work to predict six different skin lesions by imposing transformations to image features within the CNN architecture pipeline using a one-hot-encoding representation of the metadata. Also, Souza Jr. *et al*. [2024] introduced a Transformer-based model that combines both images and metadata for skin lesion classification. The model has demonstrated competitive performance by combining representations of images and metadata using transformer-based pre-trained architectures and a shallow neural network to be trained and learn the proper classification weights.

Hence, this work proposes an extension in the evaluation of LiwTERM (Souza Jr. *et al*. [2024]), a lightweight neural architecture hereby called LiwTERM-r that combines features learned by (i) a Vision Transformer (ViT) (Dosovitskiy *et al*. [2020]) 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, low computational demand, and high capability to deal wit missing information within the input data. The main contributions of this work may be summarized as follows:

- We propose a new method to handle the combination of image-based and text-based features in a multimodal skin cancer classification totally designed as a Transformer-derived model. We named this new approach LiwTERM-r (Revised Light Weight Transformer-model for Dermatological purposes), which extends the feature generalization of Transformers without requiring their finetuning.
- We evaluate the proposed LiwTERM-r method on three different datasets: the PAD-UFES-20 dataset (Pacheco *et al*. [2020a]) and a new extended version of this dataset, hereby named PAD-UFES-20+, that is around $\tilde{6}.5$ times larger, and the ISIC19 dataset. The PAD-UFES-20 and ISIC19 datasets are used as performance benchmarks for controlled scenarios since these datasets present predictable behavior concerning metadata availability. The PAD-UFES-20+ was employed to evaluate the LiwTERM-r's performance in the real scenario, where metadata is not completely available.
- 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 proposed LiwTERM method is simple to implement, lighter than transformer-based architectures, and demonstrates competitive performance even when metadata is absent within the dataset.

The remainder of this paper is organized as follows. Sections 2 and 3 introduce the proposed approach and the methodology, respectively. Section 4 presents the experiments, and Section 5 discusses the outcomes. Last but not least, Section 6 states conclusions and future works.

2 Proposed Approach - LiwTERM-r

This work proposes a model that combines features from pretrained ViTs and text-based tokenizers without incurring a transformer-based computational training cost. The generalization of Transformers for image-based tasks such as identification, description, and classification, namely ViT, are robust models for image-driven problem-solving. Even with remarkable image description potential, ViTs' drawbacks regarding

the fine-tuning process for specific tasks mostly relate to the high training computational requirements. Conversely, the inference task in pre-trained ViT models does not demand the same high computational cost but still delivers a powerful image representation.

With the massive advances in the use of Generative Pre-trained Transformers (GPT) in recent years, the text-representation field has shown accountable progress with models that encode-decode text and describe intrinsic and very difficult tasks such as sentence context and word positioning. The GPT pipeline is primarily composed of the tokenization process, which copes with the decomposition of a sentence to be further consumed as tokens. Tokens are the basic text's description units in the Natural Language Processing (NLP) field, precisely describing word and sentence positioning and delimiters (Webster and Kit [1992]), aligning raw text and context for language models (LMs) (Schmidt *et al*. [2024]). In light of the current LM, the tokenization process is a crucial part of their generalization design since such models use this technique and are built with transformer encoder-decoder designs, with bidirectional encoders (BERT-like - Devlin *et al*. [2018]) and autoregressive (GPT-like) decoders, in a sequence-to-sequence input-output. At the halfway of representation, text-based embeddings can be obtained to represent the original input sentence as high-encoded information.

In general, fine-tuning transformer-based models, including those dealing with text or images (ViT specifically), requires a huge amount of samples to perform reasonably. Thus, already pre-trained checkpoints can provide powerful feature inference for a wide range of contexts, leveraging the representation of thousands of samples previously used in the transformer tuning and benefiting the description for other simpler model training. This enables simpler setups to perform training and inference processes using transformer-based generalization over a reduced amount of available samples for such a task, a common fact in the medical field.

Also, using transformer-based models to describe both image and text may lead to the representation of two different representations to a third-and-similar (if not equal) domain. The ViT is in charge of image representation in a serialization process that considers patch positioning, tokenization, and linear projections, which can be easily comparable to the tokenization process performed by several NLP approaches. Due to this similarity in representation, combining image and text representations in a unique environment can be designed in a complementary fashion to each representation.

In this sence, we propose LiwTERM-r, a hybrid-and-shallow transformer-based model where image and text inputs feed a fine-tuned encoder to classify six skin lesions and takes advantage of both representations. The ViT and LM embeddings are combined by full connection layers from the second last layer of each architecture, feeding an encoder composed of four fully connected layers composed of ReLU, batch normalization, and dropout transformations, hereby named shallow light-weight model (SLM). A SoftMax head defines the classes at the end of SLM model for the final classification. The fine-tuning is performed only from the fully connected layer of the ViT and tokenizer that extract the image and text features, configuring a shallow-based training phase. Figure 1 illustrates the LiwTERM-r pipeline [1].

Table 1. The number of samples for PAD-UFES-20 and its extended version. Both datasets contain the same set of labels.

Label	PAD-UFES-20	PAD-UFES-20+
ACK	730	8400
BCC	845	3042
MEL	52	346
NEV	244	353
SCC	192	2282
SEK	235	689
Total	**2298**	**15112**

3 Methodology

3.1 Datasets

The proposed method has been evaluated in two public datasets named (i) PAD-UFES-20 (Pacheco *et al*. [2020a]) and (ii) ISIC 2019 (ISIC [2019]). The PAD-UFES-20 dataset is composed of a clinical skin lesion collection of images and metadata acquired in the Espírito Santo State in Brazil. In total, 2.298 clinical images were collected from smartphone devices, 21 patient clinical features, such as age, gender, anatomical region, cancer history, and skin prototype, among others, and six skin lesions, Basal Cell Carcinoma (BCC), Squamous Cell Carcinoma (SCC), Actinic Keratosis (ACK), Seborrheic Keratosis (SEK), Melanoma (MEL), and Nevus (NEV). Another dataset, ISIC 2019, comprises 25,331 public dermoscopy images, three clinical features collected from anamneses, i.e., 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 samples from PAD-UFES-20 and ISIC 2019 datasets, respectively.

To evaluate LiwTERM-r's performance in a large real-world application, we evaluate the model on an extended version of the PAD-UFES-20 dataset, which contains around 6.5 times more samples and includes real-world missing metadata (in a rate of around 14% concerning features from the clinical information). This evaluation aims to show the proposed method's performance under uncontrolled conditions and its robustness in handling real-world missing data situations. The extended version contains 15,112 samples from 5,589 patients with similar patient clinical information. Both datasets have the same skin lesion types (labels), and Table 1 presents the number of samples for each of them.

3.2 Evaluation Measures

We calculated three well-known quantitative measures to assess LiwTERM-r approaches: Sensitivity (S), Specificity (P), and Balanced Accuracy (BACC). The experimental setup also describes a statistical evaluation using Wilcoxon's signed-rank test (Wilcoxon [1945]) with a significance level of 5%.

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

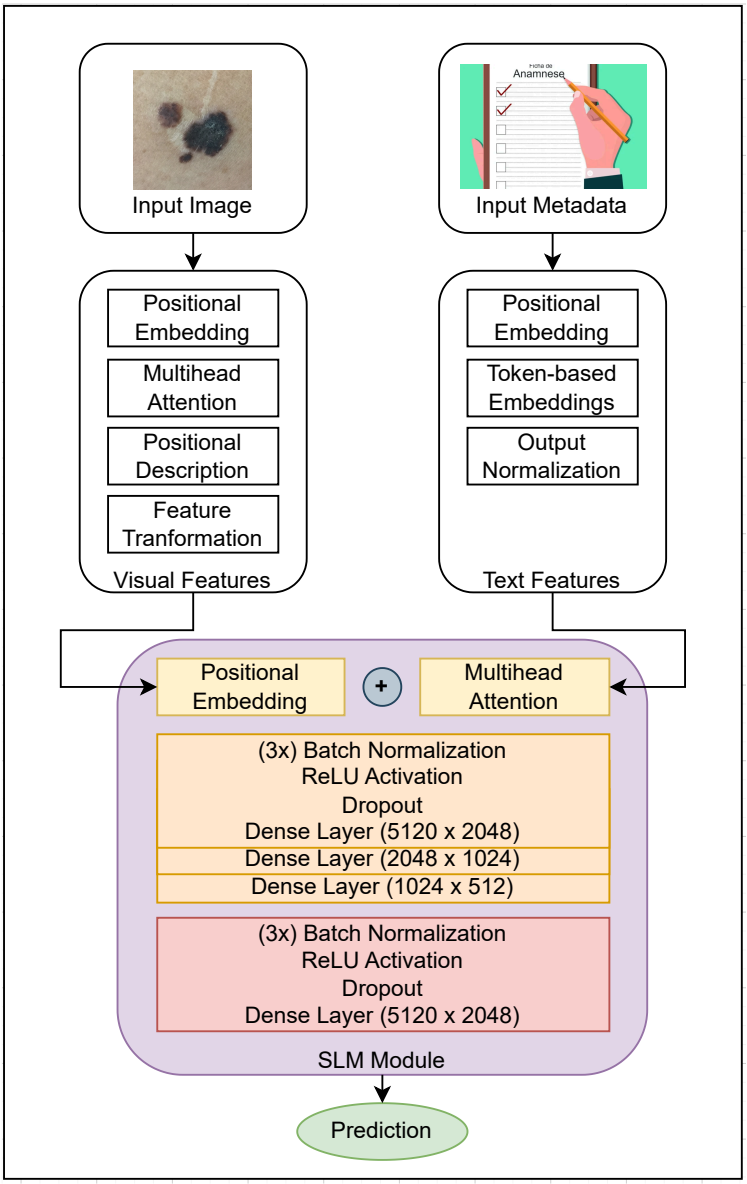
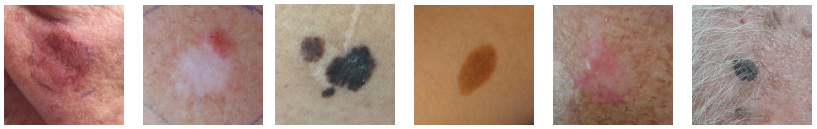
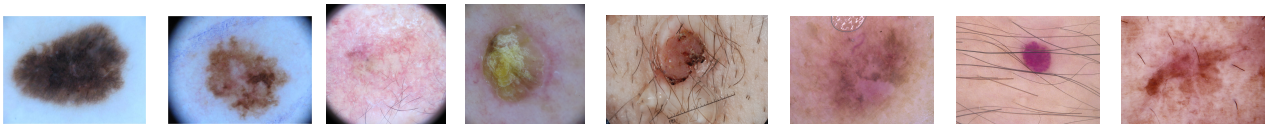


Figure 1. LiwTERM-r pipeline: the proposed model has two sections: (i) the feature extraction (with no color) and (ii) the shallow lightweight model section (with colors). The colored section concerns the trainable part of the proposed method; fed from the deep and complex ViT and tokenizer architectures, this section is in charge of learning the proper weights to provide the classification of skin lesions based on images and clinical information (from Souza Jr. *et al*. [2024]).



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

Figure 2. PAD-UFES-20 (and PAD-UFES-20+) dataset samples (from Pacheco *et al*. [2020a]).



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

Figure 3. ISIC 2019 dataset samples (from ISIC [2019]).

3.3 Experimental Delineation

LiwTERM-r has been evaluated over three experimental approaches: (i) a 5-fold cross-validation only employing the ViT bottleneck for the feature inference (only image-based features), (ii) a 5-fold cross-validation only employing the text-tokenization bottleneck for the feature inference (only text-based features), and (iii) a 5-fold cross-validation using the entire method, with feature generalized from both images and anamnese. Additionally, the baseline results (avoiding the shallow lightweight training portion) based on (iv) the classification of cancerous skin tissue only using the pre-trained ViT architecture and (v) the skin lesion classification based only on LM with the correspondent checkpoint 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 approaches were evaluated in a class-stratification fashion, balancing the number of samples of each skin lesion class for each fold.

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

3.4 Implementation Details

We employed a computer with 16 GB RAM and an NVIDIA RTX®3070 Graphics card of 8 GB VRAM for the experiments. We used the Pytorch framework for the code implementation, and it is imperative to highlight that the proposed model is designed to cope with the problem of high computational costs imposed by transformers. As one can observe, a simple computer setup has been employed for the experiments, showing the ordinary computer configuration demand our proposed model requires. We evaluated LiwTERM-r over GPU and CPU sets, with the same performance outcomes (despite the longer training time).

4 Experimental Results

LiwTERM-r focuses on three main aspects: the correct classification of skin lesions, the reduced computational cost required for fine-tuning only a portion of the model (already presented in the last section), and the training process' time consumption. Table 2 presents the model classification results on the PAD-UFES-20, ISIC 2019, and PAD-UFES-20+ datasets for all the evaluated approaches, highlighting the impact of each selected image-backbone for the feature-extraction composition. For comparison purposes, Table 2 also presents the baseline results of the evaluated backbones proposed for LiwTERM-r model. Figures 4a and 4c illustrate the overall confusion matrices of complete LiwTERM-r model over PAD-UFES-20, PAD-UFES-20+, and ISIC 2019 datasets, respectively.

5 Discussion

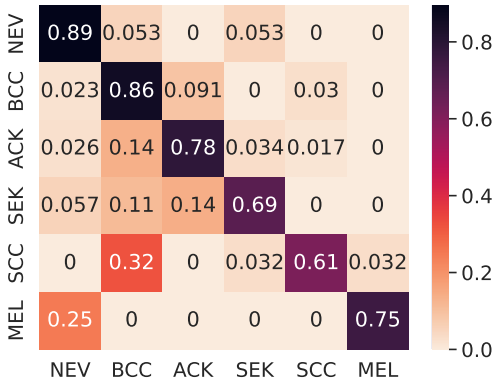
5.1 LiwTERM's Backbone Analysis

As one can observe, Table 2 highlights the efficiency of our method, where the feature-encoded information from ViT and LM models showed a complementary behavior leading to enhancements in the correct prediction of skin lesions (Fig. 4). Using only the pre-trained ViT or LM models for predicting the skin lesions was insufficient (baselines and approaches I and ii), so fine-tuning must be conducted to make such a classification feasible and deliver competitive accuracy outcomes. The introduction of the trainable shallow portion to the model could enhance the prediction results, presenting at the end the best ones when both ViT and LM models work together for the feature generalization process. Additionally, no statistical similarity was found between the best and the remaining results in the Wilcoxon text. This suggests that employing ViT and LM features outperforms all the other experimental designs, including the baselines.

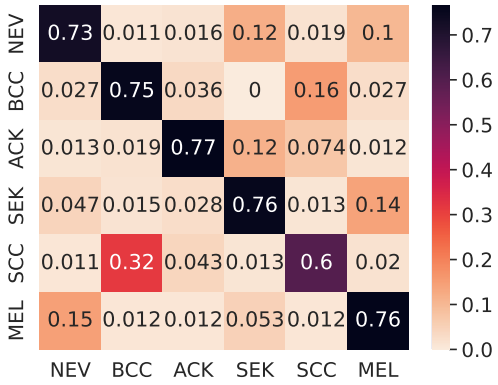
We conducted a benchmark evaluation of LiwTERM-r approaches to understand the model's memory usage. The baseline models (only ViT and LLM backbones without the SLM portion) have approximately 60M and 0.79M parameters, consuming around 200MB and 4MB of memory, respectively. When adding the SLM portion to the final LiwTERM-r architecture, we changed such scenarios to about 63M parameters and 260MB for ViT, and 1.45M parameters and 15MB for LLM. As one can observe, the addition of SLM module slightly impacted the original feature extraction size. The final LiwTERM-r 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 2 with a marginal increase in the model's size compared to the backbones.

We also conducted an evaluation of LiwTERM-r's 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-r 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-r. Finally, for PAD-UFES-20+, results of training times were approximately 5.12h for ViT+SLM, 4.84h for baseline ViT, 1.73 for LLM+SLM, 1.66h for baseline LLM, and 5.19h for LiwTERM-r. The inference times of single samples for all approaches are quite small, ranging between $10^{-4}sec$ and $10^{-6}sec$. The SLM module, responsible for the classification generalization of LiwTERM-r, adds minimal overhead to ViT and LLM backbones concerning parameters, memory, or training/inference times, making it reasonable to use even with CPU configurations. However, GPUs are preferable for reducing training time.

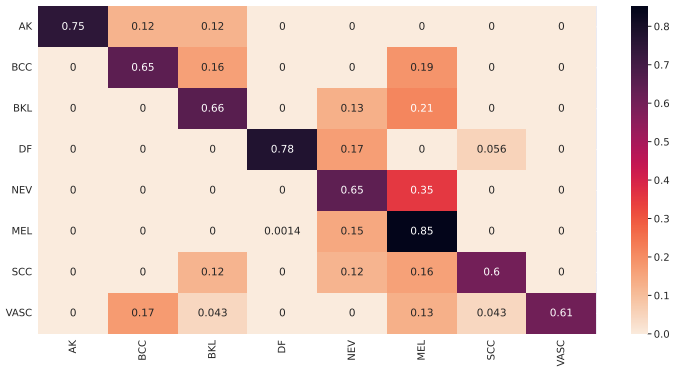
Table 3 presents a comparison between our method and the ones from state-of-the-art (Pacheco and Krohling [2021]), which employed the same evaluation protocol and datasets. As one can observe, our LiwTERM-r presents competitive performance compared to the CNN-based approaches. The Concatenation, Metablock, and MetaNet methods generalize



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



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



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

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

Table 2. LiwTERM classification results using 5-fold validation protocol over the PAD-UFES-20, ISIC 2019, and PAD-UFES-20+ datasets. Bold lines mean the overall best-obtained outcome for each dataset.

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
	PAD-UFES-20+	ViT+SLM	0.52 ± 0.05	0.64 ± 0.05	0.60 ± 0.05
		baseline	0.40 ± 0.07	0.52 ± 0.08	0.49 ± 0.07
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
	PAD-UFES-20+	LLM+SLM	0.57 ± 0.06	0.64 ± 0.07	0.60 ± 0.06
		baseline	0.47 ± 0.09	0.54 ± 0.06	0.50 ± 0.08
ViT + LM	**PAD-UFES-20**	**ViT+LLM+SLM**	$\mathbf{0.69 \pm 0.02}$	$\mathbf{0.76 \pm 0.02}$	$\mathbf{0.74 \pm 0.01}$
	ISIC 2019	**ViT+LLM+SLM**	$\mathbf{0.66 \pm 0.02}$	$\mathbf{0.77 \pm 0.03}$	$\mathbf{0.73 \pm 0.03}$
	PAD-UFES-20+	**ViT+LLM+SLM**	$\mathbf{0.65 \pm 0.04}$	$\mathbf{0.78 \pm 0.05}$	$\mathbf{0.71 \pm 0.04}$

the textual information using one-hot encoding (Pacheco and Krohling [2021]). Although it is a simple and efficient approach, it fails, for example, to handle common issues from medical anamnesis, such as out-of-vocabulary words or missing words. On the other hand, our model uses a transformer-based architecture to process the text-based information, being much more robust to these issues. We also performed a statistical analysis to compare the methods, showing that our method is statistically similar to the CNN-based approaches regarding balanced accuracy (displayed in Table 3). It is important to highlight that such similar results are achieved using the same amount of data and employing a transformer-based architecture, as it is known to be data-hungrier than CNN-based approaches.

5.2 LiwTERM Strengths and Limitations

With LiwTERM-r, we propose a transformer-based training model that combines two pre-trained deep backbones (ViT and LM architectures) with a shallow-and-lightweight trainable neural block, addressing the challenge of limited resources for training deep architectures (especially transformer-based ones). Unlike traditional transformer-based models, LiwTERM-r requires less computational resources since only the final embedding calculations from the backbones are trained along with the SLM module. Additionally, by focusing on the feature description layers and the SLM block, our model achieves competitive results even with reduced training data, a severe bottleneck for transformer-based solutions.

The use of two backbones, i.e., ViT and LM models, significantly enhances the dimensionality of skin lesion descriptions by combining different features from the same context (images and text), as data availability describes a key challenge in medical applications, especially when dealing with sensitive data that requires legal permissions. LiwTERM-r leverages transformer-based representations without needing to fine-tune such models but still delivers competitive accuracy compared to state-of-the-art studies. The model can standalone process images, clinical information, both together, or a combination of images and parts of the anamnesis, making it adaptable to practical scenarios where complete clinical data cannot be available.

Notably, the reliance on pre-trained ViT and LM models is one of the main limitations of our model. While we advocate for lightweight training behavior in comparison to fine-tuning transformers, we acknowledge the need for prior computational resources to create the checkpoints used in LiwTERM-r's backbones for the image and text feature calculation. Thus, LiwTERM-r's main goal is to leverage existing resources to optimize transformer-based model performance, which can be used in low-capable environments.

We also recognize the inference process limitation LiwTERM-r imposes. Although the model reduces training requirements compared to other transformer-based methods, it still depends on ViT and LM feature extraction to predict skin lesions. So, while our model clearly offers training advantages, it does not lessen the computational load during inference time. However, training costs are significantly higher than the cost of a single inference, still justifying LiwTERM-r's applicability.

5.3 Ablation Study - Impact of the metadata on the correct skin lesion detection

To understand the impact of the features belonging to the metadata on the correct classification of each skin lesion class, an ablation study has been conducted, imposing and removing metadata parts on the PAD-UFES-20 dataset. From what was previously presented in the literature, there are some notable features, such as age, gender, and sun exposure, that highly impact the presence of some skin lesions, such as carcinomas and melanoma (Pacheco *et al*. [2020a]; Pacheco and Krohling [2021]; Pacheco *et al*. [2020b]; Pacheco and Krohling [2020]). Interestingly, these impacts had never been deeply investigated in terms of machine learning studies, since the multi-modal-based models are starting to be recognized in the current years.

Table 3. Comparison of LiwTERM-r 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-r**	**Lightweight Transformer-based**	**0.74 ± 0.01**
	Pacheco and Krohling [2021]	CNNs	0.65 ± 0.02
	Pacheco and Krohling [2021]	**CNNs + Concatenation**	**0.76 ± 0.01**
	Pacheco and Krohling [2021]	**CNNs + MetaBlock**	**0.77 ± 0.02**
	Pacheco and Krohling [2021]	CNNs + MetaNet	0.75 ± 0.03
ISIC 2019	LIWTERM-r	Lightweight Transformer-based	0.73 ± 0.03
	Pacheco and Krohling [2021]	CNNs	0.75 ± 0.04
	Pacheco and Krohling [2021]	CNNs + Concatenation	0.77 ± 0.02
	Pacheco and Krohling [2021]	CNNs + MetaBlock	0.77 ± 0.01
	Pacheco and Krohling [2021]	CNNs + MetaNet	0.76 ± 0.01

LiwTERM model's nature is robust enough to deal with part of the metadata, and what has been investigated is the presence of each metadata feature in the correct classification of the six skin lesions. To cope with such a task, we trained and further evaluated every single model in a design stratified by class and metadata presence. For instance, to evaluate the age feature, the training data was composed of samples based on images and only age information. For evaluating only sun exposure, only such metadata information would appear as text-based feature in the training and test sets, and so on for each metadata information. This first evaluation stands for observing the impact of each of the selected metadata features in a standalone fashion. The selection of the most impactful features has been conducted based on clinical research that corroborates their importance as key markers in skin lesions, especially skin cancer appearance. The results are presented in Table 4.

Obviously, the metadata features are somehow correlated, and from the clinical background already presented in the literature, we are aware of such characteristics. To deal with that, we also conducted ablation studies to infer the best set of metadata features in the correct classification of skin lesions. This ablation study is highly desired since the use of metadata in practical observations made by the specialists assists the clinical diagnosis. Hence, we want to understand if such behavior benefits or not the correct skin lesions classification using transformer-based models. Also, it is important for us to observe the correlated metadata features in the computational validation of skin lesions. The experimental delineation was the same as the one proposed for evaluating the features in the standalone design, and the results can also be observed in Table 4.

From the computational perspective and checking Table 4, it is imperative to highlight that, in fact, some features present a higher impact on the correct classification of skin lesions than others. As one can observe, in the standalone evaluation, age, gender, and grew features present the highest BACC measures, clearly suggesting their higher impact on overall correct prediction of skin lesion classes. For the best set of features in the description of skin lesions along with image-based description, we understand that age, gender, grew, changed and bleed features are highly correlated when describing the problem with LLM models, leading to higher BACC results (even outperforming the model which uses all the metadata features). For us, this is clear evidence that our model is suitable for the use of different clinical features, but it also relies on the correlation they impose to the correct classification of skin lesions, especially the most dangerous ones such as melanoma.

Still, we acknowledge that the clinical information storing is very human-dependent, but since our model deals with tons of metadata, and we deeply investigated the impact of such features in the correct classification of ill skin, we understand the LiwTERM is robust to misrepresented clinical information, reaching competitive results even when all the features are presented as input in the textual composition that feeds the LLM backbone of the model.

5.4 Understanding the impact of LiwTERM on real scenarios - The evaluation of the extended PAD-UFES-20 dataset

The key difference between the PAD-UFES-20 and PAD-UFES-25 datasets, besides the significantly higher number of samples, is the presence of real-world missing metadata. In such a scenario, multiple samples present incomplete clinical features. For example, around 14% of the sample do not have information about the anatomic region of the lesion, which is an important feature for skin lesion classification (Pacheco and Krohling [2020]). Such cases are common in clinical practice, where patient data may be incomplete or unavailable, representing a significant challenge for CAD systems.

When evaluating the real-world missing data scenario depicted by PAD-UFES-20+ dataset, presented in Table 2, we observed that LiwTERM is robust to missing parts of metadata information, outperforming approaches where the metadata is completely absent. Interestingly, this result matches the ones obtained using only part of the metadata, presented in Table 4. The PAD-UFES-20+ dataset is highly influenced by missing metadata information at a rate of $\approx 14\%$. As previously presented in the Experimental Results and Discussion sections, the combination of image-based and text-based features is mandatory to enhance the correct classification of skin lesions, especially the most representative metadata information (Table 4). Once such features are available, it is expected to be a degradation to LiwTERM's performance, as we deeply investigated in the metadata presence ablation study.

These observations imply that the extended version of the PAD-UFES-20 dataset presents a more demanding scenario than the original dataset. Notably, LiwTERM managed to

Table 4. Ablation study of LiwTERM on different sets of metadata composing the input description along with the image backbone part of the model. This evaluation has been conducted over PAD-UFES-20 dataset. The result with a $\star$ represents the best combination of features for optimizing BACC metric.

Dataset	Age	Gender	Itch	Grew	Hurt	Changed	Bleed	S	P	BACC
PAD-UFES-20	✓	x	x	x	x	x	x	0.63 ± 0.02	0.71 ± 0.03	0.70 ± 0.03
	x	✓	x	x	x	x	x	0.60 ± 0.03	0.69 ± 0.03	0.67 ± 0.03
	x	x	✓	x	x	x	x	0.62 ± 0.01	0.67 ± 0.06	0.64 ± 0.05
	x	x	x	✓	x	x	x	0.62 ± 0.05	0.68 ± 0.03	0.68 ± 0.04
	x	x	x	x	✓	x	x	0.61 ± 0.03	0.68 ± 0.07	0.66 ± 0.06
	x	x	x	x	x	✓	x	0.65 ± 0.05	0.72 ± 0.04	0.69 ± 0.04
	x	x	x	x	x	x	✓	0.62 ± 0.05	0.68 ± 0.08	$0.67 \pm \mathbf{0.06}$
	✓	✓	x	✓	x	✓	✓	$\star\mathbf{0.74 \pm 0.03}$	$\mathbf{0.80 \pm 0.05}$	$\mathbf{0.76 \pm 0.03}$

maintain similar performance across both scenarios, demonstrating that it is more robust than the other methods proposed in the literature where the entire set of metadata features is required to feed the models' inputs.

Although the results are promising, it is important to highlight that LiwTERM is still harmed by missing metadata. As shown in Table 4, LiwTERM's performance deteriorates as the amount of missing metadata increases, and this is exactly the behavior we expected and evaluated with PAD-UFES-20+. Even with such decreasing in comparison to PAD-UFES-20 results, and still considering that the extension of such dataset presents 6.5 times more data and a rate of missing metadata around 14%, LiwTERM could still achieve promising and competitive results, especially for skin lesions that present more instances with the extended database (Fig. 4b). Thus, while LiwTERM is not a perfect solution, it represents a significant step towards more robust multimodal models for skin lesion classification. Additionally, it does not rely on imputation methods or synthetic data, which could introduce unrealistic patterns that do not reflect true medical variability.

In addition to outperforming previous methods and still being able to handle missing clinical information and present competitive results, the proposed model is straightforward to implement, leveraging the powerful representation of transformer-based architectures without the burden of fine-tuning such backbones.

6 Conclusions and Future Works

In this work, we proposed LiwTERM-r, a transformer-based model that combines two pre-trained deep backbones (ViT and LM architectures) for image and text descriptions with a shallow, lightweight-trainable neural block. Our approach addresses the challenge of limited resources for training transformer-based architectures, requiring less computational resources by focusing only on the backbones' final embedding generalization. This allows LiwTERM-r to deliver competitive results compared to other methods in the literature. Also, LiwTERM-r offers advantages over other multimodal approaches, providing a transformer-based solution suitable for low-resource setups, i.e., can be trained on a CPU, that still achieves results comparable to state-of-the-art methods for skin lesion classification. Our model introduces a new approach to combine image and text features, using ViT for images and LLMs for text, with the classification learning process dealt by a shallow neural network. Additionally, LiwTERM-r improves feature availability problems, once it still works even when some data, such as incomplete anamnesis, is missing. This makes our model adaptable to real-world and practical clinical scenarios, where not all the data besides the image can be entirely collected. 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).

Author's Contribution

Luis A. Souza Jr. and André G. C. Pacheco led the study, contributing at every step from conception through study development, coding, evaluation, and manuscript writing. Wyctor F. Rocha and Pedro H. Bouzon contributed to the coding and evaluation steps. Thiago Santos-Oliveira, Christoph Palm, and João P. Papa collaborated in the study's conception and paper's writing and reviewing.

Conflicts of Interest

The authors declare no conflicts of interest.

Availability of Data and Materials

Part of the evaluated data is available at Pacheco *et al*. [2020a]. The updated dataset will be publicly available in the near future.

References

Argenziano, G., Fabbrocini, G., Carli, P., De Giorgi, V., Sammarco, E., and Delfino, M. (1998). 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*, 134(12):1563–1570. DOI: 10.1001/archderm.134.12.1563.

Devlin, J., Chang, M., Lee, K., and Toutanova, K. (2018). BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805. Available at:`http://arxiv.org/abs/1810.04805`.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., and Houlsby, N. (2020). An image is worth 16x16 words: Transformers for image recognition at scale. *CoRR*, abs/2010.11929. DOI: 10.48550/arxiv.2010.11929.

Esteva, A., Kuprel, B., Novoa, R. A., Ko, J., Swetter, S. M., Blau, H. M., and Thrun, S. (2017). Dermatologist-level classification of skin cancer with deep neural networks. *Nature*, 542:115–. DOI: 10.1038/nature21056.

Feng, H., Berk-Krauss, J., Feng, P. W., and Stein, J. A. (2018). Comparison of dermatologist density between urban and rural counties in the united states. *JAMA Dermatology*, 154:1265—1271. DOI: 10.1001/jamadermatol.2018.3022.

Green, A., Martin, N., Pfitzner, J., O'Rourke, M., and Knight, N. (1994). Computer image analysis in the diagnosis of melanoma. *Journal of the American Academy of Dermatology*, 31(6):958–964. DOI: 10.1016/S0190-9622(94)70264-0.

Hou, W., Wang, L., Cai, S., Lin, Z., Yu, R., and Qin, J. (2021). Early neoplasia identification in barrett's esophagus via attentive hierarchical aggregation and self-distillation. *Medical Image Analysis*, 72:102092. DOI: 10.1016/j.media.2021.102092.

INCA (2022). Incidência do câncer no Brasil. Available at: `https://www.inca.gov.br/sites/ufu.sti.inca.local/files//media/document//estimativa-2023.pdf`. Last access: 06/05/2023.

ISIC (2019). Skin lesion analysis towards melanoma detection. Available at:`https://www.isic-archive.com` Last accessed: 10 March 2020.

Kharazmi, P., Kalia, S., Lui, H., Wang, Z. J., and Lee, T. K. (2018). A feature fusion system for basal cell carcinoma detection through data-driven feature learning and patient profile. *Skin Research and Technology*, 24(2):256–264. DOI: 10.1111/srt.12422.

Kittler, H., Pehamberger, H., Wolff, K., and Binder, M. (2002). Diagnostic accuracy of dermoscopy. *The Lancet Oncology*, 3(3):159–165. DOI: 10.1016/S1470-2045(02)00679-4.

Li, Y., Mao, H., and Wang, Z. (2022). A lightweight skin cancer detection model based on convolutional neural network. In *CAIBDA 2022; 2nd International Conference on Artificial Intelligence, Big Data and Algorithms*, pages 1–7. Available at:`https://ieeexplore.ieee.org/document/10104328`.

Masood, A. and Al-Jumaily, A. (2013). Computer aided diagnostic support system for skin cancer: A review of techniques and algorithms. *International journal of biomedical imaging*, 2013:323268. DOI: 10.1155/2013/323268.

OMS (2017). Radiation: Ultraviolet (UV) radiation and skin cancer. Available at: `http://www.who.int/uv/faq/skincancer/en/index1.html`. Last access: 06/05/2023.

Pacheco, A. G. and Krohling, R. A. (2020). The impact of patient clinical information on automated skin cancer detection. *Computers in Biology and Medicine*, 116:103545. DOI: 10.1016/j.compbiomed.2019.103545.

Pacheco, A. G., Lima, G. R., Salomão, A. S., Krohling, B., Biral, I. P., de Angelo, G. G., Alves Jr, F. C., Esgario, J. G., Simora, A. C., Castro, P. B., Rodrigues, F. B., Frasson, P. H., Krohling, R. A., Knidel, H., Santos, M. C., do Espírito Santo, R. B., Macedo, T. L., Canuto, T. R., and de Barros, L. F. (2020a). Pad-ufes-20: A skin lesion dataset composed of patient data and clinical images collected from smartphones. *Data in Brief*, 32:106221. DOI: 10.1016/j.dib.2020.106221.

Pacheco, A. G. C., Ali, A., and Trappenberg, T. (2019). Skin cancer detection based on deep learning and entropy to detect outlier samples. *CoRR*, abs/1909.04525. Available at:`http://arxiv.org/abs/1909.04525`.

Pacheco, A. G. C. and Krohling, R. A. (2019). Recent advances in deep learning applied to skin cancer detection. DOI: 10.48550/arxiv.1912.03280.

Pacheco, A. G. C. and Krohling, R. A. (2021). 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*, 25(9):3554–3563. DOI: 10.1109/JBHI.2021.3062002.

Pacheco, A. G. C., Trappenberg, T., and Krohling, R. A. (2020b). Learning dynamic weights for an ensemble of deep models applied to medical imaging classification. In *2020 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8. DOI: 10.1109/IJCNN48605.2020.9206685.

Scheffler, R. M., Liu, J. X., Kinfu, Y., and Poz, M. R. D. (2008). Forecasting the global shortage of physicians: an economic- and needs-based approach. *Bulletin of the World Health Organization*, 867:516–523B. DOI: 10.2471/blt.07.046474.

Schmidt, C. W., Reddy, V., Zhang, H., Alameddine, A., Uzan, O., Pinter, Y., and Tanner, C. (2024). Tokenization is more than compression. DOI: 10.18653/v1/2024.emnlp-main.40.

Sierra, S. and González, F. A. (2018). 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*, volume 2125 of *CEUR Workshop Proceedings*. CEUR-WS.org. Available at:`https://ceur-ws.org/Vol-2125/paper_219.pdf`.

Sinz, C., Tschandl, P., Rosendahl, C., Akay, B. N., Argenziano, G., Blum, A., Braun, R. P., Cabo, H., Gourhant, J.-Y., Kreusch, J., Lallas, A., Lapins, J., Marghoob, A. A.,

Menzies, S. W., Paoli, J., Rabinovitz, H. S., Rinner, C., Scope, A., Soyer, H. P., Thomas, L., Zalaudek, I., and Kittler, H. (2017). Accuracy of dermatoscopy for the diagnosis of nonpigmented cancers of the skin. *Journal of the American Academy of Dermatology*, 5444(6):A1–A50. DOI: http://dx.doi.org/10.1016/j.jaad.2017.07.022.

Souza Jr., L. A., Pacheco, A. G. C., de Angelo, G. G., Oliveira-Santos, T., Palm, C., and Papa, J. P. (2024). Liwterm: A lightweight transformer-based model for dermatological multimodal lesion detection. In *2024 37th SIBGRAPI Conference on Graphics, Patterns and Images (SIBGRAPI)*, pages 1–6. DOI: 10.1109/SIBGRAPI62404.2024.10716324.

Tuncer, T., Barua, P. D., Tuncer, I., Dogan, S., and Acharya, U. R. (2024). A lightweight deep convolutional neural network model for skin cancer image classification. *Applied Soft Computing*, page 111794. DOI: 10.1016/j.asoc.2024.111794.

Webster, J. J. and Kit, C. (1992). Tokenization as the initial phase in nlp. In *Proceedings of the 14th Conference on Computational Linguistics - Volume 4*, COLING '92, page 1106–1110, USA. Association for Computational Linguistics. DOI: 10.3115/992424.992434.

Wilcoxon, F. (1945). Individual Comparisons by Ranking Methods. *Biometrics Bulletin*, 1(6):80–83. DOI: 10.2307/3001968.