The more the merrier: the use of verbose metadata description in the multimodal classification of skin lesions

Ana T. R. S. Pereira[1], **Wyctor F. da Rocha**[1], **Pedro H. Bouzon**[1], **André G. C. Pacheco**[1], **Luis A. Souza Jr.**[1]

[1]Universidade Federal do Espírito Santo (UFES)
Programa de Pós-Graduação em Informática, Vitória – ES, Brasil

{ana.t.pereira, wyctor.rocha, p.bouzon}@edu.ufes.br
{apacheco, la.souza}@inf.ufes.br

Abstract. *CAD systems for skin cancer often rely on structured metadata with limited semantic depth. This study investigates replacing numeric/categorical encodings with verbose, anamnesis-like descriptions generated by Large Language Models (LLMs) as semantic transducers to convert structured metadata into verbose anamnesis-like descriptions for multimodal skin lesion classification. Clinical attributes from PAD-UFES-20 and ISIC-2019 were converted into natural language, encoded via SBERT, and fused with CNN visual features through the MetaBlock-SE architecture. Patient-wise 5-fold cross-validation shows that LLM-based models achieve competitive or statistically superior Accuracy and AUC compared to traditional baselines. These findings indicate that verbose metadata descriptions constitute a flexible and semantically rich alternative for multimodal skin lesion classification, particularly in scenarios with limited or semantically sparse metadata.*

Resumo. *Sistemas CAD para câncer de pele frequentemente utilizam metadados estruturados com baixa profundidade semântica. Este trabalho investiga a substituição de codificações numéricas por descrições textuais detalhadas, semelhantes a anamneses, geradas por Modelos de Linguagem de Grande Porte (LLMs) como transcritores semânticos para converter metadados estruturados em descrições textuais verbosas, semelhantes a anamneses, aplicadas à classificação multimodal de lesões cutâneas. Os atributos dos datasets PAD-UFES-20 e ISIC-2019 foram convertidos em linguagem natural, codificados com SBERT e combinados a CNNs por meio da arquitetura MetaBlock-SE. Validações cruzadas (5-fold, patient-wise) mostram que modelos baseados em LLMs alcançam Acurácia e AUC competitivas ou estatisticamente superiores ao baseline. Os resultados indicam que descrições verbosas de metadados constituem uma alternativa flexível e semanticamente rica para classificação multimodal de lesões cutâneas, especialmente em cenários com metadados limitados ou semanticamente pouco informativos.*

1. Introduction

The World Health Organization (WHO) estimates that skin cancer accounts for approximately 30% of all cancer types diagnosed worldwide, making it the most common neoplasm globally [OMS 2017]. In Brazil, the National Cancer Institute (INCA) estimates that 263,280 new cases of skin cancer will be diagnosed during the 2026–2028 triennium,

making it the most common type of cancer in the country [INCA 2026]. In the state of Espírito Santo, disease incidence follows the same pattern, with an estimated 6,930 new cases of skin cancer in 2026 alone, corresponding to approximately 42% of all cancer cases in the state [INCA 2026].

This high incidence can be partially explained by European immigration, mainly from Pomerania (a region between Germany and Poland), which occurred in the state during the 19th century [Granzow 2009]. These immigrants typically had very fair skin and relied on family farming as their primary means of subsistence [Granzow 2009]. More than a century later, their descendants largely retain the same skin phenotype and rural activities as their main source of income. Therefore, the combination of prolonged sun exposure required by rural labor and low skin pigmentation has become a major factor contributing to the high incidence of skin cancer in this community. The late diagnosis can significantly impact the remission rate, leading to increased mortality. Consequently, dermatologists' skin cancer diagnoses are based on visual analysis of suspicious lesions, combined with disease-related information obtained through medical anamnesis [1].

A great challenge is the accurate diagnosis and specialized training in dermoscopy, a non-invasive diagnostic technique that uses a dermatoscope to magnify superficial skin structures and highlight morphological features that are not visible to the naked eye [Argenziano and Soyer 2001]. Studies by [Yélamos et al. 2019] and [Sinz et al. 2017] demonstrated that dermoscopy significantly improves diagnostic accuracy; however, the human factor remains critical, as the dermatologist's experience strongly influences diagnostic performance. Thus, the high incidence of skin cancer and the shortage of medical devices and specialists, particularly in rural areas [Feng et al. 2018].

Recent Transformer-based technologies have further enhanced CAD systems by enabling the integration of image and textual data [Pacheco et al. 2020, Souza et al. 2024, Bouzon et al. 2025]. For instance, Souza Jr. et al. [Souza et al. 2024] proposed a lightweight model designed to serialize and generalize multimodal features. By leveraging clinical information and images from the PAD-UFES-20 and ISIC-2019 datasets, their approach demonstrates the strong generalization potential of these architectures in skin lesion classification.

This study investigates that question by transforming clinical metadata from PAD-UFES-20 and ISIC-2019 into verbose natural-language descriptions generated by different LLM families and parameter scales. These descriptions are encoded with SBERT and fused with CNN visual features through the MetaBlock-SE architecture [Bouzon et al. 2025]. Rather than using LLMs as diagnostic agents, we employ them as semantic transducers that enrich the representation of structured metadata. Our goal is therefore not to claim universal superiority over structured baselines, but to assess whether verbose representations can provide a flexible and competitive alternative, particularly in scenarios with limited or semantically sparse metadata.

The main contributions of this study are:

- We propose and evaluate a framework that converts structured clinical metadata

[1] A stage of the medical consultation in which detailed information about the patient's medical history, current symptoms, pre-existing conditions, risk factors, and related aspects is collected.

into verbose anamnesis-like descriptions using LLMs for multimodal skin lesion classification.
- We analyze how LLM scale and dataset metadata richness affect the stability and effectiveness of text-based metadata representations.
- We show that verbose metadata representations can achieve competitive performance against strong structured baselines, with more evident benefits in low-metadata scenarios.

2. Related Work

Deep learning, primarily through CNNs, dominates automated skin lesion classification [Qasim Gilani et al. 2023]. Since clinical diagnosis is inherently multimodal—integrating visual cues with metadata like age and anatomical site—multimodal learning has become a research standard [Pacheco and Krohling 2020]. Early methods typically employed late fusion or feature concatenation [Souza et al. 2024]; however, these simple strategies often fail to capture complex cross-modal interactions.

Early multimodal approaches typically relied on feature concatenation or late-fusion strategies [Pacheco and Krohling 2021, Souza et al. 2024]. In particular, MetaBlock-based architectures, which impose attention-based mechanisms to condition image representation on metadata, have shown that metadata-aware reweighting can improve multimodal classification, especially in the presence of missing or heterogeneous clinical attributes [Bouzon et al. 2025].

Despite these advances, most multimodal skin lesion classification methods still rely on structured metadata represented through numerical, categorical, or dense embedding-based encodings. While effective, such representations may provide limited semantic expressiveness when compared with the richer contextual structure naturally present in clinical narratives. Beyond medical imaging, a growing body of work has investigated the transformation of structured data into textual representations to enable more expressive modeling. In natural language processing, tokenization and representation strategies are essential for balancing efficiency and semantic fidelity [Wang et al. 2025].

In multimodal learning, language often serves as an intermediate representation to bridge heterogeneous data sources. Textual descriptions encode relational and contextual information that symbolic or numeric formats lack. For example, the Autoregressive Voken Generation (AVG) method [Li et al. 2025] represents images as sequences of semantic tokens ("vokens"), facilitating superior alignment between visual and textual modalities. Although this line of research has shown promising results in Neural Language Programming and general multimodal settings, its application to clinical metadata in skin lesion classification remains limited, with a few representations of scientific works using natural-language descriptions derived from patient metadata to serve as an effective alternative to conventional structured metadata representations in dermatological CAD systems.

Large Language Models (LLMs) have recently demonstrated remarkable capabilities in generating coherent and contextually rich natural-language descriptions. Based on Transformer architectures [Vaswani et al. 2017], these models learn high-dimensional semantic representations, which enable them to capture complex relationships, dependencies, and contextual cues that often elude symbolic or sparse encodings. Prior work has

explored the use of LLMs for transforming tabular, categorical, or attribute-based data into textual representations, particularly as a mean to improve downstream learning and reasoning tasks. For instance, studies on data-to-text generation have shown that neural language models can faithfully verbalize structured records while preserving salient information and introducing contextual coherence based on semantically meaningful text, suitable for embedding-based learning [Wiseman et al. 2017, Gardent et al. 2017, Liu et al. 2023, Xing and Wan 2021].

Sentence-level semantic embeddings further allow these descriptions to be mapped into dense vector spaces that preserve relational structures [Reimers and Gurevych 2019]. In multimodal settings, text serves as an intermediate semantic interface [Radford et al. 2021], offering greater flexibility and robustness to missing attributes than traditional one-hot or numeric encodings. However, the use of LLM-generated anamnesis-like descriptions as metadata representations in multimodal dermatology remains underexplored, since the effectiveness of this transformation may depend on model scale and generation fidelity.

Hence, this work aims to address such gaps by evaluating whether clinical metadata verbalized by different LLM families and scales can support multimodal skin lesion classification when encoded with SBERT and fused with visual features through the MetaBlock-SE architecture.

3. Methodology

3.1. Datasets

Two publicly available skin lesion datasets were used to evaluate the proposed approach: (i) PAD-UFES-20 [Pacheco et al. 2020] and (ii) ISIC 2019 [ISIC2019 2019]. These datasets were selected because they provide complementary multimodal settings with distinct levels of metadata richness.

PAD-UFES-20 comprises 2,298 clinical images, 21 patient clinical attributes, and six lesion classes, including Melanoma, Basal Cell Carcinoma, Squamous Cell Carcinoma, Actinic Keratosis, Seborrheic Keratosis, and others. In contrast, ISIC 2019 contains 25,331 dermoscopic images, only three clinical attributes (age, sex, and anatomical site), and eight lesion classes, including all lesion types present in PAD-UFES-20 as well as Dermatofibroma and Vascular Lesions.

This contrast is particularly relevant for the present study: PAD-UFES-20 dataset presents rich structured clinical information while ISIC 2019 presents sparse metadata. Before sentence generation, exploratory data analysis was conducted to identify inconsistencies in metadata collection, standardization, and completeness, ensuring consistency in the available clinical attributes.

3.2. Generation of Descriptive Metadata Using LLMs

Different from conventional multimodal skin lesion classification methods, we investigate an alternative representation strategy in which clinical attributes are converted into verbose anamnesis-like descriptions before being encoded and fused with visual features, instead of plain and semantically-limited optins ([Pacheco and Krohling 2021]). Our method reproduces anamnesis in free-text format, employing detailed information enriched with linguistic context to feed a multimodal model capable of processing complete

sentences jointly with lesion images. The underlying hypothesis is that simulating the detailed behavior of specialist-conducted anamnesis, the use of LLMs may yield improved performance.

Three Large Language Models (LLMs) families were employed for this purpose: DeepSeek [Liu et al. 2025], Gemma [Mesnard et al. 2024], and Qwen [Yang et al. 2025]. Multiple parameter scales were evaluated for each family to analyze how model size affects the stability, consistency, and usefulness of the generated descriptions in the downstream multimodal classification task. [2]. To reduce semantic drift, preserve data fidelity, and prevent hallucinated content, the generation process followed a controlled prompting strategy based only on the attributes available in each dataset. Missing, false, or unspecified fields were omitted rather than inferred.

3.3. Experimental Delineation

To mitigate bias, a 5-fold cross-validation protocol stratified by lesion type and patient was adopted. This strategy ensures that the data are partitioned into five distinct subsets, with each iteration using four subsets for training and the remaining one for testing. For the PAD-UFES-20 dataset, stratification was performed not only by diagnosis but also at the patient level to prevent the same individual from appearing simultaneously in both the training and testing sets, which could lead to information leakage and compromise the evaluation.

Model hyperparameters were tuned throughout the experiments. The learning rate was explored over several orders of magnitude, from 10^{-3} to 10^{-5}, to assess its impact on convergence during training. An early stopping criterion was also employed, with a patience window of 10–20 epochs, and training was interrupted when no improvement in validation metrics was observed, thereby preventing overfitting and promoting better generalization. In addition, three different CNN architectures were used as model backbones, namely CAFormer-S18 [Yu et al. 2024], MobileNet-V2 [Sandler et al. 2018], and ResNet-50 [He et al. 2016], enabling a comparative analysis of the impact of architectural choices on the final system performance.

Large Language Models were employed to generate sentences from clinical metadata, yielding the verbose metadata used to feed the classification model. Models from different families and parameter scales were considered, allowing the influence of model size and architecture on the quality of the generated textual representations to be evaluated. Specifically, the `DeepSeek-R1` models were tested in their 1.5B and 70B parameter versions; `Qwen3` in its 1.7B and 32B versions; and `Gemma3` in its 1B and 27B parameter versions. This diversity of models enabled a comparative analysis of how increasing parameter scale impacts linguistic expressiveness and the usefulness of the generated sentences as inputs for multimodal classifiers.

Furthermore, the proposed evaluation protocol focused on the binary classification task of discriminating between cancerous and non-cancerous lesions. Therefore, the diagnostic categories Actinic Keratosis, Nevus, Seborrheic Keratosis, Dermatofibroma, and Vascular Lesions were grouped as “non-cancerous”, while Basal Cell Carcinoma, Squamous Cell Carcinoma, and Melanoma were grouped as “cancerous”. Our main goal is to

[2]Sentence generator github project: https://github.com/life-ufes/Simple-RAG

assess whether verbose metadata, in conjunction with image representations, is sufficient to support the separation of cancerous and non-cancerous lesions.

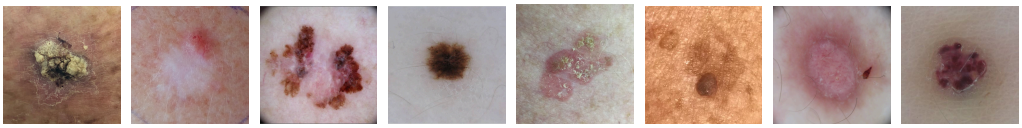

Figure 1. Samples of skin lesions depicting clinical and dermatoscopic images from both PAD-UFES-20 and ISIC 2019 datasets.

3.4. Evaluation Metrics

The experimental evaluation used three primary metrics: Accuracy (ACC), Balanced Accuracy (BACC), and the Area Under the ROC Curve (AUC). In scenarios with class imbalance (such as the one we are dealing with), ACC may not fully reflect the model's true performance. Hence, BACC was also considered. The AUC metric, in turn, evaluates the model's discriminative capability across different decision thresholds, allowing its robustness to be assessed under varying operating conditions. The combined use of these metrics provides a more comprehensive and reliable analysis of model performance in classifying skin lesions.

3.5. Classification Model – MetaBlock-SE

To evaluate the quality of the LLM-generated narratives, we employ the multimodal MetaBlock-SE architecture [Bouzon et al. 2025]. The model's classification performance serves as a proxy to determine how effectively the generated sentences represent the underlying clinical context compared to strong structured metadata baselines.

MetaBlock-SE extends the original MetaBlock design by replacing conventional structured metadata representations with sentence embeddings generated via SBERT [Reimers and Gurevych 2019]. While structured encodings are effective for strictly organized data, they may provide limited semantic depth and can be sensitive to missing or noisy attributes. In contrast, MetaBlock-SE utilizes dense, contextualized representations to preserve the clinical meaning of the anamnesis. By leveraging a semantic interface rather than sparse or rigid structured representations, MetaBlock-SE provides a richer and more resilient representation, particularly when dealing with unstructured or inconsistent data typical of real-world clinical records.

3.6. Statistical Analysis

To assess whether observed performance differences between models were statistically significant, we conducted paired Wilcoxon signed-rank tests [Wilcoxon 1945] to compare the proposed verbose-based models against each baseline and against each other, with a significance of 5%.

4. Results and Discussion

This section presents and discusses the experimental results obtained with the proposed approach, focusing on the impact of using verbose metadata generated by Large Language

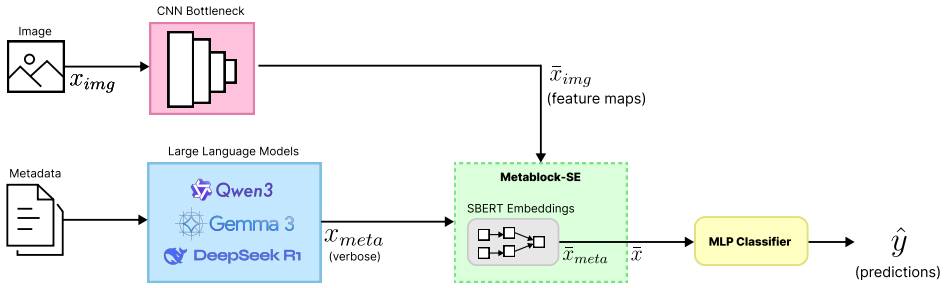

Figure 2. Overview of the MetaBlock-SE-based multimodal pipeline used for skin lesion classification.

Models (LLMs) for multimodal skin lesion classification. The analysis compares the baseline MetaBlock-SE model, which relies on structured metadata representations, with its variants that incorporate free-text descriptions of clinical information.

4.1. Sentence generation

Verbose clinical narratives were generated from structured attributes using a semi-structured prompting strategy. The objective was to transform tabular data into concise, anamnesis-like sentences while strictly maintaining data fidelity. To prevent hallucinations, the models were explicitly instructed to omit any attributes marked as `N/A`, `False`, or unspecified, rather than inferring missing information.

To ensure reproducibility and semantic alignment across different LLMs, the prompt employed a guided output schema: enforcing a consistent logical sequence while allowing the models to synthesize the attributes into a coherent medical summary. Key constraints included: (i) using the neutral term "patient" for unknown genders; (ii) prohibiting external case comparisons; and (iii) employing natural transitions between clinical fields.

> **Prompting Template:** *Generate a concise anamnesis summary using only provided data; do not infer or fabricate. Omit 'N/A' or 'False'. If gender is unknown, use 'patient'. Use placeholders: [ID, AGE, GENDER, LOCATION, SIZE, FITZPATRICK, FAMILY_HIST, ENV_FACTORS, MED_HIST, LIFESTYLE, SYMPTOMS]. **Output Format:** Patient [ID] is a [AGE]-year-old [GENDER] with a lesion on the [LOCATION] ([SIZE] mm). History: [MED_HIST], [FAMILY_HIST]. Factors: [ENV_FACTORS], [LIFESTYLE]. Symptoms: [SYMPTOMS].*

To illustrate the transformation process, an anonymized example of generated descriptions is shown below.

> **Example (PAD-UFES-20 style):** *Patient 102 is a 65-year-old male with a lesion on the back. History includes previous skin cancer. Environmental factors include chronic sun exposure.*

This controlled generation process ensured that the resulting narratives were semantically rich and compatible with the downstream SBERT encoding pipeline.

4.2. Exploratory analysis of verbose metadata embeddings

To investigate the structure induced by the LLM-generated sentences, all verbose descriptions were encoded using the *emilyalsentzer/Bio_ClinicalBERT* [3] SBERT model, producing 384-dimensional embeddings. K-means clustering was applied for $k \in [2, 10]$, and the optimal number of clusters was selected based on the silhouette coefficient. For both PAD-UFES-20 and ISIC-2019, the silhouette score consistently peaked at $k = 2$, suggesting the presence of two dominant semantic groupings in the embedding space.

Although the datasets contain multiple diagnostic subclasses, this clustering behavior aligns more closely with the binary cancer vs. non-cancer formulation adopted in our experiments. This indicates that verbose metadata embeddings tend to encode coarse-grained clinical separability rather than fine-grained subclass distinctions, likely reflecting stronger correlations among attributes such as age, lesion location, family history, and overall malignancy status. Due to space constraints, full UMAP visualizations and clustering diagnostics are available in the project repository[4].

4.3. Quantitative Results

Table 1. Classification results from PAD-UFES-20 dataset. Values in bold mean statistical similarity ($p \geq 0.05$), and $\star$ means the results from the best setup.

Model	LLM	Backbone	ACC	BACC	AUC
Baseline	-	MobileNet-V2	0.89 ± 0.01	0.89 ± 0.01	0.95 ± 0.01
		CAFormer-S18	0.89 ± 0.01	0.89 ± 0.01	0.95 ± 0.01
		ResNet-50	0.90 ± 0.01	0.90 ± 0.02	0.96 ± 0.01
Verbose Metablock-SE	Gemma3:1b	MobileNet-V2	0.88 ± 0.02	0.89 ± 0.02	0.95 ± 0.02
		CAFormer-S18	0.88 ± 0.02	0.88 ± 0.02	0.95 ± 0.02
		ResNet-50	0.88 ± 0.02	0.89 ± 0.03	0.96 ± 0.01
	Gemma3:27b	MobileNet-V2	0.89 ± 0.02	0.89 ± 0.02	0.96 ± 0.02
		CAFormer-S18	0.89 ± 0.0	0.89 ± 0.02	0.96 ± 0.01
		ResNet-50	0.89 ± 0.02	0.89 ± 0.02	0.96 ± 0.02
	Qwen3:1.7b	MobileNet-V2	0.89 ± 0.02	0.89 ± 0.02	0.96 ± 0.01
		CAFormer-S18	0.89 ± 0.02	0.89 ± 0.02	0.95 ± 0.01
		ResNet-50	0.90 ± 0.02	0.90 ± 0.02	0.96 ± 0.01
	Qwen3:32b	MobileNet-V2	0.90 ± 0.02	0.90 ± 0.02	0.96 ± 0.02
		CAFormer-S18	0.90 ± 0.03	0.90 ± 0.03	0.95 ± 0.02
		$\star$ **ResNet-50**	0.91 ± 0.03	0.91 ± 0.03	0.96 ± 0.02
	Deepseek-r1:1.5b	MobileNet-V2	0.87 ± 0.02	0.87 ± 0.02	0.94 ± 0.02
		CAFormer-S18	0.87 ± 0.02	0.87 ± 0.02	0.94 ± 0.02
		ResNet-50	0.87 ± 0.03	0.87 ± 0.03	0.94 ± 0.02
	Deepseek-r1:70b	MobileNet-V2	0.89 ± 0.02	0.89 ± 0.02	0.96 ± 0.02
		CAFormer-S18	0.89 ± 0.02	0.89 ± 0.02	0.96 ± 0.01
		ResNet-50	0.89 ± 0.02	0.89 ± 0.02	0.96 ± 0.02

Tables 1 and 2 summarize the classification performance on the PAD-UFES-20 and ISIC-2019 datasets, respectively, considering Accuracy (ACC), Balanced Accuracy (BACC), and Area Under the ROC Curve (AUC). Across both datasets, the baseline MetaBlock-SE model achieved competitive and stable performance, confirming the robustness of the multimodal architecture when integrating visual features with structured metadata representations.

[3]Link of the used sentence encoder model: `https://github.com/EmilyAlsentzer/clinicalBERT`

[4]https://github.com/life-ufes/Simple-RAG

Table 2. Classification results from ISIC-2019 dataset. Values in bold mean statistical similarity ($p \geq 0.05$), and $\star$ means the results from the best setup.

Model	LLM	Backbone	ACC	BACC	AUC
Baseline	-	MobileNet-V2	0.83 ± 0.01	0.83 ± 0.01	0.97 ± 0.00
		CAFormer-S18	0.88 ± 0.01	0.86 ± 0.01	0.98 ± 0.00
		ResNet-50	0.86 ± 0.01	0.85 ± 0.01	0.98 ± 0.01
Verbose Metablock-SE	Gemma3:1b	MobileNet-V2	0.90 ± 0.00	0.88 ± 0.01	0.95 ± 0.00
		CAFormer-S18	0.93 ± 0.01	0.91 ± 0.01	0.97 ± 0.01
		ResNet-50	0.91 ± 0.01	0.88 ± 0.01	0.96 ± 0.01
	Gemma3:27b	MobileNet-V2	0.91 ± 0.01	0.88 ± 0.01	0.96 ± 0.01
		CAFormer-S18	0.93 ± 0.01	0.91 ± 0.01	0.97 ± 0.00
		ResNet-50	0.91 ± 0.01	0.89 ± 0.01	0.96 ± 0.01
	Qwen3:1.7b	MobileNet-V2	0.90 ± 0.01	0.87 ± 0.01	0.95 ± 0.01
		CAFormer-S18	0.92 ± 0.01	0.90 ± 0.01	0.97 ± 0.01
		ResNet-50	0.92 ± 0.01	0.89 ± 0.01	0.96 ± 0.01
	Qwen3:32b	MobileNet-V2	0.91 ± 0.00	0.89 ± 0.00	0.96 ± 0.00
		$\star$ **CAFormer-S18**	0.93 ± 0.01	0.91 ± 0.01	0.97 ± 0.01
		ResNet-50	0.91 ± 0.01	0.88 ± 0.02	0.96 ± 0.01
	Deepseek-r1:1.5b	MobileNet-V2	0.89 ± 0.01	0.86 ± 0.02	0.95 ± 0.01
		CAFormer-S18	0.92 ± 0.01	0.89 ± 0.02	0.96 ± 0.01
		ResNet-50	0.91 ± 0.01	0.89 ± 0.01	0.96 ± 0.00
	Deepseek-r1:70b	MobileNet-V2	0.91 ± 0.01	0.88 ± 0.01	0.96 ± 0.01
		CAFormer-S18	0.92 ± 0.01	0.90 ± 0.02	0.96 ± 0.01
		ResNet-50	0.91 ± 0.00	0.88 ± 0.01	0.96 ± 0.00

On PAD-UFES-20, which provides a rich set of clinical attributes, the use of verbose metadata led to noticeable variations in performance depending on the LLM and backbone employed. In several configurations, particularly with medium- and large-scale LLMs such as Gemma3 and Qwen3, the performance remained close to the baseline, with AUC values consistently above 0.90. However, some configurations exhibited statistically significant differences in ACC and BACC when compared to the baseline, as confirmed by the Wilcoxon signed-rank test ($p < 0.05$).

Likewise, in the ISIC-2019 evaluation (Table 2), which contains a smaller set of clinical attributes (age, sex, and anatomical site), models using verbose metadata achieved better results than the baseline across several configurations. In addition, most LLM-based variants showed statistically significant differences, suggesting that transforming structured metadata into natural-language descriptions may be particularly beneficial in low-metadata settings.

4.4. Discussion

Overall, the results of the presented work indicate that transforming tabular metadata into natural language is not only feasible but can also achieve performance that is statistically equivalent to, or in some cases better than, strong structured baselines.

A central finding concerns the interaction between dataset characteristics and the effectiveness of metadata verbalization. In PAD-UFES-20, which provides 21 clinical attributes per patient, the baseline structured representation already offers a rich representation space. Consequently, verbose representations yielded performance that was mostly statistically equivalent to the baseline. This suggests that when metadata are abundant and well-structured, the primary advantage of verbalization lies not necessarily in boosting accuracy, but in offering a more flexible and semantically coherent representation that

remains robust to inconsistencies. In contrast, the ISIC-2019 dataset contains only three clinical attributes (age, sex, and anatomical site). In this scenario, the transformation into coherent clinical narratives led to statistically significant improvements over the baseline in several configurations, suggesting that natural language may amplify weak structured information by embedding it into a richer semantic context, allowing sentence encoders to capture interactions that simpler structured encodings may fail to model.

An important observation from the experiments concerns the impact of LLM size on description quality and usefulness. Smaller models, such as DeepSeek-R1 (1.5B parameters), more frequently generated inconsistent or semantically fragile sentences, occasionally degrading performance. This likely reflects limited contextual modeling capacity, increasing susceptibility to errors when processing missing or noisy metadata. Large-scale LLMs demonstrated greater stability, producing more coherent and clinically plausible descriptions that were effectively leveraged by the downstream classifier without substantial loss in predictive performance (limited by a certain threshold, indicating diminishing returns in classification accuracy).

These findings indicate that, rather than consistently outperforming structured encodings, the proposed approach demonstrates that verbose metadata representations constitute a viable and flexible alternative, particularly in scenarios where clinical attributes are limited or semantically sparse. For the task of clinical metadata verbalization, model fidelity and consistency are more critical than sheer generative expressiveness – something particularly relevant for deployment scenarios where computational efficiency must be balanced with reliability.

5. Conclusion

This work investigated the use of Large Language Models (LLMs) as semantic transducers that convert structured clinical metadata into verbose, anamnesis-like textual descriptions for multimodal skin lesion classification. Rather than using LLMs as diagnostic agents, the proposed pipeline exploits their capacity to produce linguistically meaningful representations that can be embedded and fused with visual features. Experiments on PAD-UFES-20 and ISIC-2019 indicate that verbose metadata can serve as an effective substitute for traditional numerical and one-hot encodings, yielding statistically comparable performance across different CNN backbones and supporting free-text as a viable metadata representation in multimodal CAD settings. We observed that the effect of metadata verbalization depends on both dataset characteristics and LLM scale. In general, large-scale LLMs produced more stable descriptions, while smaller models were more prone to semantic drift and inconsistencies. These findings suggest that LLM-based metadata verbalization offers a flexible alternative for multimodal systems, especially when metadata schemas are heterogeneous or partially incomplete.

Acknowledgments

The authors thank the Espírito Santo Research Foundation (FAPES); the Capixaba Institute of Health Education, Research, and Innovation (ICEPi); the National Council for Scientific and Technological Development (CNPq); the Department of Science and Technology of the Ministry of Health (Decit/SECTICS/MS); Brazilian National Program of Genomics and Precision Health (Genomas Brasil); and the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Finance Code 001.

References

Argenziano, G. and Soyer, H. P. (2001). Dermoscopy of pigmented skin lesions – a valuable tool for early diagnosis of melanoma. *The Lancet Oncology*, 2(7):443–449.

Bouzon, P. H. G. et al. (2025). Metablock-se: A method to deal with missing metadata in multimodal skin cancer classification. *IEEE Journal of Biomedical and Health Informatics*, 29(12):8855–8862.

Feng, H. et al. (2018). Comparison of dermatologist density between urban and rural counties in the United States. *JAMA Dermatology*, 154(11):1265–1271.

Gardent, C. et al. (2017). The WebNLG challenge: Generating text from RDF data. In Alonso, J. M., Bugarín, A., and Reiter, E., editors, *Proceedings of the 10th International Conference on Natural Language Generation*, pages 124–133, Santiago de Compostela, Spain. Association for Computational Linguistics.

Granzow, K. (2009). *Pomeranos sob o Cruzeiro do Sul: colonos alemães no Brasil*. Arquivo Público do Estado do Rio de Janeiro.

He, K., Zhang, X., Ren, S., and Sun, J. (2016). Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778.

INCA (2026). Incidência do câncer no Brasil. Instituto Nacional do Câncer (INCA). Disponível em: https://ninho.inca.gov.br/jspui/handle/123456789/17914. Último acesso em: 11 de Fevereiro 2026.

ISIC2019 (2019). Skin lesion analysis towards melanoma detection. Skin Image Collaboration. Disponível em: https://www.isic-archive.com/. Último acesso em: 29 de Maio 2024.

Li, Y., Cai, H., Wang, W., Qu, L., Wei, Y., Li, W., Nie, L., and Chua, T.-S. (2025). Revolutionizing text-to-image retrieval as autoregressive token-to-voken generation. In *Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '25, page 813–822, New York, NY, USA. Association for Computing Machinery.

Liu, A. et al. (2025). Deepseek-v3 technical report.

Liu, P. et al. (2023). Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *ACM Comput. Surv.*, 55(9).

Mesnard, T. et al. (2024). Gemma: Open models based on gemini research and technology.

OMS (2017). Radiation: Ultraviolet (UV) radiation and skin cancer. World Health Organization (WHO). Disponível em: http://www.who.int/uv/faq/skincancer/en/index1.html. Último acesso em: 05 de Junho 2023.

Pacheco, A. G. et al. (2020). Pad-ufes-20: A skin lesion dataset composed of patient data and clinical images collected from smartphones. *Data in Brief*, 32:106221.

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.

Pacheco, A. G. C. and Krohling, R. (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*. In press.

Qasim Gilani, S., Syed, T., Umair, M., and Marques, O. (2023). Skin cancer classification using deep spiking neural network. *Journal of Digital Imaging*, 36(3):1137–1147.

Radford, A. et al. (2021). Learning transferable visual models from natural language supervision. *arXiv preprint arXiv:2103.00020*.

Reimers, N. and Gurevych, I. (2019). Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084*.

Sandler, M. et al. (2018). Mobilenetv2: Inverted residuals and linear bottlenecks. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4510–4520.

Sinz, C. et al. (2017). Accuracy of dermatoscopy for the diagnosis of nonpigmented cancers of the skin. *Journal of the American Academy of Dermatology*, 77(6):1100–1109.

Souza, L. A. et al. (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. IEEE.

Vaswani, A. et al. (2017). Attention is all you need. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, page 6000–6010, Red Hook, NY, USA. Curran Associates Inc.

Wang, S., Guo, W., Chen, Z., Xu, Y., Hu, X., and Xiong, H. (2025). Less is more: Token-efficient video-qa via adaptive frame-pruning and semantic graph integration.

Wilcoxon, F. (1945). Individual comparisons by ranking methods. *Biometrics*, 1:196–202.

Wiseman, S., Shieber, S., and Rush, A. (2017). Challenges in data-to-document generation. In Palmer, M., Hwa, R., and Riedel, S., editors, *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2253–2263, Copenhagen, Denmark. Association for Computational Linguistics.

Xing, X. and Wan, X. (2021). Structure-aware pre-training for table-to-text generation. In *Findings of the association for computational linguistics: ACL-IJCNLP 2021*, pages 2273–2278.

Yang, A. et al. (2025). Qwen3 technical report.

Yu, W. et al. (2024). Metaformer baselines for vision. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(2):896–912.

Yélamos, O. et al. (2019). Usefulness of dermoscopy to improve the clinical and histopathologic diagnosis of skin cancers. *Journal of the American Academy of Dermatology*, 80(2):365–377.