XLVI CILAMCE
VITÓRIA - ES - BRAZIL - 2025
XLVI IBERO-LATIN AMERICAN CONGRESS ON COMPUTATIONAL METHODS IN ENGENEERING

Automated detection and diagnosis of skin cancer using neural networks and multi-agent systems

Andrey Alexandre de Souza Basilio[1], Daniel Cruz Cavalieri[2]

[1]*Dept. of Automation and Control Engineeering, Federal Institute of Espirito Santo*
Espirito Santo, Brazil
andreyalexandre16@gmail.com
[2]*Dept. of Automation and Control Engineeering, Federal Institute of Espirito Santo*
Espirito Santo, Brazil
daniel.cavalieri@ifes.edu.br

Abstract. This paper proposes a multi-agent system based on large-scale language models (LLMs) for automated skin cancer diagnosis from images. The architecture includes segmentation, classification, and report generation steps, integrating specialized agents based on models such as Gemini-2.5-pro and GPT-4o. The approach was evaluated using bootstrap sampling with five replicates of size 80, highlighting the efficiency of the visual agents and the consistency of the diagnoses generated. Preliminary results indicate the potential of integrating computer vision and natural language to support medical diagnosis.

Keywords: Skin cancer, Large Language Models, automated diagnosis, Multi-agents Systems, computer vision

1 Introduction

Skin cancer is widely prevalent globally, accounting for a large portion of oncological diagnoses in both developed and developing countries. According to estimates by [1], more than 30% of cancer cases in Brazil in the 2023-2025 triennium will be skin cancer. According to the Brazilian Society of Dermatology [2], the most common types are basal cell carcinoma and squamous cell carcinoma, while melanoma, although rarer, is the most aggressive variant and has the highest mortality rate. The incidence of skin cancer has increased in countries with high sun exposure, such as Brazil, due to intense ultraviolet (UV) radiation. Studies indicate that a lack of sun protection is a determining factor in this increase.

In addition to UV radiation exposure, factors such as genetic predisposition, skin pigmentation, sunburn history, and age have a significant impact on the development of skin cancer [3]. In the Brazilian context, raising awareness of early diagnosis is essential, especially in regions with limited access to specialists. Early detection is crucial to reducing mortality and treatment costs in advanced stages. In this scenario, artificial intelligence (AI)-based technologies emerge as a promising alternative, expanding access to diagnosis and increasing accuracy, even in areas with a shortage of specialists.

This research proposes the development of a multi-agent system based on Large Language Models, trained, validated, and tested with images from HAM10000 dataset provided by Harvard [4]. The goal is to maximize diagnostic accuracy and reduce the incidence of false positives in suspicious skin lesions. The proposal follows recent advances in the use of large-scale language models in medicine, which have been successfully applied to the interpretation of radiological images [5], the generation of medical reports [6], and clinical reasoning in complex diagnostic contexts [7]. Therefore, we seek to offer an innovative solution that combines diagnostic accuracy and accessibility, meeting the growing demands for effective and inclusive technologies in dermatology.

2 Methodology

The methodology of this work is structured in three main components: (i) the theoretical foundation necessary to support the choice of models; (ii) the proposed architecture of the multi-agent system; and (iii) the procedures adopted for data collection, processing and analysis.

2.1 U-Net Model for Image Segmentation

The U-Net model is a convolutional neural network architecture originally proposed for biomedical image segmentation [8]. As illustrated in Figure 1, the model features a symmetrical "U"-shaped topology composed of two main parts: the contraction path (encoder) and the expansion path (decoder).

In the contraction path, the network applies successive convolutional and max-pooling layers to extract hierarchical feature representations while progressively reducing spatial resolution. In the expansion path, the feature maps are upsampled and convolved to reconstruct spatial information and generate a final segmentation map. A key innovation of the U-Net is the use of skip connections that directly transfer feature maps from the encoder to the decoder, thereby preserving fine-grained spatial details that are often lost during downsampling.

This architecture has proven highly effective for medical imaging tasks that require accurate delineation of small or irregular regions, such as skin lesion segmentation. The model output is a binary mask $M(x, y) \in \{0, 1\}$, indicating for each pixel of the input image $I(x, y)$ whether it belongs to the lesion region. The network is trained using a segmentation loss function, such as the Binary Cross Entropy (BCE) or Dice Loss, defined as:

$$\mathcal{L}_{Dice} = 1 - \frac{2\sum_i p_i g_i + \epsilon}{\sum_i p_i + \sum_i g_i + \epsilon} \tag{1}$$

where p_i and g_i represent the predicted and ground-truth values for pixel i, and ϵ is a smoothing term to prevent division by zero.

In this work, we employ the Attention U-Net variant [9], which introduces attention gates into the skip connections, enabling the network to automatically focus on the most relevant spatial regions during segmentation. This mechanism is particularly advantageous for dermoscopic images, where the contrast between lesion and surrounding tissue can be subtle and heterogeneous.

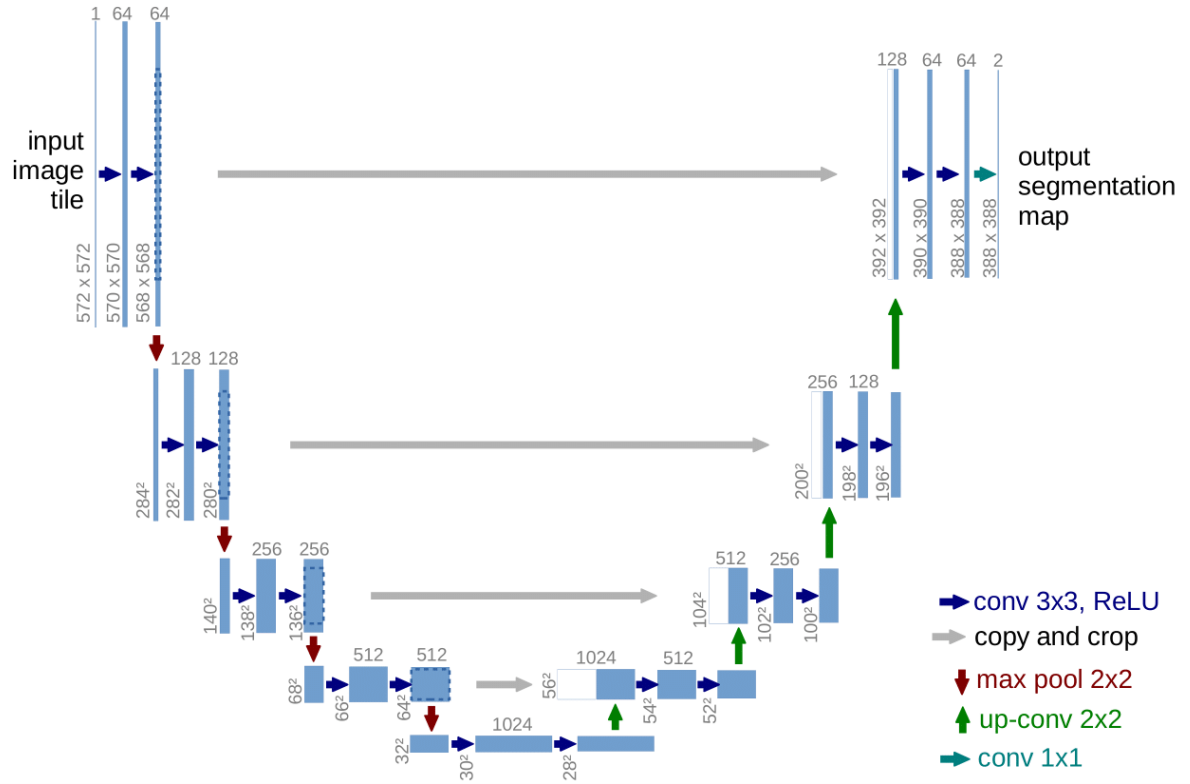


Figure 1. General architecture of the U-Net model, showing the encoder–decoder structure and skip connections. Adapted from Ronneberger et al. [8].

To train the model, dermoscopic images with manually annotated masks were used. The optimization process employed a combined Dice + BCE loss function and the Adam optimizer [10], which is widely used in deep learning due to its adaptive learning rate and momentum-based parameter updates. An empirically tuned learning rate was applied to achieve convergence stability. The U-Net family has since become a standard backbone for biomedical segmentation, being successfully applied to organs [11], tumors, and complex anatomical structures, reinforcing its robustness and versatility in medical imaging.

2.2 Large Language Models (LLMs)

Large Language Models (LLMs) are advanced neural network architectures trained on vast corpora of textual data to learn complex linguistic and semantic patterns. They are capable of performing a wide range of natural language processing (NLP) tasks, including question answering, summarization, and clinical report generation [7, 12, 13].

As illustrated in Figure 2, most state-of-the-art LLMs are based on the Transformer architecture [14], which employs a self-attention mechanism to capture long-range dependencies within text sequences. Given an input sequence of tokens $X = [x_1, x_2, \ldots, x_n]$, the model computes attention weights across all token pairs, enabling each token representation to incorporate information from the entire context.

The attention operation is defined as:

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

where Q, K, V correspond to the query, key, and value matrices, respectively, and d_k denotes the dimensionality of the key vector.

Recent advancements have expanded the capabilities of LLMs to multimodal processing, allowing them to interpret both textual and visual inputs simultaneously [13, 15]. These multimodal models, such as GPT-4o and Gemini, have shown promise in clinical contexts for generating structured medical reports that integrate textual and visual information—including the interpretation of diagnostic images, figures, and graphs [5].

In this work, we employ two distinct LLMs for diagnostic reasoning:

- **GPT-4o:** a multimodal model with advanced contextual understanding, employed to generate natural-language diagnostic reports;
- **Gemini-2.5-Pro:** a multimodal model with an extended context window, suitable for in-depth textual and visual analysis and comprehensive report generation.

To enhance the precision and factual alignment of model outputs, we apply *prompt engineering* and *Retrieval-Augmented Generation* (RAG) techniques. The RAG pipeline incorporates relevant clinical documents as an external knowledge source by embedding each document d into a vector representation $f(d) \in \mathbb{R}^n$. The system retrieves the most semantically similar documents to a given query q via cosine similarity:

$$\text{sim}(q, d) = \frac{f(q) \cdot f(d)}{\|f(q)\| \cdot \|f(d)\|} \tag{3}$$

This retrieval-augmented strategy enables the LLMs to generate diagnostic hypotheses and textual explanations that align more closely with established clinical guidelines, even in cases where visual information alone is insufficient.

2.3 Multi-Agent Systems

Multi-Agent Systems (MAS) are computational architectures composed of multiple autonomous agents that interact and collaborate to solve complex problems in a distributed and coordinated manner [16]. In this work, we propose a MAS specifically designed for automated skin cancer diagnosis, where each agent is instantiated from a Large Language Model (LLM) and specialized to operate according to the ABCD diagnostic protocol.

The architecture consists of the following stages:

- **Segmentation Agent:** isolates the lesion region from the input image using an Attention U-Net model. The resulting binary mask serves as a shared visual input for the subsequent diagnostic agents.
- **Diagnostic Agents:** three parallel LLM-based agents, each implementing the ABCD rule for dermatoscopic evaluation. Each agent is prompted to analyze one of the main diagnostic components — *Asymmetry*, *Border irregularity*, and *Color variation/Distribution*. Their individual textual outputs represent structured evaluations according to that criterion.
- **Summarizer Agent:** receives the textual analyses generated by the diagnostic agents and produces a unified diagnostic summary, integrating the evidence and formulating a consistent natural-language report.
- **Critical Review Agent:** performs a final verification step, evaluating the overall consistency, completeness, and plausibility of the diagnostic report. It simulates a peer-review or medical validation process by suggesting corrections or clarifications when necessary.

All agents are executed in isolated containerized environments and coordinated using the LangGraph library, which explicitly defines execution flows among nodes (either sequential or parallel). Figure 3 illustrates the proposed pipeline and the interaction between agents.

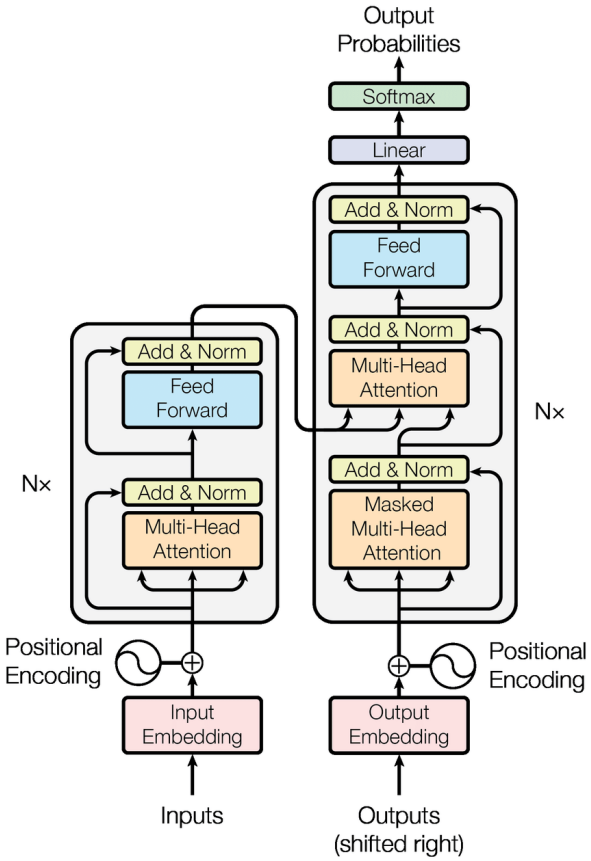


Figure 2. General architecture of a Transformer model, showing self-attention and feed-forward layers. Adapted from Vaswani et al. [14].

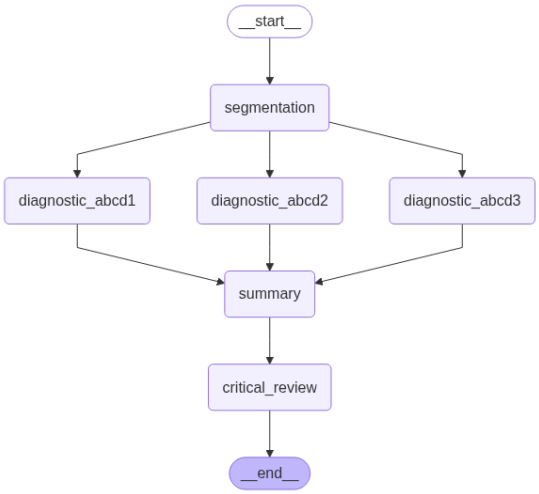


Figure 3. Execution flow of the proposed multi-agent system based on LLMs. Each node represents an agent operating either sequentially or in parallel.

The adoption of the ABCD rule as the diagnostic foundation is supported by strong clinical and computational evidence. Originally proposed by Nachbar et al. [17], the ABCD rule (Asymmetry, Border irregularity, Color variation, Diameter) has shown high predictive value in dermoscopic diagnosis of melanoma and other pigmented lesions. Its simplicity and interpretability make it suitable for computational modeling and decision-support systems. Messadi et al. [18] demonstrated that lesion segmentation combined with automated extraction of ABCD metrics can achieve accurate classification of benign and malignant lesions. More recently, Jütte et al. [19] integrated generative artificial intelligence with the ABCDE rule, showing that these morphological parameters remain highly correlated with lesion progression and malignancy risk. Additionally, Siddique et al. [20] proposed an intelligent system for multi-factor skin cancer detection, reinforcing the reliability of ABCD-based diagnostic reasoning in computer-aided systems.

The main advantage of the proposed multi-agent architecture lies in its semantic specialization. Each LLM-based agent is fine-tuned through prompt engineering to operate strictly within the boundaries of its diagnostic criterion. This modular structure allows for easy replacement or reconfiguration of agents as the system evolves. Unlike conventional deep learning models that perform direct classification through convolutional networks, this approach frames diagnosis as a language-driven reasoning process, enhancing transparency, interpretability, and clinical auditability of the diagnostic outcome.

2.4 Performance Analysis

The evaluation of the proposed system considered the traditional quantitative classification metrics of the reports generated by LLMs.

Reference Model (Baseline) As a baseline, we use a supervised-trained EfficientNet to classify segmented images as indicative of benign or malignant lesions. This model was chosen due to its widespread adoption in medical image classification tasks. The metrics extracted for comparison include:

- **Accuracy**: proportion of correct predictions to total;
- **Precision**: proportion of predicted positives that are true;
- **Sensitivity**: model's ability to detect positive cases;
- **F1-score**: harmonic mean between precision and sensitivity.

2.5 Practical Experiment

The practical experiment conducted in this research was structured into four sequential stages, as depicted in Figure 4, aiming to simulate an intelligent diagnostic pipeline based on a multi-agent system supported by large language models (LLMs).

1. Input and Preprocessing. The process begins with the reception of an image of the skin lesion, which undergoes a preprocessing step that includes segmentation of the lesion area. This segmentation aims to isolate the region of interest and provide enhanced visual information for downstream analysis, supporting both structure-focused and color-based assessment.

2. Multi-agent Analysis. The segmented image is processed through a two-layer multi-agent structure. In the first layer, three specialized agents independently analyze the lesion and generate individual diagnostic reports based on distinct heuristics and prompting strategies. These agents simulate diversified clinical reasoning, enabling redundancy and increased diagnostic robustness. In the second layer, a summarizing agent receives the outputs of the three diagnostic agents and synthesizes them into a unified diagnostic statement.

3. Academic Validation. The consolidated diagnosis is subjected to a validation step based on Retrieval-Augmented Generation (RAG). This agent uses a curated academic knowledge base to validate the consistency of the diagnosis, simulating an expert literature review. The result is a medically grounded statement that justifies or refines the synthesized diagnosis.

4. User Delivery. After academic validation, the system delivers the diagnostic and prognostic response to the end user in a clear and structured manner. This stage emphasizes explainability and traceability, essential for medical contexts.

The entire system is modular and agent-based, allowing for scalability, replacement of individual agents, and evaluation of different LLMs under identical experimental conditions. This design also facilitates future expansion with additional agents or logic layers, such as differential diagnosis or treatment recommendation.

This experimental structure demonstrates the feasibility of integrating vision, natural language understanding, and reasoning agents in a coordinated pipeline for medical diagnosis, simulating key aspects of human clinical practice.

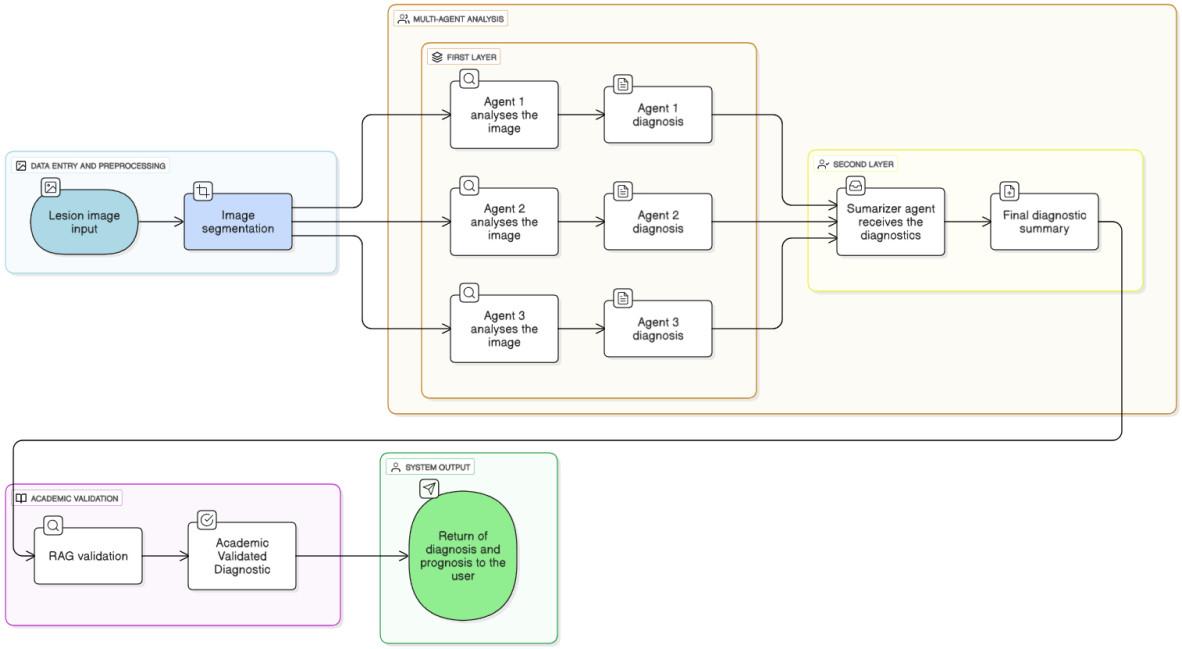


Figure 4. Execution flow of the proposed multi-agent system based on LLMs.

3 Results and Discussion

This section presents a comparative analysis of the proposed multi-agent system using LLMs and a traditional computer vision model (EfficientNet). The evaluation encompasses quantitative metrics (accuracy, precision, sensitivity, and F1-score) and qualitative aspects of interpretability and reasoning.

3.1 Quantitative Evaluation: Comparison Between Models

Table 1 summarizes the classification performance of the models tested. The EfficientNet served as the conventional baseline, while Gemini-2.5-Pro and GPT-4o were integrated into the system as vision-capable LLMs. In the case of LLMs, the final diagnostic label was extracted from the structured report generated by the system and compared to ground truth labels for evaluation.

Table 1. Quantitative performance metrics of models with standard deviation (±).

Model	**Accuracy**	**Precision**	**Sensitivity**	**F1-score**
EfficientNet	**0.84 ± 0.10**	**0.89 ± 0.13**	**0.78 ± 0.17**	**0.82 ± 0.15**
Gemini-2.5-Pro	0.35 ± 0.12	0.47 ± 0.25	0.34 ± 0.15	0.40 ± 0.19
GPT-4o (OpenAI)	0.47 ± 0.07	0.41 ± 0.21	0.31 ± 0.18	0.35 ± 0.19

Contrary to initial expectations, the LLMs underperformed in all quantitative metrics when compared to the EfficientNet. This result highlights the complexity of directly interpreting diagnostic categories from unstructured image inputs using generalist LLMs. Although GPT-4o and Gemini demonstrated high linguistic fluency and interpretability, their classification performance was limited, likely due to the lack of domain-specific fine-tuning and the high visual nuance required for skin cancer detection.

3.2 Discussion

Despite the lower quantitative results, the LLM-based system presents a distinct advantage in terms of interpretability and transparency. Each diagnostic report generated by the LLMs provided structured reasoning, including references to clinical heuristics such as asymmetry, border irregularities, and lesion color. This level of explanation offers a foundation for trust and auditability not achievable by standard convolutional classifiers.

Moreover, the multi-agent architecture enabled the simulation of peer-review-like behavior, with multiple agents contributing distinct diagnostic perspectives that were subsequently synthesized into a coherent report. This mirrors real clinical workflows and supports more robust and explainable decision-making.

The gap in performance between traditional models and LLMs also reinforces the need for domain adaptation. These results suggest that while current general-purpose vision-language models are not yet competitive in classification tasks, they could benefit substantially from supervised fine-tuning with curated dermatological datasets, enhancing both their accuracy and reliability.

4 Conclusion

This work proposed and implemented a multi-agent diagnostic system combining computer vision and large language models (LLMs) to perform automated skin cancer detection. The architecture was designed to integrate image segmentation, diagnostic reasoning based on clinical heuristics, and a summarization mechanism to produce structured, interpretable reports.

In quantitative evaluation, the EfficientNet model outperformed both Gemini-2.5-Pro and GPT-4o across all metrics, achieving 84% accuracy versus 35% and 47%, respectively. However, the LLM-based system demonstrated clear strengths in transparency and interpretability, providing detailed justifications for each diagnosis — a valuable feature in real-world clinical environments.

These findings reinforce the complementary nature of traditional and language-based approaches. While convolutional models excel at pattern recognition, LLMs bring reasoning and explainability to the diagnostic process. The limitations observed in accuracy suggest that general-purpose LLMs may require fine-tuning with domain-specific datasets to reach acceptable performance levels in medical diagnostics.

As future work, we propose:

- Implementing a fine-tuning process for LLMs using dermatology-specific corpora;
- Expanding the agent framework to include domain-specific roles (e.g., differential diagnosis, follow-up strategy);
- Integrating visual explainability modules (e.g., Grad-CAM) into the final reports.

The proposed multi-agent system represents a step forward in bridging interpretability and automation in medical diagnosis. Although the results are preliminary, they highlight both the limitations of generalist models and the promising future of specialized, explainable AI in healthcare.

Acknowledgements. The authors would like to express their sincere thanks to the Universidade Aberta Capixaba (UnAC), a program of the Government of the State of Espírito Santo, linked to the Secretaria de Estado da Ciência, Tecnologia, Inovação e Educação Profissional (Secti), for the institutional and financial support (T.O. 1064/2023) that enabled the realization of this study. We also thank the Federal Institute of Espírito Santo (Ifes) for the infrastructure provided.

Authorship statement. The authors hereby confirm that they are the sole liable persons responsible for the authorship of this work, and that all material that has been herein included as part of the present paper is either the property (and authorship) of the authors, or has the permission of the owners to be included here.

References

[1] de M. Oliveira Santos, de F. C. d. S. Lima, L. F. L. Martins, J. F. P. Oliveira, de L. M. Almeida, and de M. Camargo Cancela. Estimativa de incidência de câncer no brasil, 2023-2025. *Revista Brasileira de Cancerologia*, vol. 69, n. 1, 2023.

[2] Sociedade Brasileira de Dermatologia. Câncer da pele. `https://www.sbd.org.br/doencas/cancer-da-pele/`. Acesso em: outubro de 2024, 2024.

[3] R. Shraim, M. Z. Farran, G. He, J. Marunica Karsaj, L. Zgaga, and R. McManus. Systematic review on gene–sun exposure interactions in skin cancer. *Molecular genetics & genomic medicine*, vol. 11, n. 10, pp. e2259, 2023.

[4] P. Tschandl. The HAM10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions, 2018.
[5] S. Wang, Z. Zhao, X. Ouyang, Q. Wang, and D. Shen. Chatcad: Interactive computer-aided diagnosis on medical image using large language models, 2023.
[6] C. Li, C. Wong, S. Zhang, N. Usuyama, H. Liu, J. Yang, T. Naumann, H. Poon, and J. Gao. Llava-med: Training a large language-and-vision assistant for biomedicine in one day, 2023.
[7] K. Singhal, T. Tu, J. Gottweis, R. Sayres, E. Wulczyn, L. Hou, K. Clark, S. Pfohl, H. Cole-Lewis, D. Neal, M. Schaekermann, A. Wang, M. Amin, S. Lachgar, P. Mansfield, S. Prakash, B. Green, E. Dominowska, y B. A. Arcas, N. Tomasev, Y. Liu, R. Wong, C. Semturs, S. S. Mahdavi, J. Barral, D. Webster, G. S. Corrado, Y. Matias, S. Azizi, A. Karthikesalingam, and V. Natarajan. Towards expert-level medical question answering with large language models, 2023.
[8] O. Ronneberger, P. Fischer, and T. Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pp. 234–241. Springer, 2015.
[9] et O. al. Attention u-net: Learning where to look for the pancreas. *arXiv preprint arXiv:1804.03999*, 2018.
[10] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. *Proceedings of the 3rd International Conference on Learning Representations (ICLR)*, 2015.
[11] F. Milletari, N. Navab, and S.-A. Ahmadi. V-net: Fully convolutional neural networks for volumetric medical image segmentation. In *4th International Conference on 3D Vision (3DV)*, pp. 565–571. IEEE, 2016.
[12] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, and others. Language models are few-shot learners. *Advances in neural information processing systems*, 2020.
[13] OpenAI. Gpt-4 technical report, 2024.
[14] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, volume 30, 2017.
[15] K. Saab, T. Tu, W.-H. Weng, R. Tanno, D. Stutz, E. Wulczyn, F. Zhang, T. Strother, C. Park, E. Vedadi, J. Z. Chaves, S.-Y. Hu, M. Schaekermann, A. Kamath, Y. Cheng, D. G. T. Barrett, C. Cheung, B. Mustafa, A. Palepu, D. McDuff, L. Hou, T. Golany, L. Liu, baptiste J. Alayrac, N. Houlsby, N. Tomasev, J. Freyberg, C. Lau, J. Kemp, J. Lai, S. Azizi, K. Kanada, S. Man, K. Kulkarni, R. Sun, S. Shakeri, L. He, B. Caine, A. Webson, N. Latysheva, M. Johnson, P. Mansfield, J. Lu, E. Rivlin, J. Anderson, B. Green, R. Wong, J. Krause, J. Shlens, E. Dominowska, S. M. A. Eslami, K. Chou, C. Cui, O. Vinyals, K. Kavukcuoglu, J. Manyika, J. Dean, D. Hassabis, Y. Matias, D. Webster, J. Barral, G. Corrado, C. Semturs, S. S. Mahdavi, J. Gottweis, A. Karthikesalingam, and V. Natarajan. Capabilities of gemini models in medicine, 2024.
[16] M. Wooldridge. *An Introduction to MultiAgent Systems*. John Wiley & Sons, Chichester, UK, 2nd edition, 2009.
[17] F. Nachbar, W. Stolz, T. Merkle, A. Cognetta, T. Vogt, M. Landthaler, P. Bilek, O. Braun-Falco, and G. Plewig. The abcd rule of dermatoscopy: High prospective value in the diagnosis of doubtful melanocytic skin lesions. *Journal of the American Academy of Dermatology*, vol. 30, n. 4, pp. 551–559, 1994.
[18] M. Messadi, A. Cherifi, and A. Bessaid. Segmentation and abcd rule extraction for skin tumors classification. *arXiv preprint arXiv:2106.04372*, 2021.
[19] M. Jütte, P. Schäfer, T. Drexler, N. Krüger, and B. Schilling. Integrating generative ai with abcde rule analysis for enhanced skin cancer diagnosis. *Frontiers in Medicine*, vol. 11, pp. 1445318, 2024.
[20] A. Siddique, H. Alshahrani, M. A. Khan, M. Raza, and T. Iqbal. An intelligent mechanism to detect multi-factor skin cancer. *Computational Intelligence and Neuroscience*, vol. 2024, pp. 1–12, 2024.