Universidade Estadual de Campinas

School of Electrical and Computer Engineering (FEEC)

Seminars in Computer Engineering

IA382A - 2025S1

AI-Driven Medical Image Diagnosis: Comparing Code Generation and Synthetic Data Augmentation

Final Class Project

Name: Guilherme Vilas Boas Ferreira da Silva

RA: 298305

July 2, 2025

Contents

Abstract

This study explores the integration of multiple AI tools in a medical imaging pipeline focused on classifying dermatoscopic images as either benign (nevus) or malignant (melanoma). Using a filtered subset of the HAM10000 dataset, two large language models (Claude and DeepSeek) were tasked with generating Python scripts for model training, while synthetic images were produced via Dreamina to support data augmentation and robustness evaluation. Quillbot was used to refine scientific writing throughout the project. Experimental findings reveal that AI-generated code can lead to biased or ineffective classifiers without human oversight. Synthetic images moderately improved model generalization, though limitations in visual fidelity persist. Overall, the results emphasize both the potential and the challenges of employing AI collaboratively across the stages of medical image analysis.

1 Introduction

Skin cancer remains one of the most prevalent forms of cancer worldwide, with melanoma representing its most aggressive and fatal variant. Early detection is essential for effective treatment and significantly enhances clinical outcomes. Dermatoscopic imaging serves as a primary diagnostic tool for pigmented skin lesions; however, variability in expert evaluations, rising case volumes, and limited access to specialized care in certain regions underscore the need for scalable, automated diagnostic solutions.

This project draws upon the seminar "Revolutionizing Medical Imaging Diagnostics with AI," which highlighted the potential of artificial intelligence (AI) to accelerate image interpretation through techniques such as classification, segmentation, and reconstruction, ultimately improving the speed and accuracy of clinical decision-making. Motivated by this perspective, the present study investigates the use of multiple AI tools within an integrated workflow to develop and assess a skin cancer classification model utilizing both real and synthetically generated medical images.

The HAM10000 dataset, containing over 10,000 dermatoscopic images spanning seven diagnostic categories [1], serves as the basis for this work. To simplify the task and concentrate on the critical clinical challenge of melanoma detection, the dataset is filtered to support a binary classification problem: distinguishing between benign lesions (nevus, class nv) and malignant lesions (melanoma, class mel). This binary setup is widely adopted in the literature due to its clinical importance and clearer diagnostic boundaries [2].

Four AI tools are employed, each fulfilling a distinct role within the workflow:

. Claude and Deepseek R1: two advanced code-generation models independently used to generate Python scripts for training image classification models on the selected dataset.

. Dreamina: an AI-based image generation platform utilized to create synthetic dermatoscopic images from textual prompts, facilitating data augmentation and generalization analysis.

. Quillbot: an AI-driven scientific writing assistant applied to enhance clarity, coherence, and structure in the report's textual content.

The primary objective of this study is to examine how AI can contribute across multiple phases of an applied medical imaging pipeline—namely, model development, data expansion, and documentation. The project also enables a comparative evaluation of the output quality, effectiveness, and practical applicability of various AI tools in health-related computational tasks.

2 Methodology

The project was conducted in well-defined stages, from initial data preparation to final inference with synthetic images. The first stage consisted of downloading and organizing the HAM10000 dataset, which originally contained images distributed in two separate folders. The images were grouped into a single directory structure, accompanied by the metadata file. Since the project's objective was to perform a binary classification between melanoma (mel) and nevus (nv), only these two classes were filtered. In addition, balance between categories was ensured, removing the bias of the majority class.

With the data organized, a prompt was created for two AIs (DeepSeek and Claude) to independently generate complete scripts for training a classification model using TensorFlow. The prompt used was the following:

User Prompt

Write a complete Python script using TensorFlow and Keras to train a binary image classifier using the HAM10000 dataset. The goal is to distinguish between the classes 'mel' (melanoma) and 'nv' (nevus). The dataset should be filtered to only include these two classes, and the model should be trained on balanced data from both categories. You may decide the image resolution, preprocessing steps, model architecture, and training strategy. The code should include data loading, training, evaluation, and comments explaining each step.

The goal of leaving some elements of the prompt open was to allow each AI to make autonomous decisions about the architecture, preprocessing, and training strategy, which favored the comparative analysis of the creativity and performance of each tool.

DeepSeek's response brought a model based on EfficientNetB0 with images resized to 224×224, using ImageDataGenerator for data augmentation and normalization. The dataset was split into 70 % training, 15 % validation, and 15 % testing. The network was trained with callbacks for early stopping, checkpointing, and learning rate reduction.

Claude's answer followed a similar structure, also using EfficientNetB0, but with some differences: a two-phase training strategy was adopted (freezing and fine-tuning), BatchNormalization and Dropout layers were added, and the model evaluation included additional metrics such as precision, recall, and confusion matrix. The data was split into 70 % training, 20 % validation, and 10 % testing.

During the analysis, a critical error was identified in both approaches: although EfficientNetB0 already includes a Rescaling layer for data in the [0-255] range, both AIs applied normalization to the images, which compromised performance. This problem was fixed directly in the model suggested by DeepSeek, through the following tuning prompt:

User Prompt

I want to point out something, actually to point out an error. In augmentation you are rescaling the training and test/val data, but for the backbone used, the input is expected to be between the range 0-255. I will copy what it says on the tf keras applications EfficientNetB0 page on the tensorflow website: "Note: each Keras Application expects a specific kind of input preprocessing. For EfficientNet, input preprocessing is included as part of the model (as a Rescaling layer), and thus keras applications efficientnet preprocessinput is actually a pass-through function. EfficientNet models expect their inputs to be float tensors of pixels with values in the [0-255] range.

After this correction, there was a significant improvement in the model's performance.

The next step involved generating synthetic images for both classes. Prompts on the Dreamina platform with detailed clinical descriptions were used to simulate realistic images of benign nevus and melanoma. Due to limitations in generation due to tokens and watermarks on the images, only 25 images were generated per class. The prompts used were:

Nevus:

User Prompt

Dermatoscopic image of a benign skin lesion (nevus) on human skin, photographed under clinical conditions. The mole appears flat or slightly raised, brown in color, with regular borders and uniform pigmentation. The lesion is shown on natural skin, with visible surrounding skin texture. Realistic lighting, medical photo style.

Melanoma:

User Prompt

Dermatoscopic image of a malignant melanoma on real human skin. The lesion is asymmetric, with irregular and blurred borders, and includes different shades of brown and black. The surrounding skin is visible, showing natural texture and tone. Clinical image, under medical lighting, no artificial background.

Below is an example of each generated image:

Given the small number of images, it was decided to use them exclusively in the

Figure 1: Benign Lesion (nevus)

Figure 2: Malignant Lesion (melanoma)

inference stage, instead of retraining the models. To organize the downloaded files, a script was requested that would rename the images using cv2, with standardized names such as nv01.jpeg or mel01.jpeg, depending on the source folder. The prompt for this request was:

User Prompt

Great, I've already downloaded the images, there are 25 for the nevus class and 25 for melanoma. The images are already in the folder for each class. The problem is that their names are messed up, currently they are the same as the prompt plus .jpeg. I want them to be organized as follows: for nevus/ nv01, nv02, nv03. And for melanoma: mel/ mel01, mel02. How can I do this using cv2?

Finally, the AI was asked to generate a script to perform inference with these

images, using the saved model melanomaClassifierFixed.h5. The model accepts RGB images of size 224x224 and returns a probability (sigmoid). The script was requested with the following requirements:

Reading the images with OpenCV; Resizing to 224x224 (without normalization); Prediction of the class based on the threshold 0.5; Printing the file name, predicted class and score; Generation of a confusion matrix.

The development and testing were successfully completed, and the results obtained with the synthetic images are analyzed in the next section.

To compose the introduction and conclusion of this report, the Quillbot language model was used, with the objective of generating a cohesive, formal and technically adequate text. For this, a descriptive prompt was provided, containing a summary of the scope and objective of the work. The prompt used was the following:

User Prompt

You are a technical assistant and need to write the introduction, conclusion and abstract of a project report. The project's theme was the creation of a binary image classifier with TensorFlow, using the HAM10000 dataset to differentiate between melanoma and nevus. The focus is on exploring different architectures with the help of generative artificial intelligence, making comparisons between different approaches for training the model, evaluating performance, and later testing with synthetic images generated by AI. Write a technical paragraph that is clear, formal, and objective.

The introduction and conclusion generated were later reviewed and adapted according to the context of the work, maintaining clarity and scientific adequacy.

3 Results

This section presents the results obtained from the training and evaluation of the models generated by the AIs Claude and DeepSeek. Initially, both models were trained with the EfficientNetB0 architecture, but they presented unsatisfactory performance due to a common error: the normalization of the input images, despite the backbone already containing an internal Rescaling layer. This error negatively affected the generalization capacity of the models, as can be seen in Figures 3 to 6.

Figure 3 shows the confusion matrix of the model generated by Claude, which classified 100% of the samples as belonging to the nevus class, completely failing to identify melanomas. This behavior is corroborated by the accuracy and loss curves

presented in Figure 4, which highlight the low effectiveness of the training.

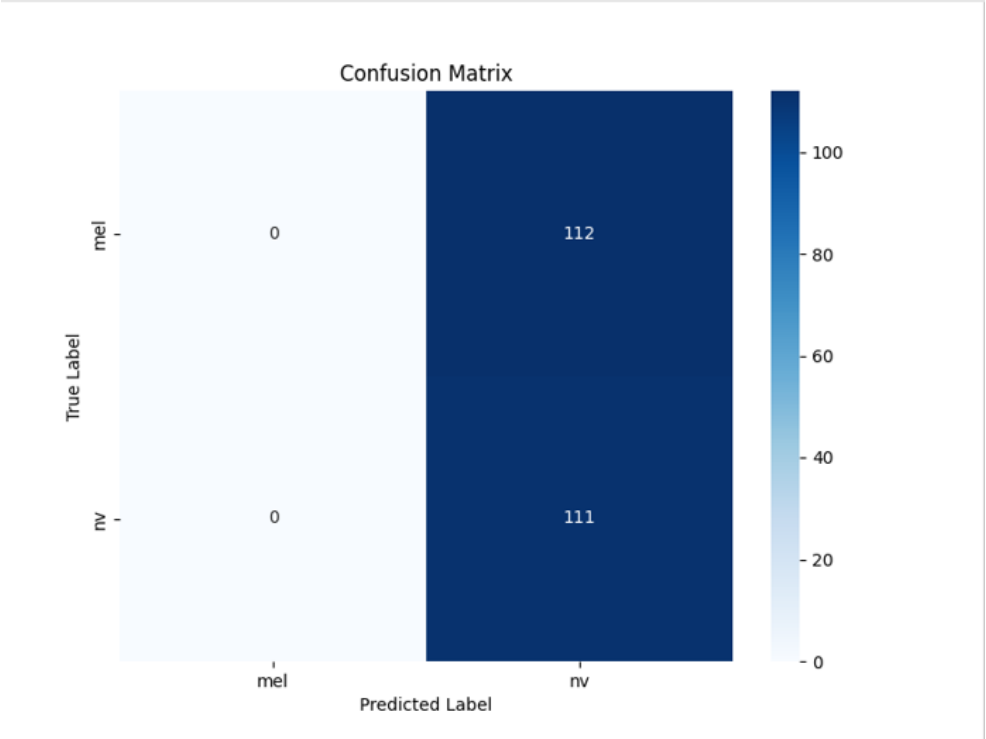

Figure 3: Claude Confusion Matrix

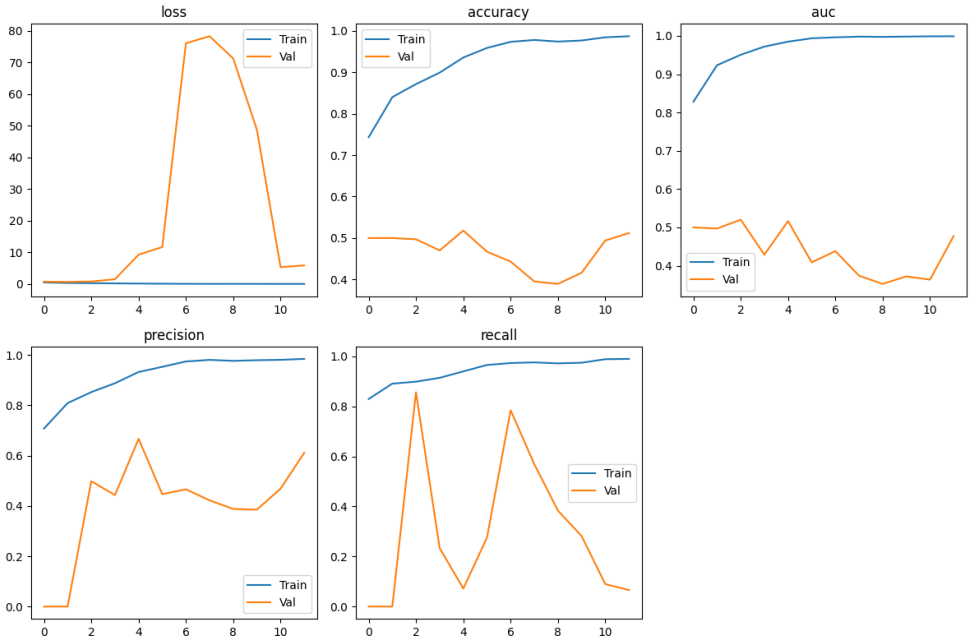

Figure 4: Claude Model Results

Similarly, 5 presents the confusion matrix of the model created by DeepSeek before

correction. In this case, the behavior was the opposite: all samples were classified as melanoma, also indicating a total failure in discriminating between classes. The training and validation curves in 6 reinforce this conclusion, showing fluctuations and a lack of significant convergence.

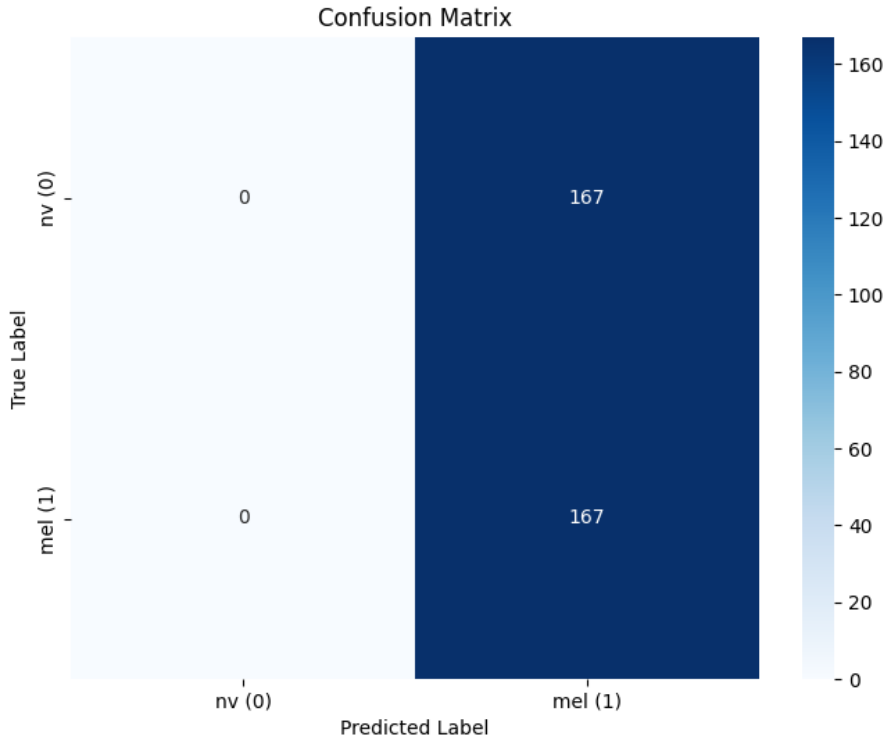

Figure 5: DeepSeek Confusion Matrix

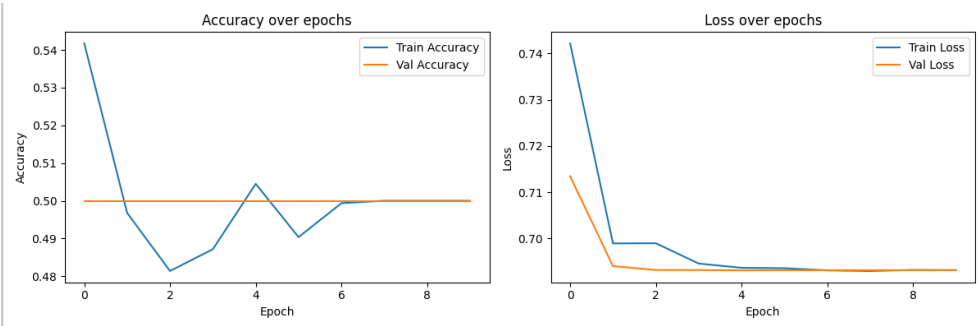

Figure 6: DeepSeek Model Results

Given these results, the DeepSeek model was adjusted to remove undue normalization. After this correction, a substantial improvement in performance was observed. The new confusion matrix is shown in 7, evidencing a more balanced separation between classes. The accuracy and loss curves in 8 show stable behavior, with progres-

sive improvement in accuracy and reduction in loss in both training and validation, indicating a more reliable and generalizable model.

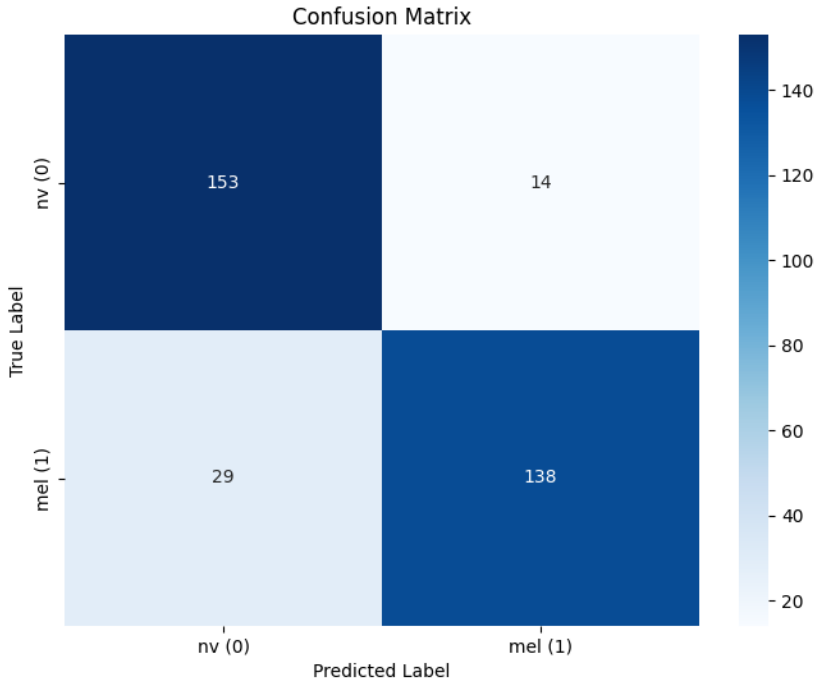

Figure 7: Fixed Model Confusion Matrix

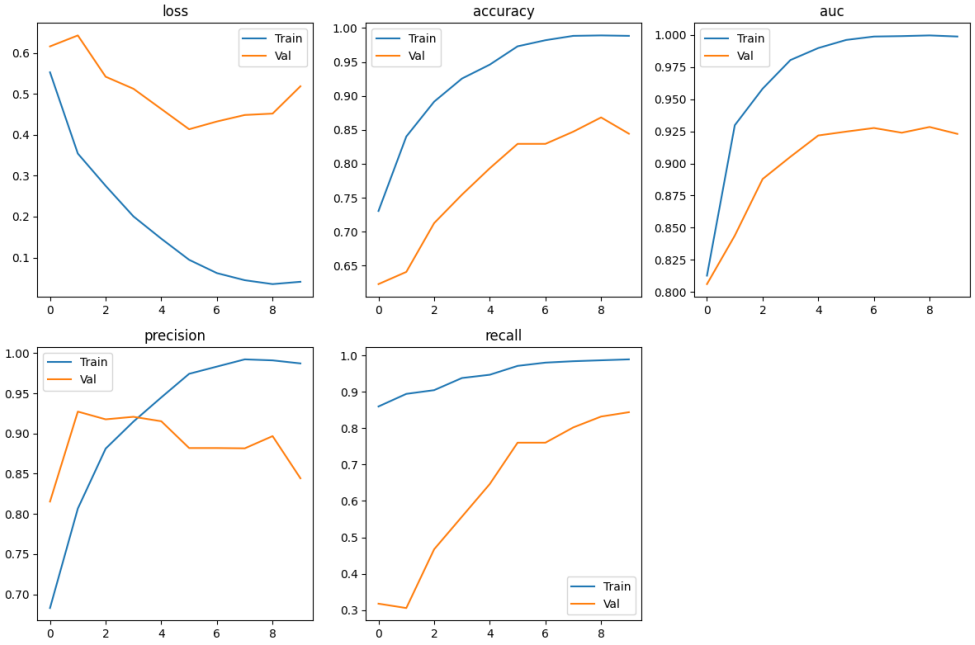

Figure 8: Fixed Model Model Results

Finally, 9 presents the confusion matrix obtained by applying the adjusted model

to the synthetic images generated by AI. Even with the limited number of samples (25 per class), the model was able to achieve an accuracy of 60%, with a precision of 0.63 for nevus and 0.58 for melanoma, and recall of 0.48 and 0.72, respectively. These results indicate that the model tended to overidentify malignant lesions, which resulted in higher sensitivity (recall) for melanoma, at the cost of a higher number of false positives for this class. Although the performance was not ideal, the test with artificial data served as a preliminary assessment of the model's generalization ability, especially useful given the limited availability of real images.

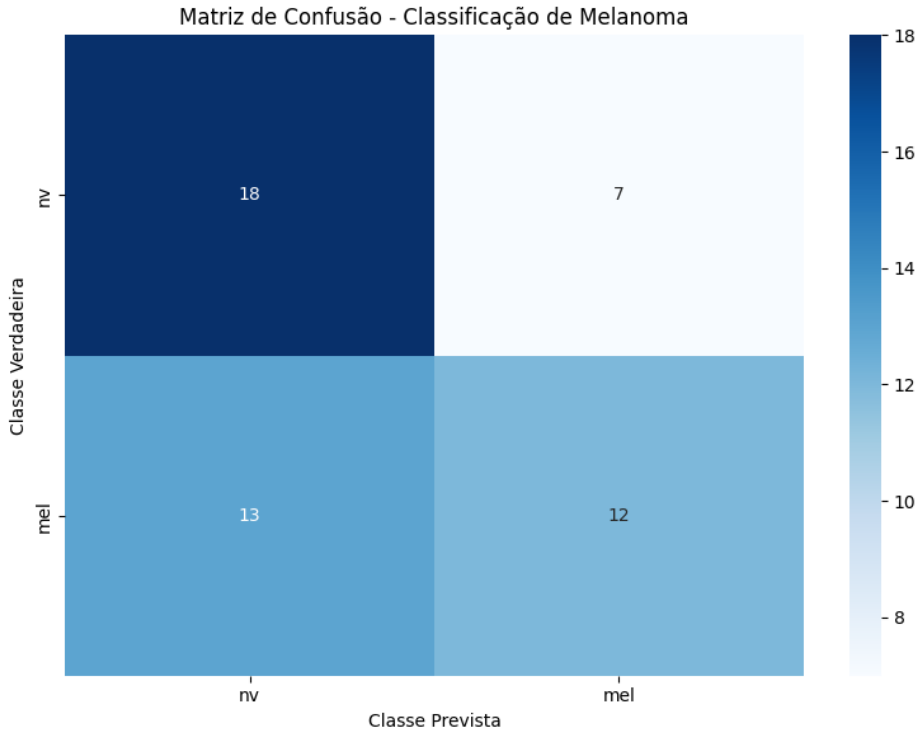

Figure 9: Generated Images Inference Confusion Matrix

4 Conclusions

This work explored the combined use of artificial intelligence tools to build and evaluate a binary classifier for dermoscopic images, focusing on the distinction between benign (nevus) and malignant (melanoma) lesions. Two code generators (Claude and DeepSeek) were used, which received the same instruction prompt to ensure a fair comparison regarding the construction of the training pipeline.

In practice, the model initially generated by Claude presented biased behavior, classifying all images as benign, while the model generated by DeepSeek presented the

opposite bias, classifying all as malignant. After careful analysis, a common flaw was identified: both normalized the input images, contrary to the official recommendation of the EfficientNetB0 architecture used. This error compromised the performance of both models. By correcting it in the DeepSeek model, it was possible to achieve significantly better results, with good separation between classes and consistent metrics during training and validation.

This episode highlighted a crucial point: in projects involving the use of AI tools, it is essential to have prior knowledge about the concepts involved — such as neural network architectures and specific preprocessing requirements —, since leaving all decisions exclusively to AIs can generate subtle but critical flaws in the final performance of the system. In this regard, both tools proved to be flawed. Despite this, both showed excellent quality in generating quality code in an end-to-end machine learning project, saving a lot of software development time.

Then, the model's performance was evaluated with a set of synthetic images generated by the Dreamina AI. Twenty-five images per class (melanoma and nevus) were obtained, manually reviewed and renamed to allow inference. Although the number of images was limited, the results showed that the model was able to partially generalize, with an accuracy of 60%, recall of 72% for melanoma and 48% for nevus. This indicates that, even with reduced volume, artificial images provided a useful additional test of the model's robustness and generalizability, albeit with limitations.

The combined use of AI tools—both for generating models and for creating data—has shown promise, accelerating development and enabling multiple iterations quickly. However, important limitations need to be considered: the structural bias of code-generating AIs, the imperfect visual fidelity of synthetic images, and the lack of clinical validation are factors that affect applicability in real-world scenarios. Additionally, the use of Quillbot proved useful for structuring the text in an objective way, ensuring fluidity and good technical writing.

As future suggestions, it is recommended to expand the model's scope to multiple classes (e.g., BCC, AKIEC, etc.), use larger datasets validated by experts, explore alternative architectures (such as Vision Transformers or EfficientNetV2), and test models with synthetic data on a larger scale, evaluating to what extent they can replace or complement real data in the training process.

The complete code, models, and synthetic dataset are available at: `https://github.com/Gui7621/skin-lesion-classification-ai/tree/main`.

5 References

[1] R. C. . K. H. Tschandl, P., "A large collection of multi-source dermatoscopic images of common pigmented skin lesions," in *Scientific Data*, vol. 5, no. 1, 2018.

[2] N. C. F. Codella, D. Gutman, M. E. Celebi, B. Helba, M. A. Marchetti, S. W. Dusza, A. Kalloo, K. Liopyris, N. Mishra, H. Kittler, and A. Halpern, "Skin lesion analysis toward melanoma detection: A challenge at the 2017 international symposium on biomedical imaging (isbi), hosted by the international skin imaging collaboration (isic)," in *2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018)*, 2018, pp. 168–172.