May 16th, 2025

CNN-SVM Comparison for Skin Cancer Detection

Final paper

Machine Learning and Deep Learning

Program: MSc in Business Administration and Data Science

Authors: Michele Daconto (176193), Alessio Desideri (176184)

Examiner: Somnath Mazumdar

Number of characters: 19,068

Number of pages: 11

Link to Script: here

Academic Year 2024/2025

Table of Contents

Abstract

The aim of this research is to design, train, and compare machine learning models capable of automatically diagnosing melanoma from dermoscopic images. Through an extensive ensembling process, we developed an end-to-end model named SkinNET80, based on a convolutional neural network (CNN). We then trained and compared an end-to-end support vector machine (SVM) and a hybrid model which combines CNN features with an SVM classifier. Among the tested approaches, SkinNET80 achieved the best trade-off between accuracy (77%) and runtime (466 sec.). SVMs achieved best metrics but required more than double the time. The primary framework used for development was TensorFlow Keras. The models were trained on over twenty thousand padded and resized images from the ISIC archive. The results obtained are fully comparable to those reported in similar studies in the field.

Keywords: Melanoma – CNN – SVM – Computer-aided-diagnosis (CAD) – Image Recognition – PCA – ISIC

1. Introduction

Melanoma is a malignant tumor of the melanin-producing cells. Although it accounts for approximately 5% of all skin cancers, it is the most aggressive form and is responsible for the majority of skin cancer-related deaths (American Cancer Society, 2024).

Until the 1980s and 1990s, melanoma diagnosis relied solely on clinical examination with the naked eye and, following surgical excision, on histological analysis under a microscope. Physicians used criteria such as the ABCDE[1] rule to differentiate suspicious lesions from benign moles. Dermoscopy, employing a magnifying lens (x20) and polarized light, only became widespread toward the end of the 20th century, while the integration of digital imaging systems and AI algorithms is a much more recent development (Esteva et al., 2017).

[1] The ABCDE rule involves examining each mole by assessing the following criteria: Asymmetry**,** Border irregularity**,** Color variation**,** Diameter (greater than 6 mm), and Evolution (changes over time).

1.1 Research Question

Using a dataset of approximately 27,000 dermoscopic images from the ISIC[2] archive, this research aims to develop, train, and compare various machine-learning models capable of accurately and confidently determining whether a skin lesion configures a melanoma or not.

1.2 Motivation

From the beginning, we were motivated to undertake a project in the field of medical diagnostics. This dermatological study drew our attention for a clear reason: melanoma becomes significantly more lethal once it metastasizes to other organs, making early and accurate diagnosis essential. The five-year survival rate is 99.6% when melanoma is still in its early stage, but drops sharply to 35.1% if diagnosed later (American Cancer Society, 2024).

Since the 1980s, the false negative rate has decreased from 18.5% to 7% (PubMed Central, 2024). We believe this rate can be reduced even further (Haenssle et al., 2020).

1.3 Related Work

In the academic domain, the majority of the domain-specific papers identified were published in conjunction with ISBI 2016[3]. The majority of these studies demonstrated that machine learning models – particularly convolutional neural networks (CNNs), which are often pre-trained, and support vector machines (SVMs) – achieve diagnostic accuracy comparable to, and in many cases exceeding, that of a benchmark group of eight expert dermatologists, who achieved a correct diagnosis rate of 70.5% (Esteva et al., 2017).

---

[2] The ISIC (International Skin Imaging Collaboration) archive is an open-access platform that annually collects thousands of annotated dermoscopic images of skin lesions to support research and the development of AI algorithms.

[3] The ISBI (International Symposium on Biomedical Imaging) Challenge 2016 evaluated algorithms for lesion segmentation, dermoscopic pattern recognition, and the benign/malignant classification of skin tumors such as melanoma. Organized by the ISIC (International Skin Imaging Collaboration), the challenge attracted participation from 38 teams, resulting in 79 submissions.

Codella et al. (2018) evaluated the performance of purely deep learning-based models, achieving diagnostic accuracy higher than that of dermatologists, but not exceeding 80%. In contrast, Kawahara et al. (2016) adopted a hybrid approach, combining features extracted from convolutional neural networks (CNNs) with support vector machine (SVM) classifiers, reaching an accuracy of approximately 90%, which outperformed that of the 'pure' models.

1.4 Conceptual Framework

Our research was carried out in four straightforward steps which brought us to three models (see Fig. 1).

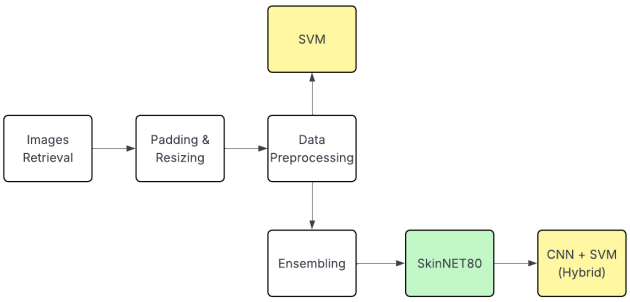

Figure 1 – Conceptual Framework

Following the preparation of the data and images for analysis through key operations such as undersampling, padding, and resizing, the subsequent focus shifted to the development of the models.

The initial efforts focused on the design and optimization of a convolutional neural network (CNN). The final model, SkinNET80, is the result of an extensive *ensembling* process. In this sense, two additional networks are also documented in the accompanying notebook (see script).

In the subsequent phase of the study, the performance of a SVM was evaluated. This model takes vectors (representing images) as inputs and, as each vector has 150,528 features we performed PCA (*Principal Component Analysis*) to accelerate the training.

Following the approach of Kawahara et al., we then assessed a hybrid model combining CNN features with an SVM classifier, training the latter on 256 parameters from the *dense* layer of the network.

2. Methodology

2.1 Dataset Description and Preprocessing

As previously mentioned, the images analyzed in our study were sourced from the ISIC archive as updated on May 11, 2025. Rather than relying on a precompiled dataset, we constructed our own from scratch. The ISIC archive website allows users to browse the full image collection, apply filters and download images along with their associated metadata. By adjusting the filtering criteria, we downloaded a total of 28,905 dermoscopic images (.JPEG), of which 17,940 were labelled as representing melanoma ('malignant') and 10,965 were labelled as benign nevi or other non-threatening skin lesions ('benign').

Once the images had been obtained, we standardized their dimensions to 224 x 224 pixels as they initially varied in size. However, we applied *padding* before resizing to avoid unwanted distortions. This technique involves transforming each image into a square by adding black borders where necessary.

The metadata included 31 columns relating to both the lesion and the patient, covering attributes such as approximate age and the anatomical site of the lesion. However, as the majority of these values were missing, we decided to concentrate solely on the images and their associated labels.

In this phase, we balanced the two classes by reducing the number of images labelled as 'malignant' from 17,940 to 10,965 using a procedure known as *undersampling* (see Fig. 2).

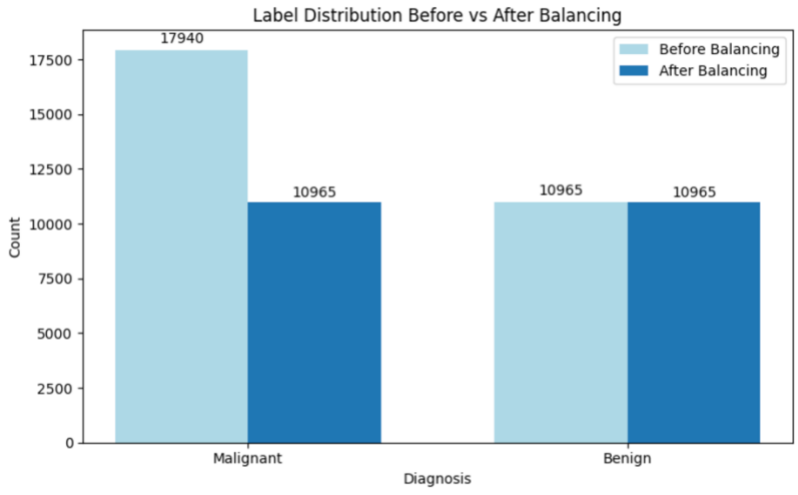

Figure 2 – Label Distribution Before vs. After Balancing

Before proceeding to model training, we defined the image *augmentation* criteria. During training, a series of random transformations are applied to the images in the training set to improve the model's robustness and generalisation capabilities. Specifically, we applied mirroring and random rotation (±60°), as well as random zooming (±10%) and contrast adjustment (factor 0.1). See Fig. A in the Appendix for a comparison between original and augmented images.

No augmentation is applied to the images in the validation set (15%) or the test set (15%), as our goal is to evaluate model performance on the original, unaltered images.

2.2 Journey to SkinNet80

To identify the optimal CNN model capable of achieving the best performance metrics, we employed an ensembling[4] approach. To provide a clear narrative of how we arrived at our final network, SkinNET80, we present two intermediate models as part of the development process.

2.2.1 Basic CNN

The first model is a basic CNN consisting of two convolutional layers, leading to a binary output through a *sigmoid* activation function. This network - like all subsequent models described - uses the *Adam* optimizer with a *learning rate* of 1e-3, aiming to minimize *loss* and maximize *accuracy*. With 20 epochs of training, it is evident that the model begins to severely overfit within the very first epochs (see Fig. B in the Appendix).

2.2.2 CNN - Dropout (0.3) & Class Weights

In the second network, we added two additional convolutional layers, for a total of four. To mitigate overfitting, we introduced a *Dropout* layer (rate 0.3) just before the output layer. This command randomly deactivates 30% of the neurons during training and helps the model generalize better. Furthermore, to reduce the number of false negatives - which are significantly more critical than false

[4] Technique that combines the characteristics of many different models.

positives - we incorporated the *class_weights* parameter to assign a higher penalty to misclassifications of malignant cases.

In this model, we also introduce *callbacks* - mechanisms that allow the network to adjust itself dynamically during training. Specifically, *EarlyStopping* halts training if the validation loss does not improve for 5 consecutive epochs, while *ReduceLROnPlateau* halves the learning rate if the validation loss fails to improve for 3 consecutive epochs.

However, this model also began to overfit after approximately 8 epochs, leading to decreased reliability in its performance (see Fig. C in the Appendix).

2.2.3 SkinNET80

After making some refinements to the previous architecture, we finally arrived at our optimized CNN: SkinNET80. To enhance the model's generalization capabilities, we increased the dropout rate from 30% to 50% and incorporated a *BatchNormalization*[5] step following each convolutional layer.

Although the validation accuracy of this model is slightly lower than that of the previous model (77% vs. 82.5%), we prefer it as it does not exhibit signs of overfitting or underfitting (see Fig. C in the Appendix), and it handles both false negatives and false positives in an acceptable way. We used a *confusion matrix* and a ROC-AUC curve to monitor these types of errors (see Fig. D in the Appendix).

2.3 SVM end-to-end

Having evaluated the performance of a CNN on our image dataset, we proceeded to build an SVM model to assess potential differences.

As an SVM requires vector inputs, we flattened each image into a vector of $224 \times 224 \times 3 = 105{,}528$ features. To accelerate the training process, we applied *Principal Component Analysis* (PCA) to retain

[5] Batch Normalization normalizes the activations of a layer within each mini batch to have zero mean and unit variance, stabilizing and accelerating training.

99% of the variance and eliminate redundant dimensions. We then standardized the features and performed grid search cross-validation[6] to identify the optimal values for the C and γ[7] parameters.

The SVM was subsequently trained and tested, yielding satisfactory results (see Fig. F in the Appendix).

2.4 CNN + SVM – Hybrid Approach

In this combined model, we leverage the feature extraction capabilities of SkinNET80 through its three convolutional layers and then use an SVM classifier for binary classification. Specifically, the input to the SVM consists of the activations from the *Dense* layer - located just before the output and Dropout layers.

If the SVM classifier proves more effective than the sigmoid function used in the original CNN, then this hybrid approach should yield improved performance metrics compared to SkinNET80. This was indeed the case (see Fig. G in the Appendix).

3. Results

3.1 Key metrics for analyzing the models

We chose to focus our model evaluation on the *recall* metric for the "Malignant" class and on the F1-score specific to *malignant* because these measures more directly reflect the clinical requirements of a diagnostic support tool. Recall, also known as sensitivity, quantifies the model's ability to capture true positives, that is, how many among the patients affected by melanoma are correctly identified. This is the most critical parameter when the goal is to minimize false negatives, meaning cases in which cancer

[6] Grid cross-validation is a procedure where you systematically evaluate all combinations of specified hyperparameters ("grid") across multiple train-validation splits (folds), computing average performance to identify the best configuration.

[7] The C parameter controls the trade-off between a wide margin and classification errors: higher values penalize misclassifications more heavily but may lead to overfitting. The γ (gamma) parameter in the RBF kernel determines each training point's reach: larger γ yields more complex, localized decision boundaries, while smaller γ produces smoother, more global boundaries.

escapes automated detection. In a medical context, the impact of a missed diagnosis can be extremely severe, with consequences for the timeliness of treatment and, consequently, for the patient's prognosis.

$$recall = \frac{TP}{TP + FN}$$

On the other hand, the *F1-score* for the melanoma class provides a single measure that combines recall and precision. Precision measures the proportion of positive predictions that are actually correct, helping to keep the number of false alarms under control and avoiding an excess of unnecessary biopsies or unfounded anxiety in patients.

$$F1 \frac{2 \cdot TP}{(2 \cdot TP) + FP + FN}$$

By integrating recall and precision into one harmonic value, the F1-score becomes particularly well suited to a context in which it is not enough merely to maximize tumor detection; it is also essential to maintain a reasonable balance between sensitivity and specificity, ensuring responsible use of clinical resources and reducing psychological burden.

3.2 A comparison between SkinNet80 and SVM models

As outlined above, we developed and evaluated three distinct approaches: (1) SkinNet80, a custom three-layered CNN trained from scratch; (2) two different SVM classifiers, one of them built on features extracted from SkinNet80. When we compare their performance metrics alongside computational cost, the SVM stands out as the most accurate, with the highest F1-scores and recall, while SkinNet80 emerges as the best trade-off between flexibility, balanced class performance, and modest training time and resource requirements.

We need to start by analyzing the difference between the final version of SkinNet80 and its previous version shown in the journey. Only by looking at the metrics we could state that the final version underperformed, but we decided to proceed with that one because the model without batch normalization tends to overfit through time. Even if the final model has lower scores in F1 and recall, we avoided overfitting.

Moving on to the comparison with the SVM models we built, we can examine both versions since they both deliver very positive results. When we look at F1 and recall, the two SVM structures perform almost identically. Fig. 3 below shows the metrics from the SVM built on features extracted from SkinNet80.

Test set classification report:

	precision	recall	f1-score	support
Benign	0.79	0.84	0.81	1645
Malignant	0.83	0.78	0.80	1645
accuracy			0.81	3290
macro avg	0.81	0.81	0.81	3290
weighted avg	0.81	0.81	0.81	3290

Figure 3 – Classification Report SVM (Hybrid)

The other SVM shows similar values. In detail, it achieves balanced performance across both classes, with an overall accuracy of 82 %. For benign lesions, it shows a precision of 0.81 and a recall of 0.84, meaning most benign cases are correctly identified with relatively few false positives. For malignant lesions, precision is 0.83 and recall is 0.81, indicating strong detection of melanomas while keeping false alarms low. The nearly identical F1-scores (0.83 vs. 0.82) and matching macro- and weighted-averages confirm that the classifier generalizes evenly without favoring one class over the other.

Test set classification report:

	precision	recall	f1-score	support
Benign	0.81	0.84	0.83	1645
Malignant	0.83	0.81	0.82	1645
accuracy			0.82	3290
macro avg	0.82	0.82	0.82	3290
weighted avg	0.82	0.82	0.82	3290

Figure 4 – Classification Report SVM (end-to-end)

SkinNet80 achieves high benign recall (0.91) but lower precision (0.70), while for melanomas it reaches precision 0.87 but recall only 0.62, meaning it misses roughly 40 % of true cancers (see Fig. E in Appendix). Its overall accuracy and F1-score of 0.76 reflect this imbalance, pointing to a need for higher melanoma sensitivity, to improve clinical safety.

Classification Report:

	precision	recall	f1-score	support
Benignant	0.70	0.91	0.79	1645
Malignant	0.87	0.62	0.72	1645
accuracy			0.76	3290
macro avg	0.79	0.76	0.76	3290
weighted avg	0.79	0.76	0.76	3290

Figure 5 – Classification Report SkinNET80

3.3 Models trade-offs

In order to choose the best model, we need to understand the trade-offs. If we consider only the recall and F1-score metrics, end-to-end SVM performs better than SkinNet80. It successfully classifies most of the two classes and it is a balanced model. Although, we need to consider that the SVM model tends to be a little overfitting, while SkinNet80 does not have this kind of problem. Moreover, the computational power required by SVM model is higher than the one required by SkinNet80. In fact, the first one takes almost double the time to complete its run (see Fig. H in Appendix) and it requires a huge amount of RAM usage, which sometimes can lead to a stop in the running.

4. Conclusion

In conclusion, we developed three distinct models to detect melanoma with varying levels of efficiency. The choice of a primary model depends on context and dataset size: while the SVMs achieve excellent performances, they demands substantial computational resources and shows mild overfitting. By contrast, SkinNet80 is slightly less precise and balanced but avoids overfitting and runs quickly.

Throughout our evaluation, we prioritized recall and the melanoma-specific F1-score to reflect the clinical imperative of catching every possible tumor. Raising recall inevitably increases false positives, benign lesions flagged as suspicious, but sharply reduces the risk of missed melanomas, which can have fatal consequences.

This trade-off suits an oncological screening scenario, though it does lead to more follow-up visits and biopsies, with psychological and resource costs. Balancing precision and recall via the F1-score prevents unnecessary alarm while ensuring timely diagnosis and responsible use of medical resources. Finally, reporting overall accuracy alongside weighted and macro averages provides a fair summary of model capabilities, always with the reminder that these tools support, rather than replace, clinical judgment.

4.1 Limitations

The main limitation we encountered was computational power: although Colab's A100 GPU is very capable, image processing and model training still took a considerable amount of time. In addition, to keep the algorithm fluid and within memory constraints, we worked with a reduced number of images,

not using the whole archive (as opposed to the methodology employed in the ISBI 2016 challenge; consequently, our results are not directly comparable to those reported in that context). We also faced RAM limitations that, on several occasions, slowed down or even halted certain preprocessing and training steps.

4.2 Future Scenarios

For future work, we would like to collect and integrate histopathological annotations, such as Breslow thickness and melanoma subtype, to train a multi-task model that simultaneously predicts both the invasiveness grade and the specific melanoma type (e.g., *superficial spreading, nodular, lentigo maligna*). This will involve curating a richly labeled dataset and adapting our CNN architecture to output fine-grained, clinically actionable classifications.

References

[1] American Cancer Society, *Cancer Facts & Figures 2024*, Atlanta: American Cancer Society, 2024.

[2] Codella, N. C. F., Rotemberg, V., Tschandl, P., Celebi, M. E., Dusza, S. W., Gutman, D., ... & Halpern, A., *Skin lesion analysis toward melanoma detection: A challenge at the 2017 International Symposium on Biomedical Imaging (ISBI),* 2016 IEEE 13th International Symposium on Biomedical Imaging (ISBI), 2018.

[3] Esteva, A., Kuprel, B., Novoa, R. A., Ko, J., Swetter, S. M., Blau, H. M., & Thrun, S., *Dermatologist-level classification of skin cancer with deep neural networks,* Nature, 2017.

[4] Haenssle, H. A., Fink, C., Schneiderbauer, R., Toberer, F., Buhl, T., Blum, A., ... & Reader study level-I group, *Man against machine: diagnostic performance of a deep learning convolutional neural network for dermoscopic melanoma recognition in comparison to 58 dermatologists*, Annals of Oncology, 2020.

[5] Kawahara, J., BenTaieb, A., & Hamarneh, G., *Deep features to classify skin lesions,* 2016 IEEE 13th International Symposium on Biomedical Imaging (ISBI), 2016.

[6] PubMed Central, *Trends in melanoma detection accuracy: a 40-year overview*, PMC Research Highlights, 2024.

[7] Chollet , F., *Keras: The High-Level neural networks API of TensorFlow*, TensorFlow Core Team, 2024.

[8] Harangi, B., *Deep learning-based skin lesion segmentation and classification of melanoma using support vector machine (SVM),* ResearchGate, 2019.

[9] Ali, A., Khan, M. A., Anwar, S. M., & Tariq, U., *Skin lesion classification using hybrid convolutional neural network and support vector machine*, Applied Sciences, 2023.

Appendix

Fig. A – *Original vs. Augmented Images*

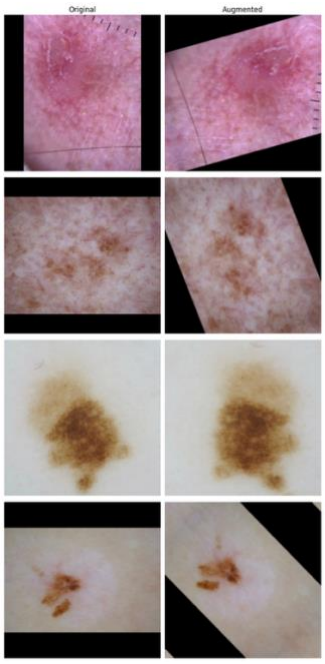

Fig. B – *Overfitting in Basic CNN*

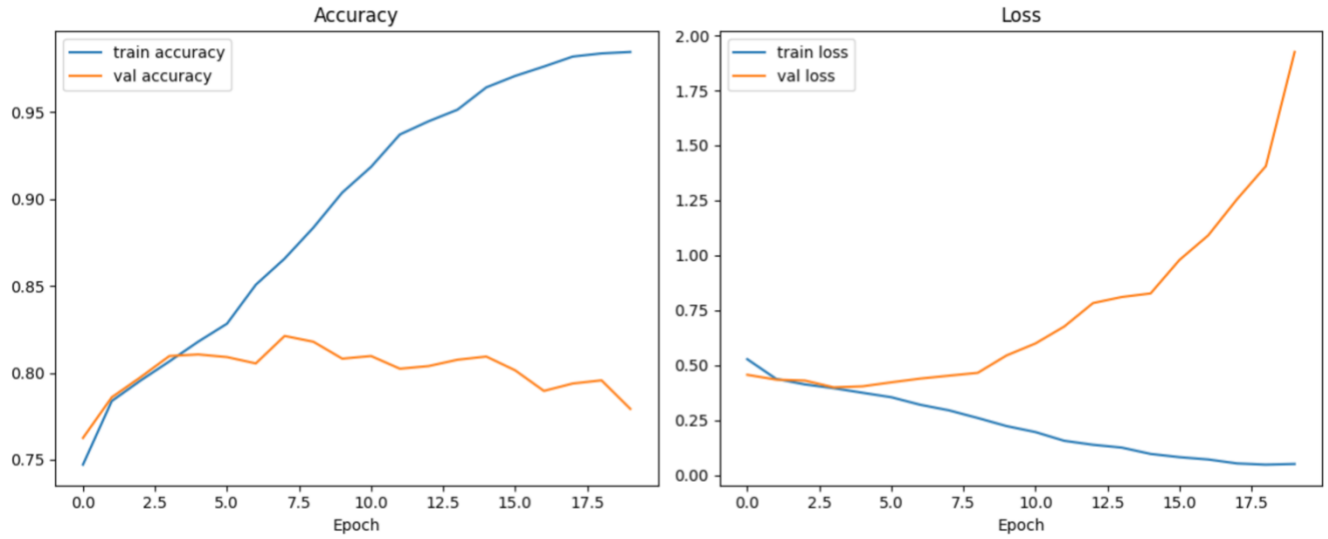

Fig. C – *Overfitting in CNN w/ Dropout & Class Weights*

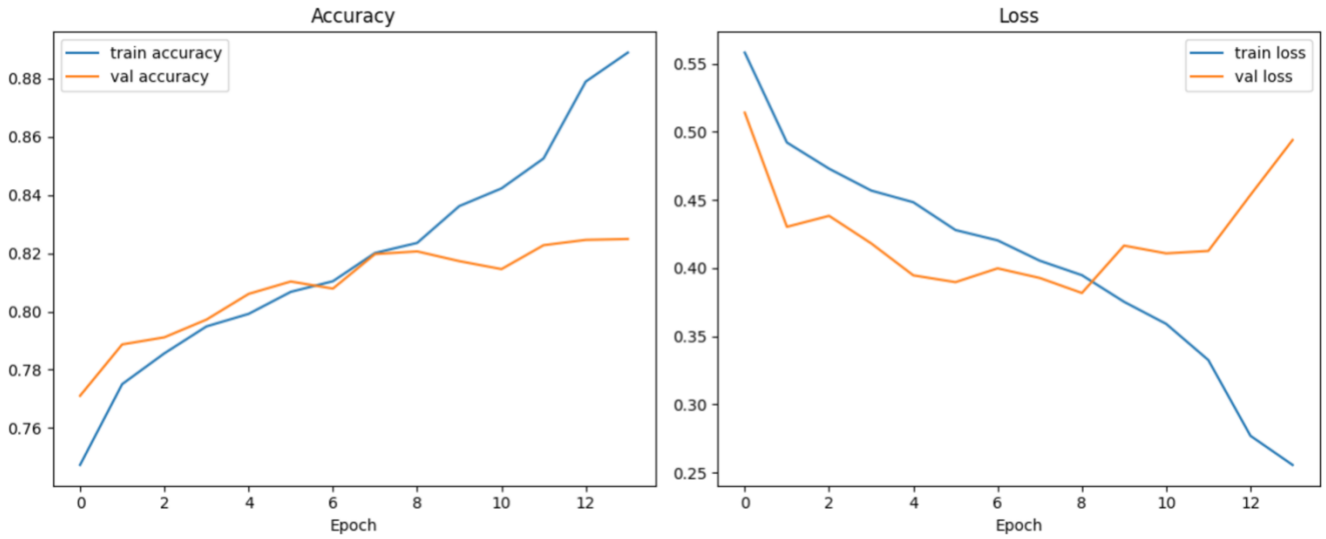

Fig. D – No overfitting in SkinNET80

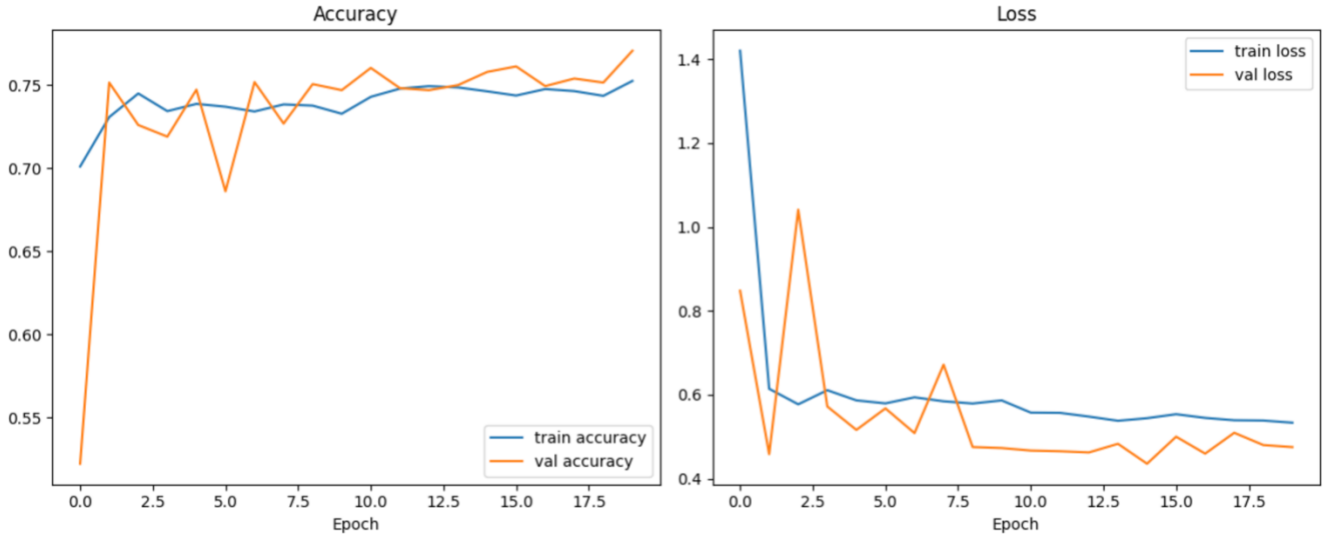

Fig. E – Confusion Matrix & ROC-AUC Curve of SkinNET80

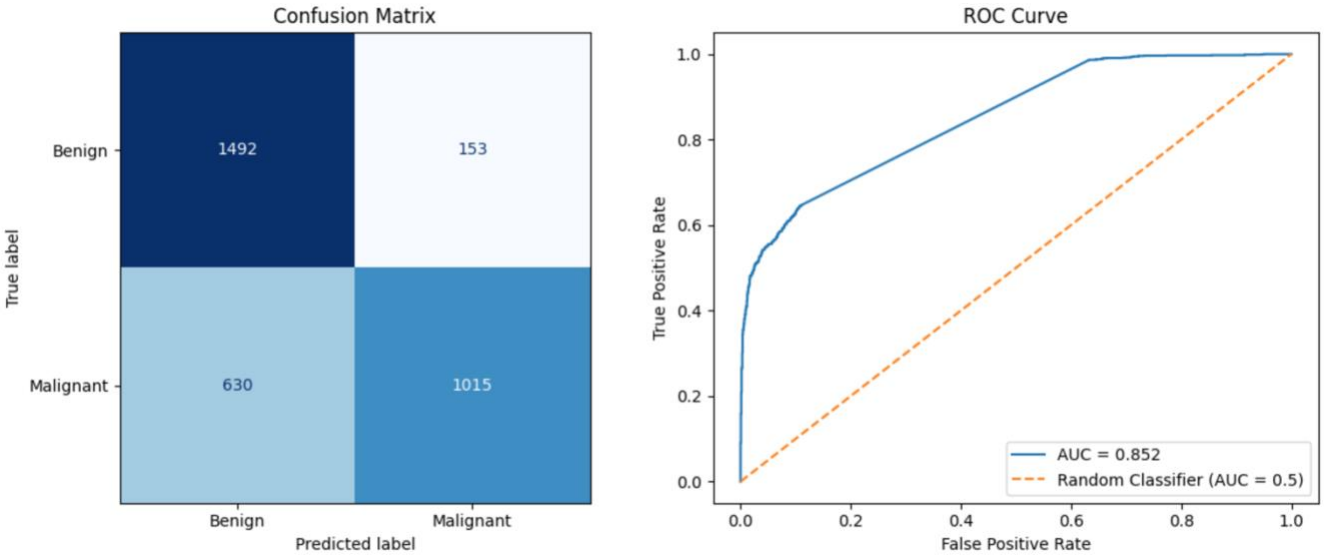

Fig. F – Confusion Matrix & ROC-AUC Curve of SVM

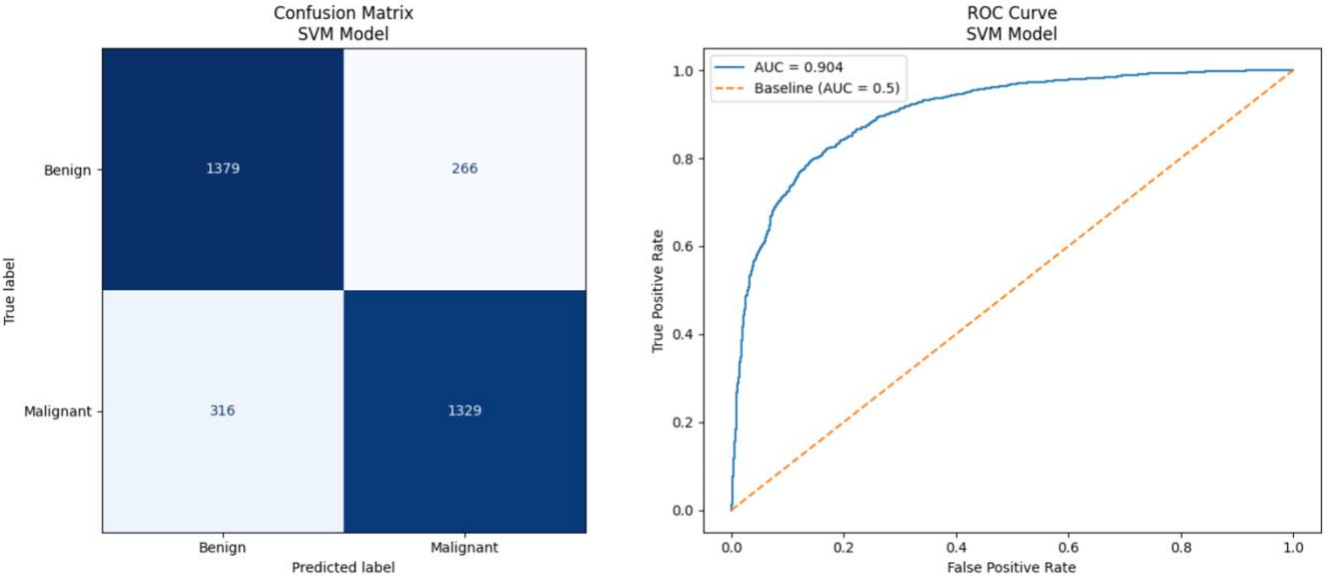

Fig. G – Confusion Matrix & ROC-AUC Curve of SVM (Hybrid Approach with CNN)

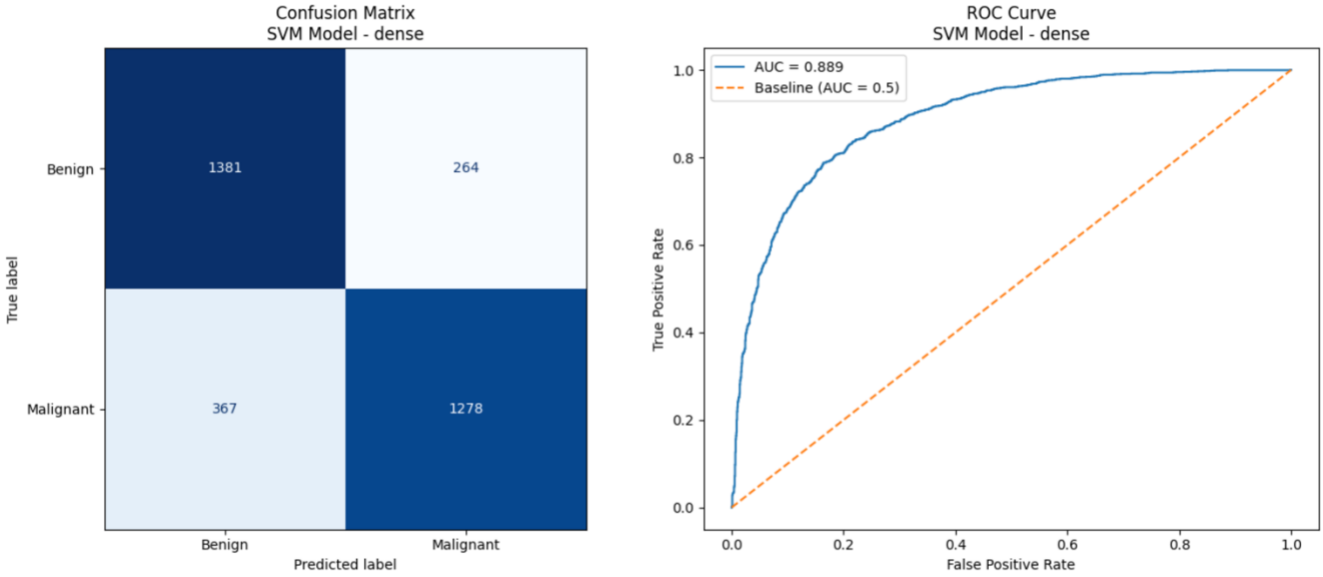

Fig. H – *Running Times*