Skin Classification: A Comparative Analysis of Machine Learning and Deep Learning Approaches

Xavi Beltran[1], Clara Lisazo[1], and Luisana Alvarez Monsalve[1]

[1]University of Cassino, Erasmus Mundus Joint Master Degree in Medical Imaging and Applications (MAIA)

June 13, 2023

Abstract

Skin cancer is a common worldwide disease with an increasing frequency, that demands the development of automated methods for early detection and classification of skin lesions. This project compares two approaches in the context of the ISIC2017 Challenge: utilizing deep learning in combination with machine learning classifiers and relying solely on deep learning models. The first method extracts high-level features using pre-trained networks, followed by classification with machine learning algorithms. The second strategy focuses solely on deep learning for both feature extraction and classification. By contrasting the two methods, their relative strengths and limitations in terms of precision, efficiency, and generalizability are determined. The results demonstrate that the deep learning technique outperforms machine learning, with a balanced accuracy of 0.72. However, further improvement is needed to use this information for real-life diagnostic and treatment applications.

Keywords - Skin classification, Machine Learning, Deep Learning, Transfer Learning, ISIC2017, DenseNet201, Skin cancer, Melanoma, Seborrheic, VGG16

1. Introduction

Skin cancer represents the most prevalent worldwide malignancy and its incidence shows no signs of plateauing [1]. This disease comprises the uncontrolled growth of abnormal skin cells and it is predominantly caused by exposure to ultraviolet (UV) radiation from the sun or tanning beds, though genetic and environmental factors can also play a role [2]. Due to its increasing incidence, skin cancer became a concern for global health institutions since the 1980s. The World Health Organization (WHO) estimates that skin cancer will be the most prevalent cancer in the world by 2030, with over 20 million new cases diagnosed each year [3].

The ISIC2017 Challenge [4] aligns with the worldwide concern of skin cancer by providing an opportunity for the development of automated systems that assist in the detection and classification of skin lesions. The challenge acknowledges the significance of early detection and accurate classification of skin cancer, as they have a direct impact on patient outcomes. By participating in the ISIC2017 Challenge, researchers contribute to the advancement of computer-assisted diagnosis for skin cancer, potentially revolutionising the field through the development of robust models that aid dermatologists in the early detection and management of skin lesions.

In this project, we aim to compare the performance of two distinct approaches in the context of the ISIC2017 Challenge: using deep learning (DL) as a feature extractor in conjunction with traditional machine learning (ML) algorithms for the classification, and relying solely on deep learning algorithms for both feature extraction and classification.

The first method extracts high-level features from skin lesion images using deep learning models, such as Densenet201. The extracted features are then incorporated into standard machine learning algorithms, such as support vector machines (SVMs) for its classification. This method combines the ability of deep learning models to extract features with the interpretability and efficacy of conventional machine learning algorithms.

The second strategy focuses solely on deep learning methodologies across the entire pipeline. This approach leverages the powerful learning capabilities of deep neural networks, allowing models to autonomously learn discriminative features from the raw images of skin lesions. By eliminating the need for explicit feature engineering and relying solely on deep learning algorithms, this strategy may capture more complex patterns and result in enhanced classification performance.

By contrasting the two methods, we hope to determine their relative strengths and limitations in terms of precision, efficiency, and generalizability. The comparison will cast light on the performance tradeoffs and the most effective strategy for skin lesion classification within the context of the ISIC2017 Challenge.

2. ISIC 2017 Dataset

In the fields of dermatology and computer vision, the ISIC 2017 dataset is a widely-used resource. It was designed to aid in the study of automated skin lesion analysis and melanoma detection. This dataset consists of a large collection of dermoscopic images of skin lesions captured from a variety of angles and illumination conditions.

The training dataset of the ISIC 2017 challenge, which contains a total of 2,000 images, was used to implement this project. Each image in the dataset is classified as melanoma, seborrheic keratosis, or benign nevus (see Fig. 1).

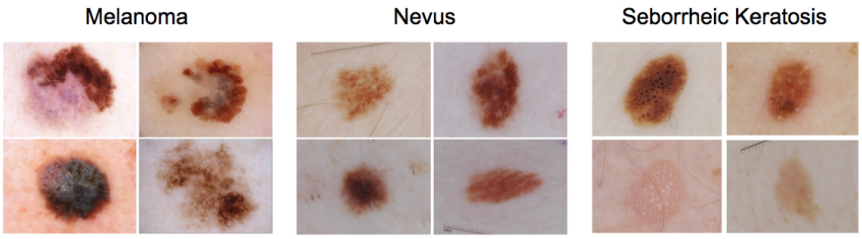

Figure 1: Label distribution of ISIC 2017 challenge.

3. Material and methods

In this study, our aim is to compare two different methods of skin classification, in which we classify into 'melanoma', 'seborrheic' and 'benign', using the train dataset of the ISIC2017 challenge. Firstly, an

approach consisting of the combination of deep learning to accomplish the feature extraction of the images, and machine learning to perform the classification task, is presented. Then, a second approach focused on a fully automated deep learning approach to perform the classification, is provided.

3.1. Approach 1: Combination of Deep Learning and Machine Learning

As mentioned before, the purpose of this method is to integrate the power of deep learning (to perform tasks such as feature extraction) with the classification power of machine learning algorithms in order to successfully classify skin lesions. To achieve the final result of this approach, multiple steps have been taken (see Fig.2).

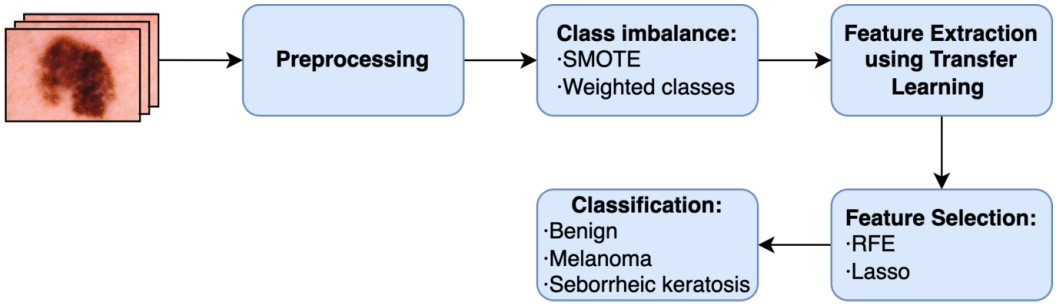

Figure 2: Followed steps to implement the approach 1.

These steps are as follows:

3.1.1. Preprocessing

Some preprocessing has been carried out to prepare the images for feature extraction. The use of a pretrained network in the *imagenet challenge*[5] required the preprocessing of the images in accordance with the pre-processed methodology employed during the network's training. In order to achieve the stated objective, we have employed the default preprocessing function provided by Keras (the corresponding function of the architecture that we are using). A visual example of this preprocessing is shown in the Fig.3 .

3.1.2. Class Imbalance

Due to the fact that the dataset provided exhibits class imbalance, with 1367 samples belonging to the benign class and only 373 and 254 samples belonging to the melanoma and seborrheic classes respectively, certain techniques have been used to address this issue. Those techniques are the followings:

- **SMOTE:** The Synthetic Minority Over-sampling Technique (SMOTE) algorithm is a commonly used approach to tackle the issue of class imbalance in the field of machine learning. The objective of this algorithm is to mitigate this concern by producing artificial instances for the minority class, thus achieving a state of equilibrium in the distribution of classes.
- **Weighted classes:** This approach involves addressing the imbalance of data by assigning different weights to the classifier, based on the respective number of samples in each class. Subsequently, minority classes such as 'melanoma' or 'seborrheic' are assigned a higher weight, whereas majority classes such as 'benign' are assigned a smaller weighting.

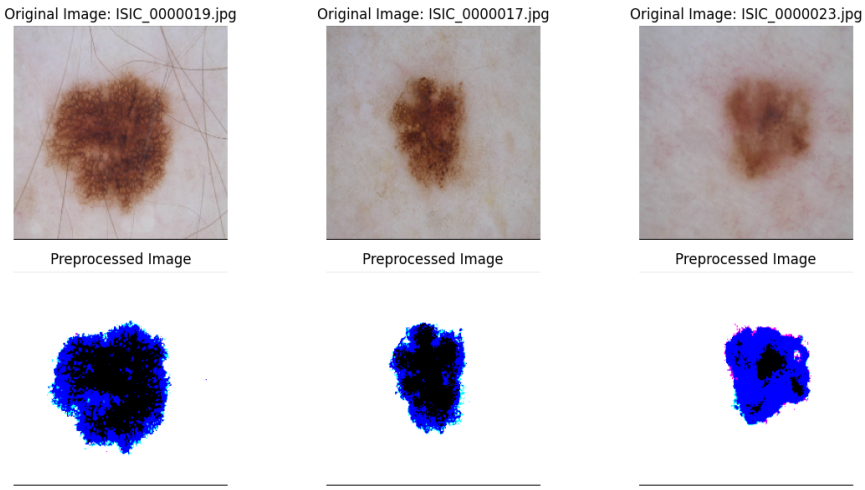

Figure 3: A visual example of the preprocessing.

3.1.3. Feature Extraction

As stated in the previous sections, in order to accomplish the feature extraction of the images, a pretrained network in the imagenet challenge has been used. Among a large number of models, DenseNet201 was chosen to carry out this assignment (see Fig 1). This network is a deep learning model that belongs to the DenseNet family of convolutional neural networks (CNNs) and represents an extension of the original DenseNet architecture. Contains 201 layers and is widely known for its impressive performance and accuracy in various image recognition challenges.

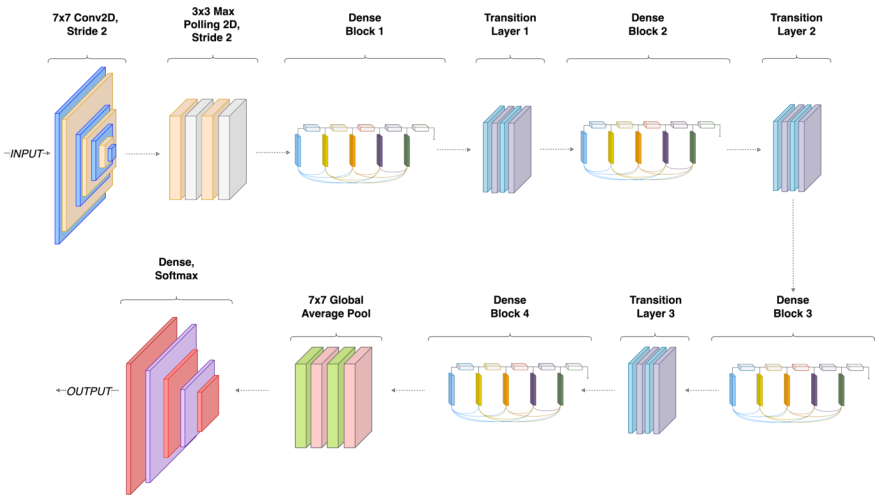

Figure 4: Example of the DenseNet201 architecture.

3.1.4. Feature Selection

In order to enhance the performance of this method and better select the features used for the classification of our model, a phase of feature selection is implemented. To complete this task, two distinct approaches were utilised. These techniques are as follows:

- **Lasso:** In machine learning and statistics, Lasso, also known as the least absolute contraction and selection operator, is a well-known method for selecting features. The Lasso algorithm encourages sparse coefficient estimates by incorporating a penalty term into the standard linear regression model. This technique is used as the initial phase in our approach's feature selection strategy.
- **RFE:** Recursive Feature Elimination (RFE) is a powerful feature selector method that eliminates less essential features iteratively and ranks the remaining features based on their contribution to the model's performance. The process is repeated until a predetermined number of features or a predetermined threshold is reached. This technique is utilised as the final feature selector in our approach, with a predetermined number of 75 features to retain. In addition, SVM and logistic regression have been used as estimators for the RFE model in this approach.

3.1.5. Classification

For the classification assignment,the following algorithms were utilised:

- Support Vector Machines (SVM)
- Logistic Regression
- K-Nearest Neighbors (KNN)
- Random Forest

A hyperparameter optimisation phase has been conducted for all classifiers in order to determine the optimal configuration for the task we wish to accomplish. Following that, in order to avoid overfitting, a cross-validation stage of 10 folds has been carried out. Moreover, for the training of the classifiers, one versus all classification technique has been specified. This method converts a multi-class classification problem into multiple binary classification problems. In each binary classification, one class is regarded as the positive class, while the remaining classes are merged into a single negative class, thereby permitting the classification of each class against all others.

3.2. Approach 2: Deep Learning classification

The objective of this technique is to effectively classify skin lesions, similar to the previous method, but this time using only the power of deep learning. Several steps were executed to achieve the desired result. These steps are as follows:

3.2.1. Preprocessing

Similarly to the prior method, since we are utilising pretrained networks, a preprocessing phase in accordance with the preprocessed methodology employed during the network's training, is required. To accomplish the stated objective, we have employed the Keras default preprocessing function for each of the tested architectures in this project.

3.2.2. Data augmentation

Due to the fact that the dataset provided was highly unbalanced, a data augmentation phase was performed before training the models. To achieve this objective, the images of the minority classes ('melanoma' and 'seborrheic') have gone through a number of transformations. These transformations include rotations, zoom ranges, and horizontal and vertical flips (see Fig.5).

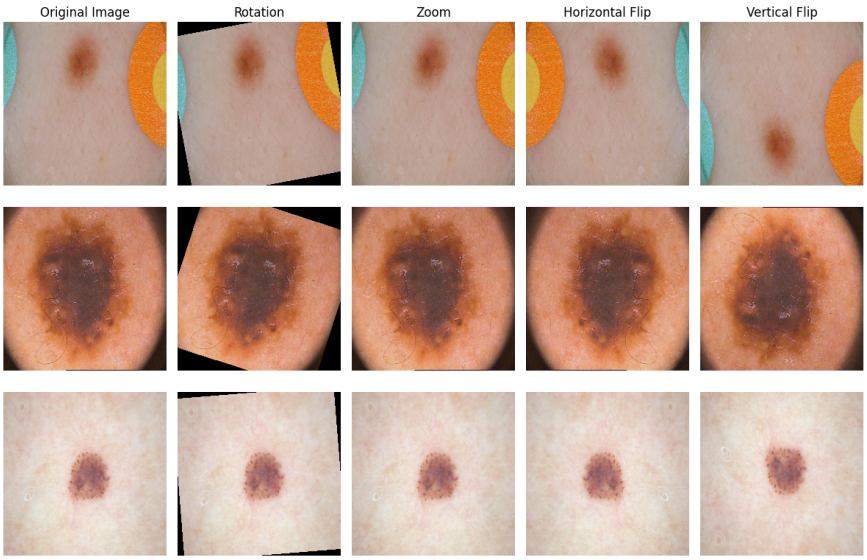

Figure 5: Example of the data augmentation performed in approach 2.

Once this phase was applied, the data were precisely balanced, with roughly the same number of samples for each class.

3.2.3. Model Creation

In this approach, as mentioned previously, multiple pretrained networks have been evaluated to determine which one best suits the task we wish to accomplish. Specifically, this approach have tested the following networks:

- Inception Resnet v2
- VGG19
- VGG16
- DenseNet169
- DenseNet201
- EfficientNetB0
- ResNet50v2
- ResNet152V2

Due to the fact that all of these networks were trained for the *imagenet* challenge, it was necessary to modify their architecture in order to classify the data into a different number of categories. In Fig.6 we can observe an example of these modifications applied to the *VGG16* model's architecture. The base of the model is composed by the specific layers of the model we are employing. However, the top layers of that architecture were modified to adapt the network for the classification task we wish to complete. The same modifications that were applied to the VGG16 architecture were also made to the previously mentioned architectures.

In addition, the same compiling configuration was used for all the models. That configuration is as follows:

- **Optimizer**: The optimizer used was the *Adam* optimizer with a learning rate of 0.0001.
- **Loss**: The loss employed was the '*categorical cross-entropy*', which is widely used for multi-class classification tasks.
- **Metric**: The metric used to evaluate the model during the training was the '*accuracy*' metric.

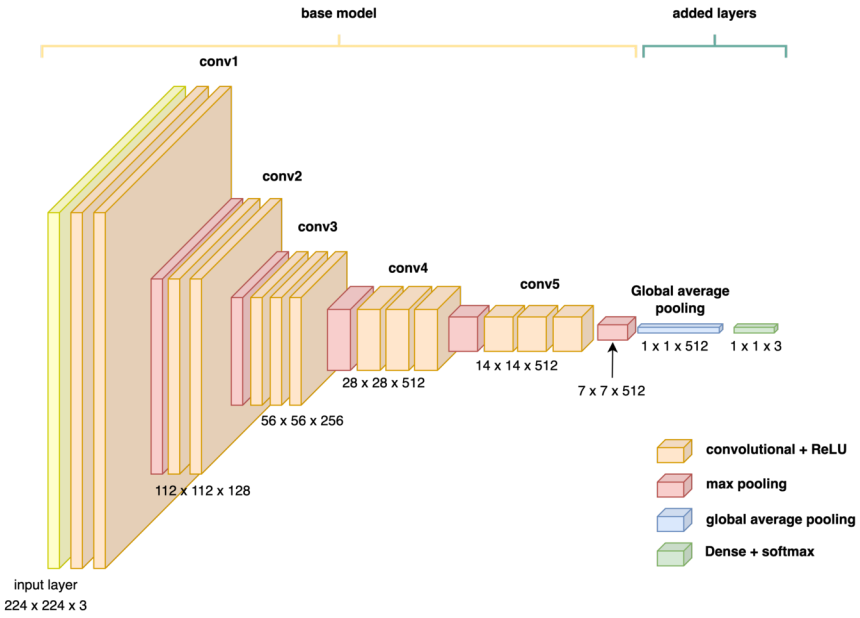

Figure 6: VGG16 model's architecture.

3.2.4. Training of the models

Throughout the training of the model, several parameters are used to improve the process. These parameters are as follows:

- **Epochs:** The number of epochs determines how many times the complete training dataset will be iterated during the training procedure. This method sets the number of epochs between 25-50, depending on the model.
- **Class weights:** Even though the data was balanced during the data augmentation phase, this may not be sufficient to address the problem. In order to assist that issue, class weights have been implemented.

The following weight combination was used to train our models:

  - Benign: 0.5
  - Melanoma: 2
  - Seborrheic: 0.7

  These weights were chosen empirically.

- **Early stopping callback:** In order to help to prevent overfitting, the EarlyStopping callback has been used. This callback will monitor the validation loss during training. If the validation loss does not improve for a certain number of epochs (in our case, 5), training will be stopped.

.2.5. Metrics

In order to quantify and compare the performance of the different approaches, balanced accuracy (BA) has been used as the main metric to evaluate and compare model's performance. This metric is widely used for multiclass classification. The formula followed to compute it, is the following:

$$\frac{1}{M}\sum_{m=1}^{M}\frac{tp_m}{n_m}$$

Where:

- M is the number of classes.
- tp_m is the number of true positives for the class m.
- n_m is the total number of samples of the class m.

3. Results

As previously stated, in this project several configuration have been performed for both DL and DL+ML. In the *Table 1* we can observe the quantitative results of the DL+ML approaches. These results consist of using several combination of class imbalance approaches together with two different classifiers. In addition, the ROC curve of each of the evaluated methods shown in the *Table 1*, are presented in the Fig.7.

Table 1: Quantitative results of Approach 1

Classifier	Class imbalance	Feature selection	BA	AUC	Average TPR	Average TNR
Logistic Regression	SMOTE	Lasso + RFE	0.64	0.72	0.64	0.72
Logistic Regression	**Weighted classes**	**Lasso + RFE**	**0.68**	0.74	**0.67**	**0.80**
SVM	Weighted classes	Lasso + RFE	0.67	0.75	0.67	0.82
Random Forest	Weighted classes	Lasso + RFE	0.65	0.73	0.62	0.74
KNN	Weighted classes	Lasso + RFE	0.55	0.66	0.54	0.76

In the *Table 2* we can observe the quantitative results of the DL approach. In order to not to make this report too extensive, only the best configuration of each model is presented in the table. Following that table, the accuracy and loss of the training and validation sets, are shown in the Fig. 8 and Fig. 9, respectively. Moreover, the ROC curve of each of the different architectures tested in this project, are presented in the Fig 10.

Table 2: Quantitative results of Approach 2

Architecture name	BA	AUC	AVG TPR	AVP TNR
INCEPTION_RESNET_V2	**0.72**	**0.81**	**0.72**	**0.90**
VGG16	0.68	0,77	0.67	0.86
VGG19	0.60	0.71	0.6	0.81
ResNet152V2	0.70	0.79	0.70	0.87
ResNet50V2	0.68	0.78	0.68	0.87
ResNet101V2	0.71	0.81	0.71	0.89

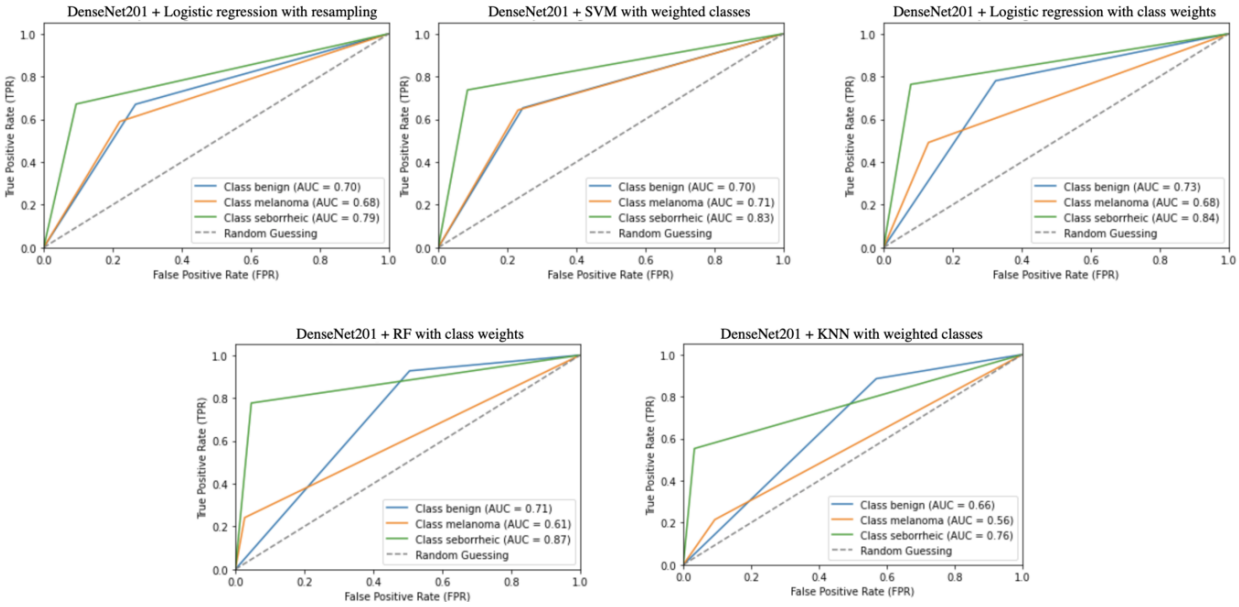

Figure 7: ROC Curve of each of the followed methods in the approach 1 of this project.

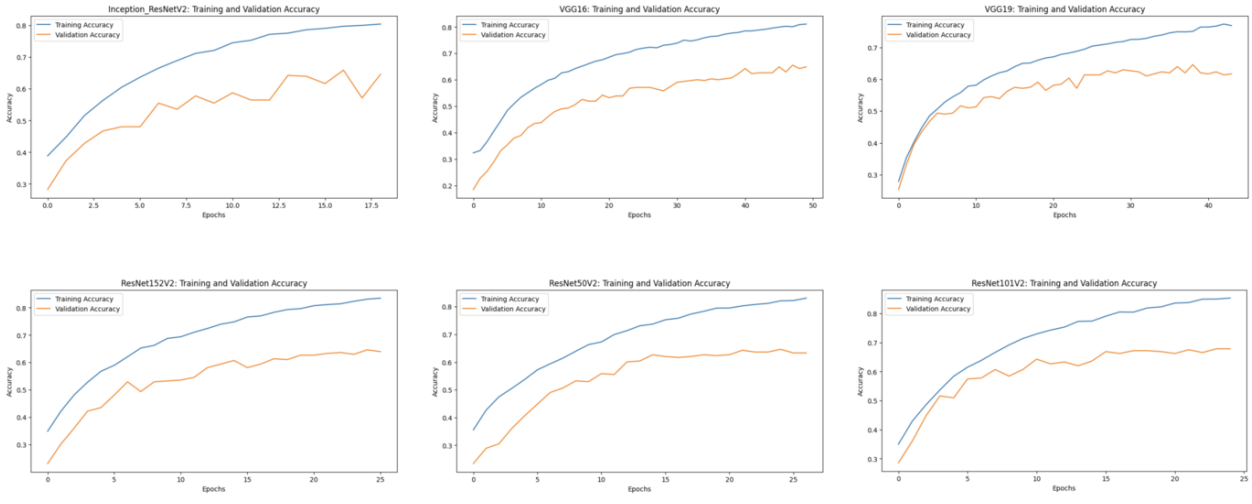

Figure 8: Training and validation accuracy of each of the pre-trained models employed.

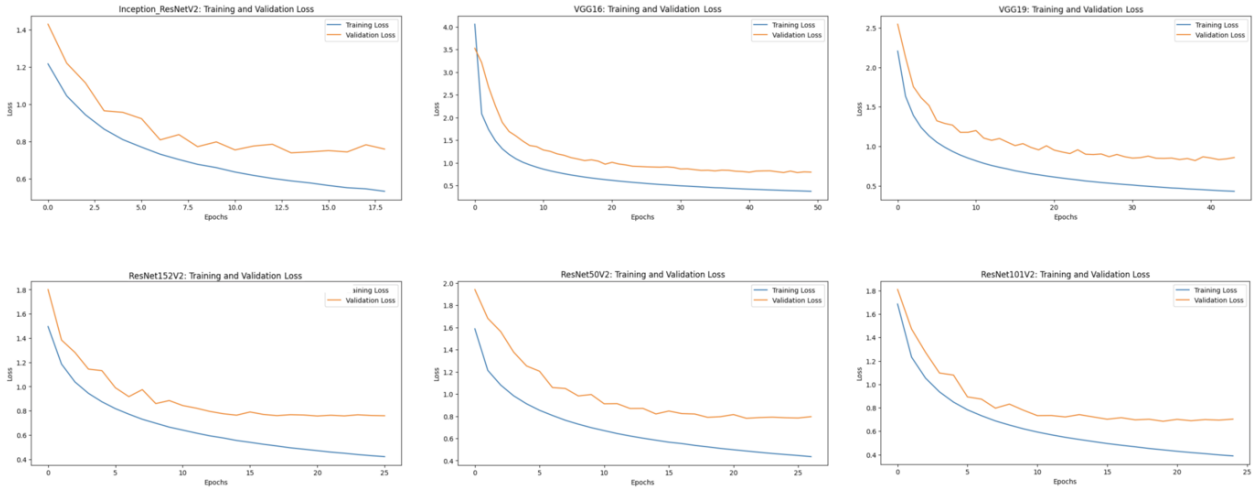

Figure 9: Training and validation loss of each of the pre-trained models employed.

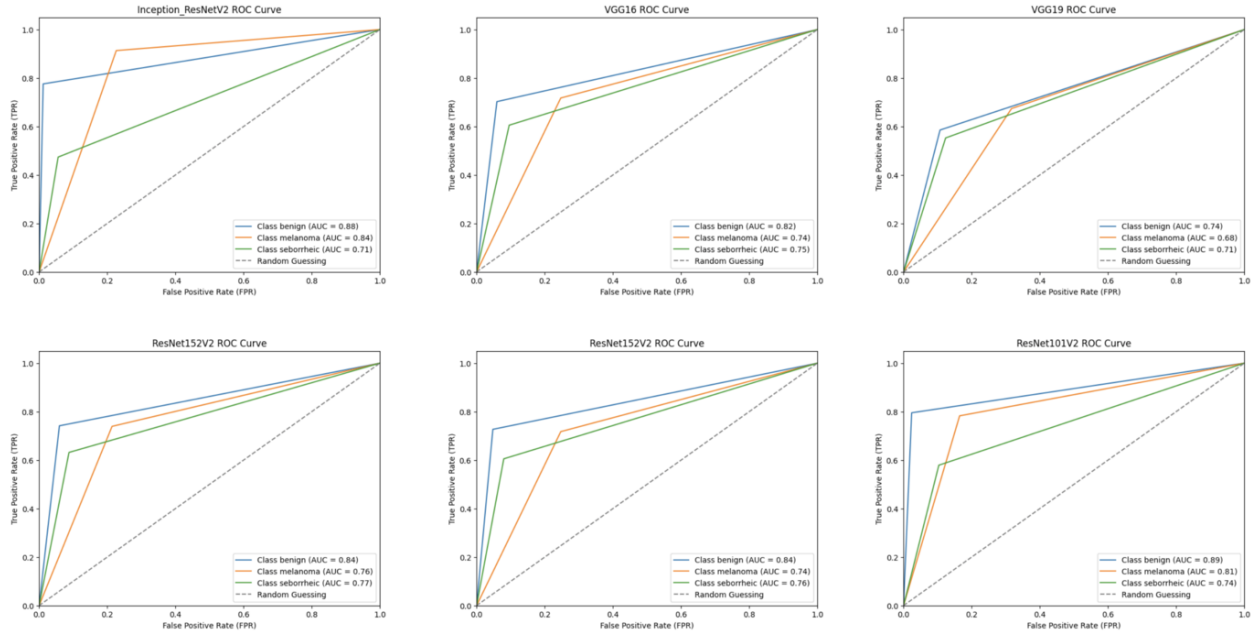

Figure 10: ROC Curve of each of the pre-trained models employed.

4. Discussion and Conclusion

This project presented two distinct skin classification techniques. The first method presented consists of using the power of DL together with ML classifiers. As can be shown in Table 1, with this method we have obtained good results reaching maximum balanced accuracy values of 0.68. The second and last method consisted of utilising deep learning to perform the task we wish to accomplish. As can be shown in Table 2, with this method we have obtained satisfactory results reaching maximum balanced accuracy values of 0.72.

As we can observe in the *Table 1* and *Table 2*, the deep learning approach has outperformed the machine learning one. Transfer learning for feature extraction together with machine learning techniques, has demostrated a very good performance. However, as expected, the performance obtained with deep learning was slightly higher. This difference can be due to various reasons, one of them could be the nature of the algorithms. In the case of the full deep learning approach, the algorithm incorporates both the feature extraction and classification stages in a single end-to-end learning process. This end-to-end learning allows the model to optimize the feature extraction and classification jointly, leveraging the interdependencies between these stages. In contrast, the ML approach separates feature extraction and classification, which may result in suboptimal performance if the selected features do not adequately capture the underlying patterns in the data.

In addition, deep learning algorithms, specifically deep neural networks, are designed to automatically learn hierarchical representations of data. Deep learning models are composed of multiple layers of interconnected nodes (neurons), each performing simple computations. These models learn to extract increasingly complex and abstract features from the input data as information passes through these layers.

In conclusion, deep learning strategies have demostrated to outperform machine learning approaches. Although ML techniques have their advantages, the transformative power of deep learning has made it the method of choice for confronting difficult problems in the field of medical imaging. However, even though the performance obtained with this algorithm seems promising, still has to be improved in order to use this information for the diagnostic, treatment and follow-up of real patients, for whom even the slightest error in clinical fields can be fatal.

5. Future work

Despite the fact that the project's outcomes have been promising, they can still be enhanced. Obtaining a larger dataset would be one method to enhance them. For deep learning approaches, the amount of data is crucial, as the larger the dataset, the more data the neural network has to learn from and, consequently, the greater its efficacy. In addition, once we have a larger dataset, we will be able to train neural networks from scratch, enabling them to learn information solely from the images we provide, as opposed to using information learned from the imagenet challenge, as was the case with pretrained networks.

Another way to improve the performance would be implementing cross-validation while training our neural networks. This would enable us to manage overfitting. In this project, we were unable to do so due to the limited GPU capabilities of Google Colab. Increasing computational power would enable us not only to execute cross-validation, but also to implement more complex models in order to better achieve the main task of this project.

References

[1] Randy Gordon. Skin Cancer: An Overview of Epidemiology and Risk Factors. *Seminars in Oncology Nursing*, 29(3):160–169, aug 2013.

[2] Rebecca L. Siegel, Kimberly D. Miller, and Ahmedin Jemal. Cancer statistics 2020. *CA: A Cancer Journal for Clinicians*, 70(1):7–30, jan 2020.

[3] Hyuna Sung, Jacques Ferlay, Rebecca L. Siegel, Mathieu Laversanne, Isabelle Soerjomataram, Ahmedin Jemal, and Freddie Bray. Global Cancer Statistics 2020: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries. *CA: A Cancer Journal for Clinicians*, 71(3):209–249, feb 2021.

[4] ISIC Challenge. https://challenge.isic-archive.com/landing/2017/. Accessed on Tue, May 30, 2023.

[5] ImageNet. https://www.image-net.org/challenges/LSVRC/. Accessed on Thu, June 01, 2023.

[6] Peter Goldreich and Pawan Kumar. Wave generation by turbulent convection. *The Astrophysical Journal*, 363:694, nov 1990.

[7] Pawan Kumar, Peter Goldreich, and Richard Kerswell. Effect of nonlinear interactions on p-mode frequencies and line widths. *The Astrophysical Journal*, 427:483, may 1994.