Deep Learning for Melanoma Classification: A Study Using Skin Lesion Images

Miguel Marinho Ferreira da Silva

Dissertation written under the supervision of professor Pedro Afonso Fernandes

Dissertation submitted in partial fulfilment of requirements for the MSc in Business Analytics, at the Universidade Católica Portuguesa on June 1, 2023.

Deep Learning for Melanoma Classification: A Study Using Skin Lesion Images

Miguel Marinho Ferreira da Silva

Abstract

Cutaneous melanoma is considered the skin cancer with highest mortality rate and has been gaining the attention of the medical community due to its rapidly increasing incidence. Advancements in computational technologies have paved the way for innovative image detection methods that can be transferable to medical applications, significantly enhancing the potential for early intervention in melanoma diagnosis. To make diagnosis more accurate and to further increase survival rates, this study employs deep learning techniques on an extensive dataset derived from multiple sources. Utilizing Microsoft Azure Cloud as the computational infrastructure, trial and error approach was employed by hyperparameterizing several convolutional neural networks (CNN) where the decision criteria were choosing the one with highest F_β Score. MAR-MELA-CNN is an innovative ensemble model incorporating six state-of-the-art pre-trained CNN architectures: Xception, VGG16, ResNet50, NASNetMobile, MobileNetV2, and InceptionV3. The primary goal of this research is to further understand CNN's efficiency in the diagnosis of melanoma and to furthermore measure its performance on a merged dataset. The proposed algorithm achieved a F_β score of 85%, an area under the curve (AUC) score of 93%, and an average precision (AP) score of 92%, promising diagnostic tool for cutaneous melanoma compared to traditional methods. Further improvements lay in the improvement of the architecture, expansion of the computational instances as well as of the dataset. Another field of future work could be devising a strategy for real-time implementation of this model in a hospital setting, as it could be of vital importance to provide swift support to doctors.

Keywords: cutaneous melanoma, deep learning, convolutional neural networks, F_β score, medical imaging.

Deep Learning for Melanoma Classification: A Study Using Skin Lesion Images

Miguel Marinho Ferreira da Silva

Resumo

O melanoma cutâneo é considerado o cancro de pele com a maior taxa de mortalidade e tem vindo a ganhar a atenção da comunidade médica devido ao seu rápido aumento de incidência. Os avanços tecnológicos contribuíram para métodos inovadores de detecção de imagens transferíveis para aplicações médicas, aumentando significativamente o potencial de intervenção precoce no diagnóstico de melanoma. Para tornar o diagnóstico mais preciso e aumentar a taxa de sobrevivência, este estudo emprega técnicas de aprendizagem profunda num conjunto alargado de dados provenientes de várias fontes. Utilizando a infraestrutura computacional Microsoft Azure Cloud, a abordagem de tentativa e erro foi utilizada ao hiperparametrizar várias redes neuronais convolucionais, sendo o critério de decisão a escolha daquela com a maior pontuação F_β. MAR-MELA-CNN é um modelo ensemble que incorpora seis arquiteturas pré-treinadas: Xception, VGG16, ResNet50, NASNetMobile, MobileNetV2 e InceptionV3. O objetivo principal desta investigação é potenciar a eficiência das CNNs no diagnóstico de melanoma e medir o seu desempenho num conjunto de dados unificado. O algoritmo proposto alcançou uma pontuação F_β de 85%, AUC de 93% e uma precisão média de 92%, tornando-se uma ferramenta promissora para o diagnóstico de melanoma em comparação com os métodos tradicionais. Os desenvolvimentos futuros incluem a melhoria da arquitetura e a extensão das ferramentas computacionais e do conjunto de dados. Outro campo de trabalho futuro poderia ser a criação de uma estratégia de implementação em tempo real deste modelo num hospital, já que pode ser de vital importância para fornecer apoio imediato aos médicos.

Palavras Chave: melanoma cutâneo, aprendizagem profunda, redes neuronais convolucionais, pontuação F$_\beta$, imagiologia médica

Contents

List of Tables

List of Acronyms

Concepts

ML Machine Learning

DL Deep Learning

PTSD Post-Traumatic Stress Disorder

AAPC Average Annual Percentage Change

Models

SVM Support Vector Machines

GBM Gradient Boosting Machine

LOGIT Logistic Regression

RF Random Forest

XGB Extreme Gradient Boosted Trees

CNN Convolutional Neural Network

KNN K-Nearest Neighbour

CART Classification and Regression Trees

ID3 Iterative Dichotomiser 3

LSTM Long Short-Term Memory

GRU Gated Recurrent Unit

BILSTM Bidirectional LSTM

BIGRU Bidirectional GRU

ANN Artificial neural networks

MLP Multilayer Perceptron

DLNN Deep Learning-Based Neural Networks

ELM Extreme learning machine

Methods

RFE Recursive Feature Elimination

RELU Rectified Linear Unit

SMOTE Synthetic Minority Over-sampling Technique

CNNR Condensed Nearest Neighbor Rule

GLCM Gray Level Co-Occurrence Matrix

CLAHE Contrast Limited Adaptive Histogram Equalization

MRF Markov Random Field

LBP Local Binary Pattern

SMOTE Synthetic Minority Over-sampling Technique

Metrics

AUC Receiver Operating Characteristic

AUC Area Under the Curve

ACC Accuracy

F1 F1 Score

TPR True Positive Rate

FPR False Positive Rate

TP True Positive

FP False Positive

TN True Negative

FN False Negative

1 Introduction

Cutaneous melanoma is considered the skin cancer with the highest mortality rate. Due to it not being easily detectable in an early stage as it approximates other skin lesions in appearance, this could translate into a spread to other parts of the body (Chang et al. (2022), American Cancer Society (2019)). The new cases of melanoma that may be diagnosed in the United States during 2023 are about 100 thousand, a 26% increase from 2013. The consistently increasing prevalence of melanoma over time is a cause for global concern, as not only it translates to a public health liability, but to a weakness to society as a whole. It is key to note that the five-year survival rate for melanoma is relatively high when the disease has not spread, drastically decreasing when it has metastasized to other parts of the body, thus exemplifying the urgency of early diagnosis (American Cancer Society, 2019).

Melanoma is one the most aggressive types of skin cancer capable of infiltrating deep skin tissue and spreading to other parts of the body, consequently, it is critical to understand the significance of the early diagnosis of melanoma. Prompt discovery of the tumor drastically increases the chances of a successful treatment, having a 98.4% five-year survival rate if efficiently found with the comparison of a 22.5% survival rate if it has spread to other parts of the body (Holtel, 2022). Although complications arise, as even experienced dermatologists struggle with the identification of melanoma due to its complex nature, studies show that 10% of the time there is a misdiagnosis in the first clinical visit (Sondermann et al., 2016). Main adversities comprise of it being almost indistinguishable from common benign skin growth (Vocaturo et al., 2019), certain areas in the body are overlooked (Holtel, 2022), tumors have varied physical appearance (Mayo Clinic Staff, 2022), limited access to healthcare (i.e. Pandemic (Grady, 2020)), and when it is tested it takes a significant amount of time before the result is received.

The importance of this research will be translated into all the ways deep learning could revolutionize the healthcare sector in terms of early detection of this disease. The introduction of machine learning has provided pioneering applications in all sectors, being highlighted in the medicinal context when assessing predictive techniques. Furthermore, with continuous research applied to machine learning, deep learning methods were introduced and have had favorable results with regard to image-based classification in various sectors of healthcare, providing a promising basis for the application to melanoma detection.

Deep learning is comprised of artificial neural networks, which contain in them an algorithm that has been gaining traction recently with regards to image recognition quality, convolutional neural networks (CNN) (Vocaturo et al., 2019). This method consists of layers that apply convolutional filters to detect features, which are then combined and transformed through convolution, pooling, and activation function put through a final layer that extracts probabilities of output classes. The study of the applications of this modeling technique has been limited in the context of melanoma due to consistent issues encountered by researchers, namely the impact of data limitations. And importantly, as mentioned by Balkenende et al. (2022):

> *A big downside of DL is its need for large datasets to achieve satisfactory performance.*

As it is perfectly put in the above statement, the efficiency of these algorithms is often dictated by the availability of data. The purpose of this research is to further investigate and understand the applications of CNN, more specifically developing an efficient, extensive algorithm that detects and classifies melanoma. Moreover, to provide a bigger framework that can be picked up for further study, with the introduction of a mixed dataset originating from various sources, to mitigate the common researcher struggle (Balkenende et al. (2022)), henceforth testing the limits of the quality of this deep learning technology when presented with data sourced from multiple origins.

Furthermore, a perspective that is not highlighted significantly in literature is the importance of choosing the correct metric in the context of medicine. It is crucial to recognize that there is a higher societal impact when diagnosing a person that truly has melanoma inaccurately (False Negative) than misclassifying someone as ill when not having cancer (False Positive). Thus, FP and FN cannot be weighed at equal cost as highlighted by Panwar et al. (2020). If an individual falls under the false negative category, the disease may spread unnoticed, hindering a potential successful recovery and decreasing the patient's survival chance from 98.4% to 22.5% (Holtel, 2022).

Ergo, highlighting the importance of a machine learning model having high recall with the intent of minimizing false negatives. Although, it is critical to understand that the initial intuition of maximizing recall may not be the most appropriate approach, as it can lead to a high number of false positives. Thus, it is proposed to assess the model performance through the F$_\beta$ Score, which is the weighed harmonic mean of precision and recall. Defining β will allow for the desired customizable emphasis on recall without disregarding the realistic importance of the impact of having high false positives:

$$F_\beta = (1 + \beta^2) \cdot \frac{Precision \cdot Recall}{(\beta^2 \cdot Precision) + Recall}$$

Accordingly, the main contribution of this paper can be summarized in the following:

- Are Deep Learning algorithms effective at predicting outcomes based on human images?
- Can CNNs achieve high model quality when applied to a heterogeneous set of image datasets with varying imaging characteristics, such as different perspectives and cameras used for acquisition?

Ipso facto, with innovative deep learning algorithms it will potentially increase the accuracy of diagnosis minimizing misdiagnosis, and increase efficiency as CNN can analyze image scans significantly faster than a human dermatologist. Additionally, these algorithms could be beneficial for countries where healthcare access is limited where skilled dermatologists and expensive technology are not present, and lastly, could be applied to diagnose tumors that are in remote areas of the body and require complex procedures.

2 Background

2.1 Medical Context

Cancer is a group of diseases characterized by the uncontrolled growth and spread of abnormal cells in the body (Centers for Disease Control and Prevention, 2023a). These abnormal cells can invade and damage healthy tissue, including organs, further stimulating mutations that hinder cell growth. There are numerous varieties of cancer, each possessing its own unique set of characteristics and risk factors. Early detection through prompt diagnosis within a feasible time frame can greatly reduce the severity of the condition and mitigate associated risks, thereby highlighting the importance of timely identification.

The majority of cancer fatalities, about 71%, occur from cancers for which there is no established screening protocol (American Cancer Society, 2023). On the other hand, when it is diagnosed, as reported by the Surveillance, Epidemiology, and End Results (SEER) Program Stat Database, in the stage that has no spread, there is a higher chance of survival with a rate of 89% compared to a much lower survival rate of 21% when cancer has advanced, spread and metastasized (National Cancer Institute NCI, 2018). This context exemplifies the importance of early diagnosis prior to metastasis when assessing cancer in general.

According to the Centers for Disease Control and Prevention (CDC), cancer is the second leading cause of death in the United States. In 2021, an estimated 605,213 individuals in the United States died from cancer, accounting for approximately 25% of deaths among leading illnesses. Within the categories of cancer, lung cancer, colorectal cancer, and pancreatic are amongst the most common types that lead to death in the United States (Centers for Disease Control and Prevention, 2023b). According to the American Cancer Society, it is visible that since 2019 to the predicted values of 2023, there has been a sustainable upward trend with regard to fatalities. Considering the 2023 projections, there will be an 11.1% increase in cancer fatalities when compared with 2019 data and a 28% with 2010 data. It is also observable that from 2019 there was an average annual percent change (AAPC) of 2.7%, which consolidates the perspective of a consistent increase (American Cancer Society, 2023) .

Skin cancer can be classified into three main categories: basal cell carcinoma (BCC), squamous cell carcinoma (SCC), and melanoma, where the development of the latter occurs in the cells (melanocytes) that produce melanin. Furthermore, this skin injury can visibly appear as a mole, a pigmented lesion alteration or even appear as normal-looking skin generally having a uniform color (tan, brown or black) sized at around 6mm. Although most melanoma lesions appear in visible locations, there are also instances of hidden melanoma which appear in areas barely exposed to the sun. These locations are usually very remote such as under the nail, genitalia, inside the body, and in some cases, the eye. It is important to highlight that there is no cause for the diagnosis of melanoma, however, exposure to ultraviolet light is highly correlated identification of this type of skin cancer. Thus, people with less melanin in their skin, light-colored hair, those that sunburn easily, or have more moles than average are at additional risk (Mayo Clinic Staff, 2022).

In consonance with Seebacher (2022) and National Cancer Institute (2022), there are many staging systems that hospitals use to classify the progression of this cancer due to its complexity, but the most conventional system is as follows. The most common staging system for melanoma is the TNM system, which looks at three key factors: primary tumor (T: TX, T0, T1, T2, T3, T4), regional lymph nodes (N: NX, N0, N1, N2, N3), and distant metastasis (M: MX, M0, M1). T indicates the tumor size and thickness, TX implies the tumor cannot be measured and T0 that it cannot be found, the following stages grow depending on tumor size. N assesses the number and positioning of lymph nodes that have cancer present. NX implies cancer in lymph nodes cannot be measured and N0 that it cannot be found in the lymph nodes, the following stages grow depending on the number of lymph nodes affected by cancer. M assesses whether it has metastasized to other parts of the body, MX translates that metastasis cannot be found, M0 that there is no spread to other sectors of the body while M1 indicates positive metastasis. Melanoma in situ, also known as stage 0 melanoma, exists when there are cancer cells on the top layer of skin, known as the ”epidermis”. The higher the stage number, the more advanced the cancer has become in terms of size and spread, spreading to nearby lymph nodes/tissues before ultimately reaching the fourth stage of metastasizing to other parts of the body.

A recent study showed perspective of the trend of invasive melanoma segmented by lesion size (Sacchetto et al., 2018). The sizes assessed are in situ, thin (< 1mm), and thick (> 1mm). Using half a million records of data regarding skin lesions covering 18 European cancer registries from 1995-2012, this study painted a clear picture of an upward trend regarding invasive cutaneous melanoma and non-invasive cases in the assessed years. Overall, with regards to invasive melanoma, there was an AAPC of 4.0% for men and 3.0% for women. Distinguishing between magnitude, it was concluded that thicker lesions had increased incidence but with smaller impact when compared with thinner lesions (10% men; 8.3% women). In contrast, in situ had the steepest positive trend with an AAPC of 7.7% in men, and 6.2% in women. Thus, it was concluded that although melanoma is comprised of small amounts of cases it must not be disregarded due to it being a minority since it has been steadily increasing. Following this, with properly established agreements with medical institutions (for data gathering purposes), this could be a basis for further study in terms of DL applications. While melanoma may not be as prevalent as other types of skin cancers, it cannot be disregarded due to its smaller proportion.

> *Melanoma accounts for only about 1% of skin cancers but causes a large majority of skin cancer deaths.* (Grady, 2020)

With the continuous appearance of illnesses, attention to healthcare is paramount to the sustainability of the Human race. This is reflected in the aforementioned statement, where it is evident that although not frequent, it represents the highest death toll within its category. However, during times of crisis, the capacity of healthcare systems to effectively address the needs of those affected by cancer may be compromised. Furthermore, as examined in the previous pandemic, medical systems worldwide collapsed in the face of adversity and left cancer patients marginalized since attention shifted. Not solely because of the excessive amount of people that had to be attended to, conversely, there was a dire responsibility to avoid Covid-19 infections among cancer patients since this had fatal implications, all because the "health care system just cracked open and swallowed them up" (Grady, 2020). Thus, cancer patients were not prioritized as they did not fall under the emergency category, hence, this begs the need for some sort of automation in all medical sectors, even if it only impacts 1% of the cases (American Cancer Society, 2019).

The American Cancer Society predicts that in 2023, the United States will experience an estimated 99,780 new cases of melanoma skin cancer. This projection includes 57,180 diagnoses among male individuals and 42,600 among female individuals. Out of the 12,470 predicted deaths from skin cancer, 64% of them are expected to be caused by melanoma (American Cancer Society, 2023). The 5-year survival rate for melanoma that has not metastasized to other lymph nodes is estimated to be as high as 98.4%, which decreases to 63.7% in the event of lymph node involvement and further decreases to 22.5% when the disease has spread to other parts of the body (Holtel, 2022). In addition, when compared to the 2013 figures, the number of individuals predicted to be diagnosed with melanoma was 76,690 (9,480), which in retrospect is a 26% increase in only 10 years (Shoieb et al., 2016).

Currently, the most common method of diagnosing melanoma involves extracting the mole and sending it to pathology where it is assessed if it is malignant, which can be a long painful process that may not always be available. Furthermore, dermatologists have repeatedly expressed difficulty in the identification of early malignant melanoma from a mole (Chang et al., 2022), especially if it is in a remote physical location. Consequently, it must be emphasized that a conventional automated, and accessible way to diagnose melanoma is crucial to avoid metastasis, with the goal of maximizing the survival rate of every individual with minimal pressure on hospitals.

2.2 Data Science Applications in Medicine

Data science is currently the most innovative topic in all industries and has recently been the target of use in the healthcare sector. The exponential growth of healthcare advancements in computational power is undeniable, machine learning has become an indispensable tool in the medical field having the potential to make healthcare more efficient, accurate, and faultless.

The most relevant context where data science must be applied in medicine is predictive analytics. Considering that data availability has been gradually growing, we can apply efficient machine learning algorithms to identify future outcomes based on historical records, thus, allowing for data-driven clinical decision-making. Furthermore, it has been taking on greater significance in the context of biomedical research. In a more practical context, machine learning can be used to predict patient outcomes with high confidence, such as risk modeling and stratification, personalized screening, diagnosis, prediction of response to therapy, and prognosis (Baccouche et al., 2020).

By analyzing health records (blood pressure, body temperature to sugar level), these innovative techniques can identify patterns and correlations that can originate in data-driven clinical decision-making, inherently improving patient outcomes (Narasimman, 2023). Furthermore, these revolutionary procedures can be used for medical imaging deep learning applications such as X-ray screenings, MRIs, tumor detection, ultrasound, ct scans, and lesion detection applied to skin cancer, more specifically melanoma. With continued research, this can lead to improved accuracy and efficiency in melanoma detection, helping to save lives and to moderate disorder in hospitals when in times of crisis (Grady, 2020).

It is important to understand that when contextualizing machine learning in medicine False Positives and False Negatives cannot be weighed at equal cost (Panwar et al., 2020), especially with an aggressive cancer like melanoma. There is a greater societal impact in diagnosing someone that truly has a disease incorrectly (False Negative) than misclassifying someone that is not ill (False Positive). Considering that in the case of false negatives the person will unknowingly have cancer, this can lead to delayed treatment, cancer spreading, and becoming more difficult to treat. In some cases, a false negative result can mean the difference between life and death for the patient especially in light of the survival rate, as statistics show that it goes from 98.4% to 22.5% if there is spread to other parts of the body (Holtel, 2022). Thus, highlighting the importance of a machine learning model having high recall with the intent of having minimal false negatives. On the other hand, the model having high false positive values will lead to unnecessary and potentially harmful procedures, such as biopsies or surgeries. Although this misdiagnosis has a stressful impact on patients, in retrospect, it is not as impactful as having cancer go undetected by the model.

To further understand to what extent there are applications of machine learning in the sector of medicine there must be an appropriately conducted research. Survival analysis has been an important area of study in medicine, where various predictive analytics methods have been applied to analyze death, disease progression, and the probability of survival over time. A basic example of these types of predictive methods would be linear modeling. Although there are limitations of such models, stemming from the model's difficulty in establishing patterns that have irregular distributions, pinpointing elements affecting a specific subgroup, and applying predictive methods. Thus, machine learning is a key topic as it enables the creation of algorithms that accurately categorizes individuals based on their diverse and unequal risk factors (Galatzer-Levy et al., 2017).

2.2.1 Machine Learning Approach in a Medical Context

Galatzer-Levy et al. (2017) conducted a study regarding the diagnosis of post-traumatic stress disorder (PTSD) using a Support Vector Machines (SVM) model. A total of 152 participants were considered for the purpose of this paper, after being admitted into the emergency room (ER) following the trauma. With the purpose of making a consolidated study, appropriate machine learning practices were applied. Considering the dataset had a substantial amount of columns, RFE was used. To ensure stability in the columns chosen, a measure was developed to indicate the percentage of times a feature was selected through RFE across 5 × 10-fold cross-validation runs. Thus, a "stable feature" was considered one that is selected at least 55% in RFE runs. The prediction outcomes were determined by considering different sets of information. With only background information, the model had a prediction result of 64% AUC. The addition of variables from the emergency room improved the value to 82% AUC. Further data collection, through one week, resulted in a prediction result of 88%. The highest prediction result of 93% AUC and 70% recall was obtained by utilizing data collected over one month. These results suggest that machine learning is a powerful tool for diagnosing PTSD, especially when combined with appropriate feature selection methods. Although when assessing the recall metric, which is the most relevant metric in medicine, it has relatively weak results.

In the study by Chekroud et al. (2016), a machine learning approach was developed with the core goal of identifying symptomatic remission after taking citalopram, an antidepressant, after 12 weeks. The data was collected from 1949 patients with depression and through feature selection, out of 164 reported variables 25 were considered the most influential hence being selected. The classification-based model that was utilized was Gradient Boosting Machine (GBM), which achieved 64.6% accuracy, and a 70% AUC. In an attempt to validate this model in other datasets, the statistically significant accuracy when applied to the escitalopram group was 59.6% in the Combining Medications to Enhance Depression Outcomes (COMED) study and 59.7% for the combination of escitalopram and bupropion in the same study. Additionally, the authors also used this model to attempt to predict the scores of a questionnaire used to measure the severity of depressive symptoms and found substantial overlap between the predictors of remission/non-remission and final scores. This study suggested sufficient predictability in the diagnosis of depression when using the classical ML approach, although having very limited data can easily bias results.

In similar and more recent depression research by Wallert et al. (2022) predicting remission, the findings go in line with the previously stated results. Conversely, to Chekroud et al. (2016), remission considerations are in view of guided therapy rather than antidepressants. This study has 894 patients with mild-to-moderate major depressive disorder (MDD) with over 1000 variables to which through psychological/psychiatric expertise, statistical interpretation was narrowed down to 69 predictors. Following that, the RFE method was applied to nullify human bias retaining 45 predictors. Applications of several machine learning models led to many results, but none where there were significantly high AUC percentages. LOGIT logistic regression, RF random forest, and XGB extreme gradient boosted trees had 65.4%, 68.7%, 66.6% AUC respectively where naturally, the RF model was chosen as the final one. Seemingly enough, it appears that there are limitations to automated diagnosis within each disease which is visible when assessing depression, although this could be due to the dataset size, as the paper suggests an improvement could be its extension. Furthermore, it is mentioned that an image based deep learning approach to this mental illness could be more effective in its classification, potentially in a CNN approach.

Vaishnavi et al. (2022), conducted a study predicting mental health illness using several machine learning techniques. After thoroughly assessing a set of data containing 1259 points with 27 attributes, the authors treated the data and through feature importance selected the most relevant attributes to mitigate multicollinearity, resulting in 8 columns. After the initial steps were performed, the modeling stage was initiated using Logistic Regression, K nearest neighbour classifier, decision tree classifier, random forest classifier to the final stacked ensemble algorithm model. Following the model hyper-tuning, in reference to the accuracy metric, the results were of similar nature: 79.6%, 80.4%, 80.7%, 81.2%, and 81.7% respectively. Although, when assessing the ROC_AUC metric there was a slight increase in value: 86.0%, 87.0%, 86.0%, 90.0%, and 86.0% respectively. It is concluded from this study that these classifiers provide optimal values for prediction with regard to the various problems within mental health. All things considered, it is interestingly mentioned that a big limitation is the dataset and that further study should be carried out when more resources are available.

In the study orchestrated by Bosl et al. (2011), predictive techniques were utilized to categorize infants into either the high-risk group for Autism Spectrum Disorder (ASD) or the control group. There were 79 infants as samples, where each human's data 6, 9, 12, 18, and 24 months were treated as unique datasets. Furthermore, 46 were considered high risk for ASD while 33 were healthy controls, based on multiscale entropy calculated from the brain activity recording (EEG) data when in a resting state. The K-Nearest Neighbors (KNN), SVM, and Naïve Bayes algorithms were all employed for classification with the aim of comparison and concluding which is the best classifier for the data. The results showed that both KNN and SVM produced an accuracy of 77% at 9 months of age. However, the highest accuracy which is marked at 80%, was obtained at 18 months of age through the Naïve Bayes algorithm. A key point made by the author is the importance of these results as they are statistically significant, and when compared with other ages they had high accuracies but were not significant.

From another perspective, the study by Omar et al. (2019) applied ML methods regarding the binary diagnosis of autism disorder. The model was trained on three datasets with varying ages, having children aged from 4 to 11 years old, adolescents aged from 12 to 16 years old, and adults aged 18 or older. Moreover, to validate the model's learning, the algorithm was tested on a subsample of this same dataset and on a real dataset of 250 individuals. The algorithms applied were the Decision Tree-Classification and Regression Trees (CART) which were later improved to the Random Forest-CART and eventually to be merged with the Random Forest-ID3. When applied to the training dataset the best-merged model had 92.26%, 93.78%, and 97.10% in accuracy on the three subsets respectively, and when this best model was applied to the real dataset the values of accuracy dropped to 77.26%, 79.78%, 85.10%. It is also mentioned that the latter prediction may be hindered since the data was collected through a survey which can lead to human bias as answers honesty comes into play, furthermore, data is limited. The result of this study offers a streamlined and efficient method for identifying the presence of autism in individuals of different age groups.

In response to Covid-19, Panwar et al. (2020) applied a deep learning algorithm, nCOVnet, to apply image classification using X-Ray images. The authors highlight the importance of automation in the context of having to manually interpret X-rays and give a medical diagnosis, and how this can be enhanced through deep learning capabilities, hence introducing nCOVnet. Transfer learning is also applied with the Visual Geometry Group 16 layers (VGG16) as the base model, the layers of the nCOVnet consist of average pooling, flatten, dense connected layer of 64 units, the Rectified Linear Unit (ReLU) activation function and finally applied 0.5 dropout. The model predicted correctly 97.62% of patients with Covid-19, implying incorrect positive diagnosis was only 2.38%, furthermore, 78.57% of patients with negative Covid-19 results were properly diagnosed providing an overall model accuracy 88.10%. An interesting interpretation of results is brought to light as there is greater societal impact in diagnosing someone that truly has Covid-19 incorrectly (False Negative) than misclassifying someone that is not ill. Further concluding that if the patient will be sick unknowingly there could be severe consequences. What is also highlighted as a limitation that is made in many studies, is the common lack of consideration of data leakage as datasets many times come with multiple photos of the same diagnosis, inferring that the model can be learning from results rather than from training.

In a study by Baccouche et al. (2020), a deep learning model was developed for classifying heart diseases. To balance data, techniques like Synthetic Minority Over-sampling

Technique (SMOTE) and Condensed Nearest Neighbor Rule (CNNR) were used, along with RFE to highlight key features. For the given task, an ensemble learning framework consisting of two different neural network models was constructed and the final prediction would be the concatenation of the outputs (highest performance of each model). Five different architectures were evaluated for the models including CNN, Long Short-Term Memory (LSTM), Gated Recurrent Unit (GRU), Bidirectional LSTM (BiLSTM), Bidirectional GRU (BiGRU), and multilayer perceptron (MLP). To mitigate overfitting there was epoch limitation and dropout rate inclusion, Adadelta was the chosen optimizer and learning rate was valued at one. The researchers used 800 data points to predict four types of heart diseases. When dealing with results it is demonstrated that CNN and BiLSTM models outperform other models in the context of F1 and accuracy for three out of four heart diseases. For the fourth disease, CNN and BiGRU models performed best. Bidirectional recurrent neural network models, BiLSTM and BiGRU, improved scores to about 92-94%. The ensemble model, as the authors claim, merges the strengths of both models thus outperforming the individual models. The CNN model outperforms the MLP model as it can go deeper and optimize parameters, LSTM and GRU models outperform the unidirectional models as it takes advantage of two-directional context information. In comparison to SVM models which had a maximum accuracy of 68%, the models being evaluated were highly successful.

A study conducted by Wang et al. (2017) compared the performance of DL models with non-DL models with the objective of classifying prostate cancer. Through magnetic resonance imaging, the acquisition of images was executed, gathering a total of 2602 images. After applying 10 fold cross-validation, CNN was the main model that was hyper-tuned to the following parameters: gamma 0.1, momentum 0.9, weight decay 0.1, and maximum training iteration 1000. The model consisted of five convolution layers and two inner product layers. Methods applied were a max-pooling layer to reduce the size of the feature map and a non-linear ReLU layer after each convolution. SVM is the non-DL model being used, and further data augmentation was applied to both models being studied. With context to the traditional machine learning model, a method for image recognition named bag-of-word was used, and scale-invariant feature transform feature extraction. Results consisted of the deep learning application outperforming the machine learning model, with an overall performance of 84% to 70% AUC. When contextualizing recall, 69.6%, and 49.4% are the respective value implying low correct identification of prostate cancer, as false negatives are high. The author concludes the importance of DL in image-based issues.

Research was carried out by Khan et al. (2019) introducing a DL framework for breast cancer detection using transfer learning. Data was gathered from a benchmark dataset and from a hospital, with a total of 8000 images. In the context of pre-processing methods, data augmentation was introduced with the goal of increasing the data samples and mitigating overfitting. Based on GoogleNet, VGG13 and ResNet did not vary drastically. With regards to accuracy and considering an 80-20% train-test data split, 93%, 96%, and 96% were the achieved values respectively, underperforming the proposed framework of 97.7%. When assessing the recall metric the results were also very high: 94%, 93%, 98%, with the final proposed method having 97%. The conclusion is that transfer learning had successful outcomes when used in the context of DL and outperformed previously mentioned literature, highlighting its importance.

2.2.2 Machine Learning Approach in the Context of Melanoma

In a research paper written by Vocaturo et al. (2019), an overview is provided based on several papers of literature review in the context of machine learning methodologies when tackling the automated malignant melanoma (MM) detection issue. KNN, decision trees, logistic regression, artificial neural networks (ANN), multi-variate logistic regression, and neural networks are the main mentioned strategies. When applying KNN to images of malignant melanoma, through radiometric and shape features the model classified recall of 87% with a specificity of 92%. Moreover, the author also found that when using the MED-NODE dataset after applying pre-processing techniques, using the decision tree model accuracy had 82.35%. In other cases it was observed that by applying the same model using DERMIS and DERMQUEST dataset as training data the accuracy jumped up to 92%, Vocaturo thus concludes that decision trees indicate that this approach had great performance compared with most techniques in the scope of melanoma diagnosis. Multi-variate logistic regression was also studied in an attempt to predict the risk factors of MM with their relative position to the body, based on several examined risk factors. A past of sunburn at an anatomical position in the body was correlated to the development of MM when compared with other body sites. In other discovered information, following the same approach was modeled a survival prediction related to cutaneous melanoma. Categorizing by tumor progression, 122 invasive lesions were found pre-metastases, thusly improving survival probability within patients, as concluded by Vocaturo. Using artificial neural networks the proposition of early-stage detection of melanoma applying first-order extraction using a multilayer perceptron neural network achieved an accuracy of 83.86%. Furthermore, concluding that CNN is the highest-performing image processing algorithm based on the artificial neural networks architecture. It was also found that applying the multiple instance framework there were positive results in medical image classification, proving its worth. With context to the issue at hand, in a small but balanced dataset, accuracy, recall, and specificity had 92.50%, 97.50%, and 87.50% respectively as values. It was claimed that this spoke to the intuitive extrapolation that this technique can be a good basis for the research of more sophisticated approaches. Finally, the author found that CNN has been a target of continuous study in the application of medical image classification applied to health issues. Vocaturo mentions a study that applied VGG16, VGG19, and GoogleNet pre-trained models through fine-tuning on the DermNet dataset achieving a 73.1% accuracy on the latter and 69.5% top five accuracies on the OLE dataset. Another similar study analyzed by the author postulates that by using a pre-trained model for skin injury diagnosis, it achieved an 85.8% accuracy in a five-class predictive model, which truly helps in case of data limitations/quality. It is also mentioned that when assessing quality in a CNN model, a non-augmented dataset will underperform an augmented dataset in the context of skin lesion classification using the accuracy metric.

There was a very interesting melanoma detection study by Chang et al. (2022) with regard to predictive applications. Prior to applying machine learning to this study, it was discovered many relevant applications of machine learning in melanoma detection such as using CNN, stacked CNN, and traditional machine learning algorithms. After merging the two public datasets (ISIC 2019+2020: 2299 records) there was still a class imbalance issue where the target class, malignant melanoma (450), had significantly lower data points. Thus, through SMOTE, oversampling techniques were applied to mitigate this issue. The base models being applied to model training consist of a dermatologist handcrafted method, and three deep learning methods that are already pre-trained with a large dataset: VGG16, InceptionV3, InceptionResNetV2, and a MELA-CNN. The latter was developed by the authors using InceptionResNetV2's architecture as the core of its creation. After combining the five initial approaches with an XGB (for performance measurement purposes) 6.4%, 28.2%, 30.9%, 29.5%, and 75.6% were the F1 scores, respectively. Moreover, by combining metadata through the best performing XGB classifier, with MELA-CNN the performance was improved up to 80% also being highlighted the importance of metadata. Following this to mitigate class imbalance various oversampling techniques were applied to which K-means SMOTE had the best performance improving the model to 86.1% F1 score. Finally, to check the validity of the previously assessed strategies statistical significance was confirmed of a p-value <5%. While this study provides valuable insights, it does not provide information with regard to recall. Shoieb et al. (2016)'s approach on the other hand highlights the importance of diagnosing the disease correctly thus dealing with the misclassification of positive cancer patients.

Results from an investigation by Murugan et al. (2021) indicated similar successful techniques. From the ISIC dataset (1000 data points) four modeling techniques attempted to classify seborrhoea keratosis, melanoma, basal cell carcinoma, and benign lesion. Through image pre-processing, image segmentation, and feature extraction (Gray Level Co-Occurrence Matrix GLCM) the probabilistic neural network model had a 63.2% accuracy, RF 76.3%, SVM 87.8%, and a stacked SVM + RF 89.3%. Furthermore, with regards to recall 61.1%, 74.2%, 85.7%, and 88.56% respectively. Naturally, according to the authors, GLCM is the best performer primarily attributed to the advancements made in border detection and the selection and implementation of a set of distinguishing and functional features.

In their study, Premaladha and Ravichandran (2016) examined the effectiveness of various classification techniques for this same issue. Basing themselves on the public datasets, 992 dermoscopy images were collected with a goal of a binary diagnosis of malignant to benign melanoma. Furthermore using Contrast Limited Adaptive Histogram Equalization (CLAHE) and median filtering as a preprocessing technique and normalized Otsu's algorithm contributed to focusing on the area with skin damage followed by feature extraction and classification. The models used and resulting accuracies and sensitivities were deep learning-based neural networks (DLNN): 92.89% 94.83%, Hybrid AdaBoost: 91.73% 94.08%, adaptive neuro-fuzzy inference system: 90.44% 92.68%, SVM: 90.39% 92.65%, ANN: 90.12% 92.65%, Gentle AdaBoost: 89.91% 92.32%, Modest AdaBoost:

86.84% 89.5%, and Real AdaBoost: 86.52% 89.16% respectively, which, in contrast with Chang et al. (2022)'s best model, is outperformed by the DLNN. A limitation of this research is the limited amount of imagery used for model training.

The findings of Al-Hammouri et al. (2021) showed success in regard to melanoma classification. Using the MED-Node dataset and the methods of pre-processing, segmentation, and feature extraction (GLCM) the data was input into the classification methods: Extreme learning machine (ELM), KNN, RF, SVM. ELM is a type of algorithm that uses a single hidden layer neural network with randomly generated weights for fast and efficient training of large-scale datasets. The associated respective accuracies with eleven features were 97%, 94%, 97%, and 94% in comparison with the lower results of 91%, 95.5%, and 94% when using five features, respectively. The author compares their results in KNN, RF, and SVM with other papers concluding these models outperformed the respective referenced papers. Furthermore, concluding that the ELM is the model with the highest performance and was significantly faster with regards to running time when compared with RF, about 120 times faster. Although the study provides a good base in the context of predictive accuracy, it is not highlighted the importance of recall in melanoma detection.

Following Shoieb et al. (2016), perspective is given on the utilization of a newly created algorithm using various datasets. As base techniques, image enhancement, and preprocessing are used, while the latter is to remove background noise in the skin, region of interest segmentation is also a core step: Hue, Saturation, and Value (HSV) to convert images into color space. Furthermore, using texture features, since in lesion classification it is not appropriate to be dependent on merely color features, and finally region clustering using K-means. The technique that follows is a feature extraction using the common convolutional layers extractor, pooling, and non-linear layers. Finally, once all features are extracted they are classified through an SVM model.

Using the DermIS, DermQuest, and DermNet datasets this same model achieves different results. In the first dataset, considering it is a binary classification, it identifies 87.5% of patients with melanoma correctly and has an overall accuracy of 93.5%, while in the second dataset, it identifies 94.12% of patients with melanoma accurately, also having the aforementioned value as overall accuracy. In the final dataset, although identifying four different types of skin diseases, melanoma has the highest accuracy of 98%

and correctly recognizes melanoma in 88.89% of the cases. In like manner, the authors compare to previous research concluding their model outperforms significantly previous cutting-edge computerized systems for skin diagnosis by approximately 11% in accuracy and recall. Concluding this paper with a very interesting insight into prospects of future research including the ambition of expanding the capability of this type of diagnosis to mobile support.

In consonance with Pacheco and Krohling (2021), a very interesting approach called the MetaBlock has been suggested for data classification, which utilizes metadata (i.e. patient information) to improve the most impactful features that are extracted from images during the classification process. Naturally, this algorithm enhances model classification by processing metadata (Chang et al., 2022). Basing themselves on the ISIC 2019 and the PAD-UFES-20 dataset with three clinical features, using the pre-trained architectures of EfficientNet-B4, DenseNet-121, MobileNet-v2, ResNet-50, and VGG-13. After applying pre-processing techniques, this approach was compared among models that did not have metadata, the ones that were created with the concatenation approach, MetaBlock, MetaNet, and within all transfer learning architectures. In the ISIC data, it is noted that the additional clinical features provide support to the images, as reflected in the highest accuracy values (74,8%), using the EfficientNet. When assessing the PAD-UFES, results show that including metadata improves the models' performance significantly, with an improvement of at least 8.2% in balanced accuracy when using the MetaBlock approach, further highlighting how crucial the inclusion of metadata is. Furthermore, the Metadata Processing Block also has a statistically significant value on both datasets. The metadata impact is twice the percentual value as Chang et al. (2022) where there is a 4% increase, in comparison with Pacheco and Krohling (2021) where it was an 8.2% increase in model quality. Pacheco does not include an analysis of recall and focuses on accuracy, AUC, and balanced accuracy which provides some limitations with regard to medicinal context, due to the lack of focus on false negatives.

Applying machine learning methods to binary and three-class classification in the context of thickness is a research that was presented by Saez et al. (2016). The database had 250 dermoscopic images comprised of the three classes categorized by the size of melanoma: < 0.76 mm, 0.76-1.5 mm, > 1.5 mm. With regards to methodology, feature extraction and texture features (GLCM) are the mainly mentioned image pre-processing techniques. These images would then be input into several models, highlighting the highest performer, the Logistic regression using Initial variables and Product Units (LIPU). Although model performances were similar, the LIPU model demonstrated superior model quality compared to all other models with a 77.6% accuracy rate in both concerns. When compared with a previous study discussed in the literature review it was 8.9% higher. The authors claimed that this difference would not be of fair comparison as the dataset used was private and smaller, hence could have data bias. This research gives a good bridge to Sacchetto et al. (2018)'s medical research and consolidates the need to focus on thin melanomas, as they are more prominent.

3 Resources

3.1 Research Goal

The goal of this research is to explore the effectiveness of using deep learning techniques to detect melanoma using skin lesion imagery, contextualizing the importance of the recall metric and providing a large merged dataset as a basis for algorithmic modeling. The following research questions are being considered to draw conclusions on this topic:

- Are Deep Learning algorithms effective at predicting outcomes based on human images?
- Can CNNs achieve high model quality when applied to a heterogeneous set of image datasets with varying imaging characteristics, such as different perspectives and cameras used for acquisition?

3.2 Data

3.2.1 Dataset Description

For the present study, official data was gathered from a total of 11 sources of public data and merged into one final dataset containing 68014 images, thus introducing the MAR-MELA database. The description of every individual dataset is presented as follows:

1. International Skin Imaging Collaboration: ISIC is a large association that provides a public dataset available to anyone that would like to conduct dermatology research. A merge of the 2017, 2019, and 2020 datasets was conducted resulting in 57988 total extracted images for this research.

2. $\mathbf{PH^2}$: Consists of 200 dermoscopic images that originate from the Dermatology Service of Hospital Pedro Hispano, Matosinhos, Portugal. The dataset is subcategorized into 3 categories: common nevi (80), atypical nevi (80), and melanoma (40), although for the purpose of this research, only melanoma images are considered.

3. SKINL2: The light fields of skin lesions were captured at Centro Hospitalar de Leiria, Portugal. Dermatoscopic images of each injury with a resolution of 1920

× 1080 pixels are provided with an 8-bit depth RGB. For this research, only 53 melanoma images were extracted from this dataset out of the 376 images.

4. 7-Point Criteria Evaluation: This dataset was specifically designed for assessing the accuracy of computerized image-based prediction of the 7-point checklist for identifying malignant skin lesions. This database has 1011 images, where 252 are melanoma while the remaining are differing regular moles and other diseases. Only melanoma and benign injuries were considered for this research.

5. DermNet: The data consists of 19,559 images of 23 types of skin diseases, each image has 3 channels (RGB) and varying resolutions. Only 1295 images are effectively used for the purpose of this research, out of which 188 images are melanoma, while the remaining display 1043 benign injuries.

6. Waterloo Dataset: Developed by WATERLOO UNIVERSITY, data is comprised of a subset of dermquest and dermIS datasets, although only dermquest is considered. The latter consists of 137 images where 76 are melanoma positive and 61 are not melanoma, of which all are considered.

7. DermIS: The DermIS dataset is a comprehensive dermatological image collection based on the Dermatology Online Atlas and the Pediatric Dermatology Online Atlas, including data on nearly all skin diseases. A total of 1000 images were extracted, comprising 500 images of melanoma and the remaining images portraying benign injuries.

8. MED-NODE: Consists of 70 melanoma and 100 nevus images from the digital image archive of the Department of Dermatology at the University Hospital Groningen. All 170 pictures are considered for this research.

9. FitzPatrick17k: The Fitzpatrick 17k dataset consists of 16,577 clinical images with skin condition labels and skin type labels based on the Fitzpatrick scoring system. For the intent of this investigation, a total of 573 images featuring melanoma and 2457 benign skin lesions were taken into account.

10. SD-260: This image bank is an extension of the SD-198 image bank, containing 260 classes of diseases and 20600 images in total. For the purpose of this analysis, 373 melanoma and 5812 benign injuries are considered.

3.2.2 Dataset Approach

The MAR-MELA dataset is introduced in this research and presented as the merge of the aforementioned datasets. Due to the nature of the attributes of the individual datasets, manual categorization was created with the purpose of efficient data usage. Thus, two key variables of the dataset are presented:

- *MEL*: Melanoma is detected (MEL = 1) ;
- *Benign*: Benign disease is detected (Benign = 1);

It is important to understand what is the underlying decision-making factors that define the two target columns. The data gathered comes organized in different ways and provides different information, with some datasets having over 100 categories of skin diseases. Hence, it is very important to highlight that the categorization was manually decided. Considering the purpose of this study is to diagnose melanoma at an early phase, any stage of melanoma imagery is considered for this column. In order to classify for the *Benign* category the following categories were considered: any nevus, any benign injuries, dermatofibroma, seborrheic keratosis, miscellaneous, acne, rosacea, dermatitis, eczema.

Class	Initial Count	Final Count
Melanoma	7226	7226
Benign	60788	8000
Total	68014	15226

Table 1: Distribution of Categories of MAR-MELA.

Looking at the main statistics of the two target variables, there is a clear indication of a class imbalance. This issue must be addressed as data is skewed towards one class, in this case, the non-melanoma class, making it difficult for the model to learn how to detect the melanoma class. Furthermore, it is not justified to extensively increase training times for a category that is not part of the most relevant objective of this research. Thus, downsampling techniques were applied to the Benign class reducing the number of images to 8000.

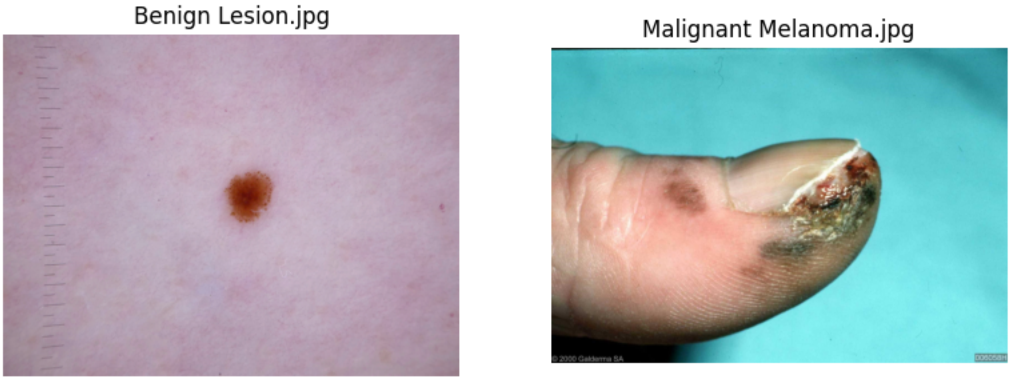

Figure 1: Example of both categories originating from different sources.

4 Methodologies

Due to the nature of imaging processing, and since we are dealing with heavy data loads a cloud computing solution was used for faster and more efficient processing. For the purpose of this research, several Azure servers were set up. The presented algorithm was executed on an Azure-based virtual machine running on Windows 10 operating system. The compute instance selected was Standard_NC6, equipped with an NVIDIA Tesla K80 GPU, which has a virtual machine size of 6 cores, 56 GB of RAM, and 380 GB of disk space.

4.1 Metrics

4.1.1 Description

Prior to immersing ourselves in the proposed algorithm aimed at achieving efficient detection of melanoma, it is of paramount importance to employ appropriate metrics that can effectively and accurately assess the model's performance. Consequently, an overview of the used metrics will be introduced and briefly discussed in this chapter. Contextualizing for this use case, true positives and true negatives will translate to the number of cases where melanoma was correctly identified in contrast to the number of times non-melanoma was correctly identified. Furthermore, false positives will indicate the number of benign injuries that are wrongly identified as melanoma, and false negatives the number of melanoma cases incorrectly identified as non-melanoma.

Accuracy is the conventional metric used for evaluating model performance, as it reflects the model's ability to correctly predict all classes, expressed as a percentage. Mathematically, it can be represented as follows:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} = \frac{True\ Predicted\ Cases}{All\ Cases}$$

The truly predicted melanoma cases summed with the truly predicted non-melanoma cases divided by all the cases will correspond to the accuracy metric.

Recall (Sensitivity) measures how well a model can detect positive instances, if this value is high it can correctly detect melanoma well. It is expressed below:

$$Recall = \frac{TP}{TP + FN} = \frac{True\ Predicted\ Melanoma\ Cases}{All\ Melanoma\ Cases}$$

Specificity evaluates the algorithm's ability to detect negative instances. It is calculated as follows:

$$Specificity = \frac{TN}{TN + FP} = \frac{True\ Predicted\ Non - Melanoma\ Cases}{All\ Non - Melanoma\ Cases}$$

Precision is the proportion of positive predictions that are actually true positives. It is calculated as follows:

$$Precision = \frac{TP}{TP + FP} = \frac{True\ Predicted\ Melanoma\ Cases}{All\ Predicted\ Melanoma\ Cases}$$

F$_\beta$ Score is the weighed harmonic mean of precision and recall and is calculated as follows:

$$F_\beta = (1 + \beta^2) \cdot \frac{Precision \cdot Recall}{(\beta^2 \cdot Precision) + Recall}$$

- $\beta > 1$: More emphasis on Recall
- $\beta = 1$: Equal emphasis on Precision and Recall (F1 Score)
- $0 \leq \beta < 1$: More emphasis on Precision

To better contextualize the formula β, interpretation must be understood. As β increases and approaches $+\infty$, the value of the formula gets closer to the real value of Recall as is visible in the limit below:

$$\lim_{\beta \to +\infty} F_\beta = \lim_{\beta \to +\infty} (1 + \beta^2) \cdot \frac{Precision \cdot Recall}{(\beta^2 \cdot Precision) + Recall} = Recall$$

As β decreases and approaches zero, the value of the formula gets closer to the real value of Precision as is visible in the limit below:

$$\lim_{\beta \to 0} F_\beta = \lim_{\beta \to 0} (1 + \beta^2) \cdot \frac{Precision \cdot Recall}{(\beta^2 \cdot Precision) + Recall} = Precision$$

4.1.2 Discussion

It is important to highlight that although accuracy is a metric that allows for a simplistic overview of model performance, it is very limited in this case. Moreover, it is relevant to highlight the importance of FP and FN not having an identical impact on society. The reason behind this dissimilarity stems from the fact that if a model classifies a patient as FN, melanoma will spread unnoticed and could cause serious health complications. Following this assessment, it is concluded that accuracy is not the appropriate measure to prioritize models as it weighs these two classes the same.

Specificity is also important as a low value in this metric implies many individuals are misclassified as having melanoma while truly not having it. Although this could generate anxiety, unwarranted biopsies, and further testing, in this particular situation, recall is more crucial than specificity as it involves saving lives.

Henceforth, recall would be the most appropriate performance assessment as it addresses the aforementioned issue, minimizing false negative cases. Consequently, a high recall could potentially prevent melanoma from going unnoticed mitigating a potentially delayed treatment and poor health outcomes. Although the initial intuition is to maximize recall to fit the goal of correctly classifying as many melanoma cases as possible, this is not the most correct approach. This is due to the increased values of false positives that would originate from the increased number of melanoma classifications (i.e. achieving 90% recall would not be as valuable if our precision is at 50%, as true benign would be misclassified as melanoma very often). The F_β Score, however, takes into account both these metrics and allows for a customizable weighing on precision and recall. More specifically, we can adjust a weighed balance between FP and FN, prioritizing FN, while not disregarding FP. This will translate to achieving more realistic results, as it acknowledges the medical consequences and approximates to the desired trade-off between these two metrics. The proposed methodology will maximize the F_β Score in the MAR-MELA-CNN, for this purpose, β will equate to two as it provides a reasonable weighing towards recall without going to an extreme biased extent.

4.2 MAR-MELA-CNN

The proposed algorithm presented in this chapter was developed through a meticulous process of trial and error and was chosen based on the highest F_β Score. Several modeling techniques were performed to discover the highest-performing model. Following a thorough process of hyper parameterization, MAR-MELA-CNN was ultimately derived as the highest-performing model. The method of trial and error included experimenting with multiple convolution layers, data augmentation values, varying optimizer parameters, number of epochs, callbacks, and image sizes

The experimentation process comprised varying filter sizes, including 64, 128, 256, 512, 1024, 2048, and 4096 which inherently allowed to unmask the highest-performing model through differing levels of complexity. This was also reflected when different image-altering techniques were explored. This included changing the contrast and brightness, applying CLAHE (Contrast Limited Adaptive Histogram Equalization), and using HSV (Hue, Saturation, Value) filters with the ultimate goal of region of interest segmentation, as highlighted in literature. Furthermore, multiple image sizes were tested, including 150x150, 224x224, 299x299, 331x331, and 512x512 as a higher image size could translate to more areas that the CNN could learn complex features from.

Moreover, various learning rates were deployed with the objective of seeing which is the appropriate scaled measure of which the model gradually learns the images from $1 \cdot e^{-3}$, $1 \cdot e^{-4}$, $1 \cdot e^{-5}$, $2 \cdot e^{-5}$. This process helped maintain a balance between how fast the model converges to the optimal point with the actual model performance. This process introduced the ReduceLROnPlateau callback as it was a way to dynamically adjust the learning rate based on the model's performance, followed by ModelCheckpoint also being added to guarantee the optimal model is saved at the highest F_β Score. To conclude, various values of epochs were also utilized varying from the basic 10 to 500 epochs, which extensively fluctuated the training duration time without providing a significant performance increase.

4.2.1 Validation Approach

The technique used in this research is based on the utilization of a validation set to evaluate the performance of the algorithm. This was preferred over the traditional k-fold cross-validation due to the nature of deep learning models and their resourceintensive functionalities, implying inefficiency in terms of training times. The data split distribution of the 15226 images is comprised of 80% being allocated for training, 10% for validation, and the remaining 10% for testing. A validation method is crucial for an efficient algorithm that eschews overfitting due to constant validation after each epoch, allowing for performance monitoring.

4.2.2 Data Preparation

Data augmentation plays a crucial role in deep learning applications as a preliminary step, aiming to enhance the diversity of the training data, improve model generalization, and further mitigate overfitting. The images fed into the ImageDataGenerator are input with the size of 224 × 224 pixels, and are validated on the test and validation datasets where only rescaling is performed.

The following parameters are used for the augmentation in the training generator:

- **rescale**: Rescales the images by a factor of 1/255, which normalizes the pixel values to the range [0, 1].
- **rotation_range**: Randomly rotates images within a range of 5 degrees.
- **zoom_range**: Randomly zooms in or out on images within a range of 0.1.
- **channel_shift_range**: Randomly shifts the color channels of images within a range of 0.001.

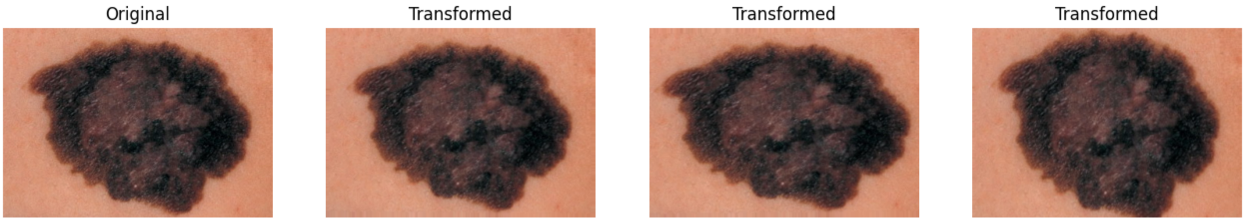

Figure 2: Example of an augmented Melanoma injury.

4.2.3 Architecture

The base of this algorithm is an ensemble model built using several pre-trained CNNs that were estimated on the ImageNet dataset (Baccouche et al., 2020). The concept of using numerous models as a base for image recognition stems from the fact that we can leverage the feature extraction capabilities that these models already have developed and apply them to this specific melanoma use case. The MAR-MELA-CNN architecture utilizes six pre-trained CNN models to form an ensemble:

- **Xception:** A deep CNN architecture based on depthwise separable convolutions, proposed by Chollet (2016).
- **VGG16:** A simple 16-layer deep CNN developed by the Visual Geometry Group at the University of Oxford, introduced by Simonyan and Zisserman (2014).
- **ResNet50:** A 50-layer deep residual network proposed by He et al. (2015), which addresses the vanishing gradient problem and allows for deeper networks by introducing residual connections.
- **NASNetMobile:** A smaller version of the NASNet architecture, designed using Neural Architecture Search (Zoph et al., 2017).
- **MobileNetV2:** A lightweight CNN architecture proposed by Sandler et al. (2018).
- **InceptionV3:** A deep CNN architecture utilizing parallel convolutional layers with different filter sizes proposed by Szegedy et al. (2015).

The ensemble model is constructed as a fusion ensemble, where each pre-trained model processes the input independently and their outputs are concatenated for the final decision-making. The ensemble iterates through each pre-trained model, applying a global average pooling layer that averages the values of each feature map thus translating to the task of dimensionality reduction. Each pooled output is then passed through a 256 unit fully connected dense layer using the ReLU activation function, subsequent to a 30% dropout layer aimed at mitigating overfitting of the model as it randomly nullifies input units. A 128 neuron dense layer is applied to further refine the learned feature representations, to which is passed on to an operation that collects all the predictions made by these models and concatenates the outputs. A final fully connected dense layer is applied to the concatenated layer using the Sigmoid activation function to be able to extract the binary classification outcome probabilities, inherently making the final prediction.

4.2.4 Model Compilation and Training

The ensemble model is compiled using the Adam optimizer with a learning rate of 2×10^{-5}, $\beta_1 = 0.9$, $\beta_2 = 0.999$, and $\epsilon = 1 \times 10^{-7}$, and AMSGrad being set to True. Adam optimizer was adopted as it has been evidently used as the conventional optimization algorithm due to its ease of use and efficiency. Furthermore, the learning rate is set to a low value to ensure weights are updated without making abrupt changes that could harm the model's performance, thus ensuring a more robust, low-deviated learning curve. Moreover, it is relevant that AMSGrad was set to true as to guarantee that this optimizer has a significantly better convergence to mitigate the optimization issue this algorithm often runs into. In order to successfully compile this model, a binary cross-entropy loss function is used for training as the task at hand is a binary classification.

A checkpoint callback is used during training to save the model with the highest validation F_β score, this is relevant to avoid potential overfitting thus saving the best model with the highest F_β Score. The ensemble model is then finally trained for 10 epochs where after each iteration of the model studying the training set, weights are readjusted and performance is measured on the validation data.

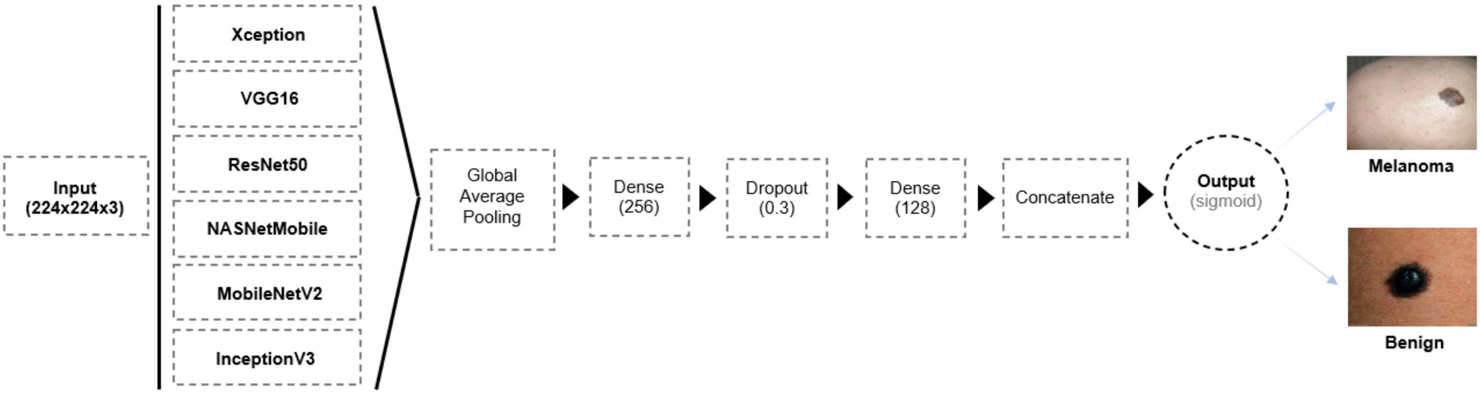

Figure 3: Overview of MAR-MELA-CNN architecture.

5 Findings

5.1 Models Assessment

Following the established methodology, it is relevant to assess the results achieved by each model in the trial and error phase. Additionally, a comprehensive analysis of the MAR-MELA-CNN will be studied. This in-depth comparison will inherently allow the identification of the most suitable method for the given research problem and draw meaningful conclusions. Considering the vast amount of CNN models with varying architectures that were attempted, only the four best-performing models with the highest F_β Score are pondered for this analysis.

The Basic Model is sequentially constructed based on a Conv2D layer with 64 filters, followed by another Conv2D layer with the same number of filters which thereafter is applied a MaxPooling2D layer. The architecture that follows is essentially the same pattern, with the exception of the filter size that will double until reaching 1024 filters, after which it is then flattened and two dense layers are applied with dropout and batch normalization.

The Basic GigaModel is an enhanced version of the aforementioned algorithm, basing itself on a larger Conv2D layer with 128 filters, and follows again the same pattern until reaching 2048 filters. Moreover, the following difference is the dense layers are eight times larger than the Basic Model's dense layers and the number of epochs increases from 10 to 50.

The Heavy Layered Model, however, has a considerably more complex and deeper architecture as it employs more data augmentations, uses 512x512 pixels as image size, and further increases the layerization of the model. The model initially implements a double-layered Conv2D with 32 filters which is followed by a BatchNormalization and MaxPooling2D layer. A triple-layered Conv2D with 64 filters ensues, of which the filter size is doubled until reaching 512 filters, including the same aforementioned anti-overfitting methods after each convolution layer. After the final layer, the model once again uses a Flatten layer which is then followed by 512 dense layers with dropout.

Model	Accuracy	Precision	Recall	F_β Score
Basic GigaModel	80%	82%	70%	72%
Heavy Layered Model	81%	78%	77%	77%
Basic Model	80%	73%	79%	78%
MAR-MELA-CNN	85%	82%	86%	**85%**

Table 2: Summary of Model Performance Metrics

The table presented above illustrates the results of the highest-performing models.

The Basic Model and the Basic GigaModel have a relatively similar output in terms of accuracy, both rounding at around 80% performance, however showing discrepancies in recall and precision. GigaModel is clearly outperformed by the Basic Model in the recall metric, as the latter scores 79% compared to the 70%. Nevertheless, GigaModel outperforms Basic in the precision metric by the same percentage points. These two value differences highlight the importance of using F_β Score in an analysis where distinct weights are attributed to recall and precision. GigaModel excels in precision but underperforms in recall compared to the Basic Model, and vice versa, and as F_β score weighs recall higher, this translates into the superior F_β of Basic Model, scoring 6% higher. Furthermore, an interesting insight that can be inferred is that the extension of the model's architecture in filters, neurons, and training time does not directly correlate to the model learning more and outputting better predictions, but rather can lead to overfitting the model.

The Heavy Layered Model achieved an F_β Score of 77% of which all the other metrics share this value on a similar level, except for accuracy which stood out with a higher score of 81%. As many robustness techniques were employed, this facilitated the model to have more generalizability and perform better on unseen data. However, Heavy Layered shows some difficulty in the detection of false negatives as there is a slight discrepancy between accuracy and recall, but overall this model proves its consistency in comparison with the others. Considering this model is also very large like Basic Giga, the 5% improvement in F_β performance could be attributed to the increased image size of 512x512 pixels, which in turn would allow for the model to gather more complex information about the injury.

The MAR-MELA-CNN model showed the best performance among the four models with an accuracy of 85%, precision of 82%, and recall of 86%, highlighting the F_β Score of 85%. Thus, applying to this research, our final deployed model classifies true melanoma and true non-melanoma classes 85% of the time, 82% of all melanoma predictions made by the model are actually melanoma, and out of all true melanoma cases, 86% are true melanoma predictions. The F$_\beta$ Score of 85% reflects the model's ability to prioritize the identification of true melanoma cases (recall) while still maintaining a strong level of precision, thus ensuring MAR-MELA-CNN effectively can classify melanoma.

This superior performance out of all four models can be attributed to the deployed ensembling architecture used, leveraging the strengths of multiple diverse pre-trained transfer learning models in the melanoma classification use case.

Below a short visualization is provided comparing all the models in the relevant proposed metrics.

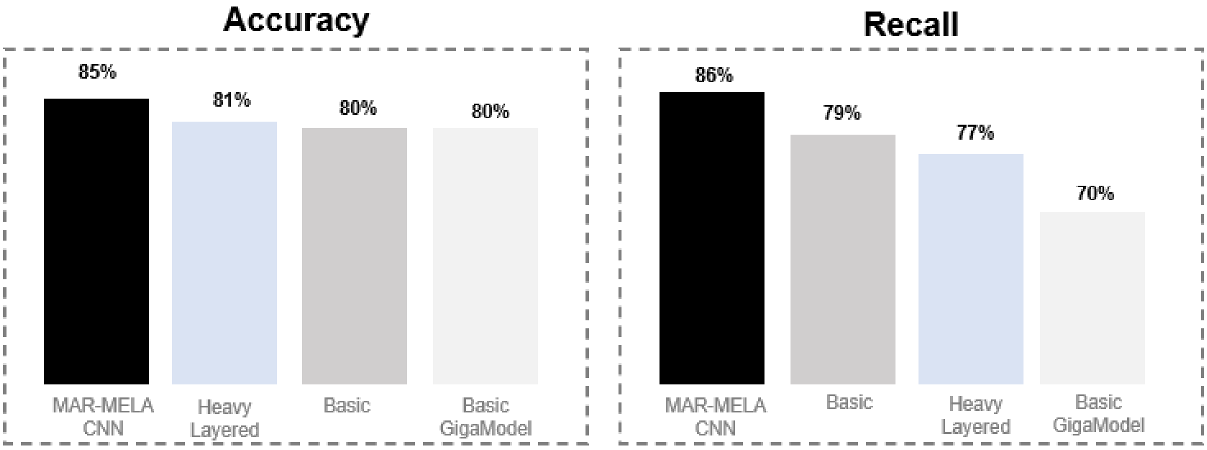

Figure 4: Visual comparison of the performance metrics

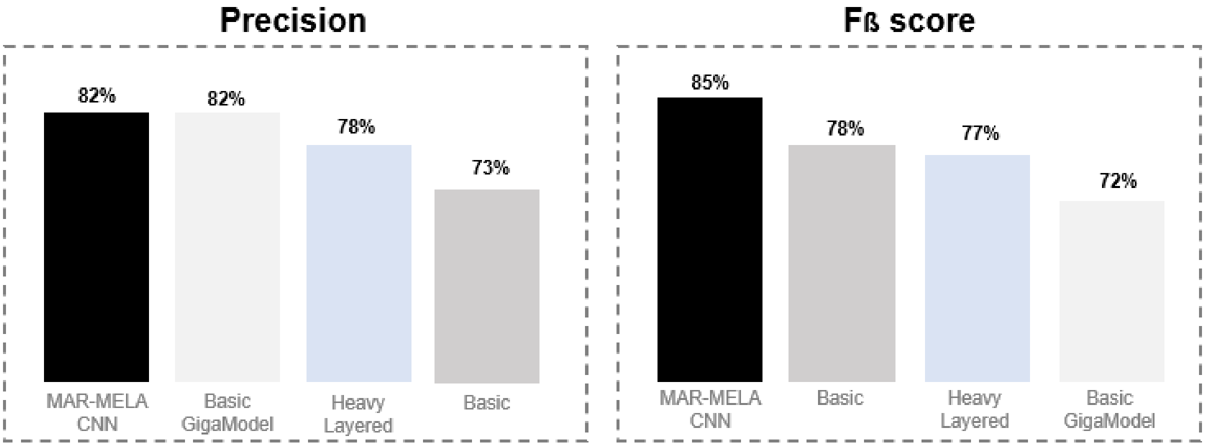

Figure 5: Visual comparison of the performance metrics

5.2 MAR-MELA-CNN's results

To further analyze the modeling outputs it is below presented a detailed analysis of the results obtained from the MAR-MELA-CNN.

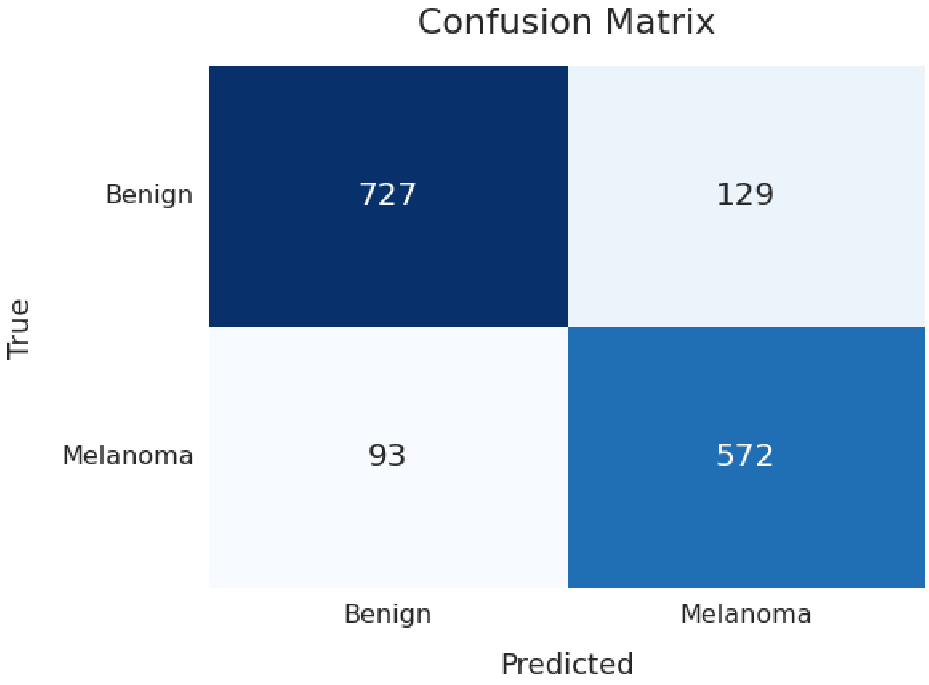

Figure 6: Confusion Matrix of MAR-MELA-CNN

Following the presented confusion matrix the independent interpretation of each component will be elucidated. The model correctly identified 727 instances as non-melanoma cases (TN) and accurately classified 572 cases as melanoma (TP). Naturally, the latter is the most significant value as early detection and treatment of this disease are crucial for the survival rate of patients, as mentioned in literature. Furthermore, this result translates the model's ability and effectiveness in identifying melanoma cases, which is the objective of this research.

However, the model failed to identify 129 cases of benign injuries, misclassifying them as melanoma (FP), although it is always always a modeling objective to minimize false returns, in this case, it is not as important as minimizing false negatives. Moreover, in a real-world scenario, 129 people would have been falsely alarmed to get biopsies to evidently test for melanoma, when in reality they had a mere benign mole. Furthermore, the algorithm also misclassified 93 true melanoma patients (FN), which raises serious concerns. Applying this situation to a real-life scenario the 93 people that effectively relied on this model, would get wrongly induced into having a mere benign injury when in reality they would have cancer. The implications of this are very serious as they will not be alerted and will miss out on early treatment, leading to lower survival rates and a worsened prognosis. Thus, in this modeling application around $\frac{93}{665}$= 13.9% of true melanoma patients will not be alerted of having cancer implying worse health outcomes for these individuals.

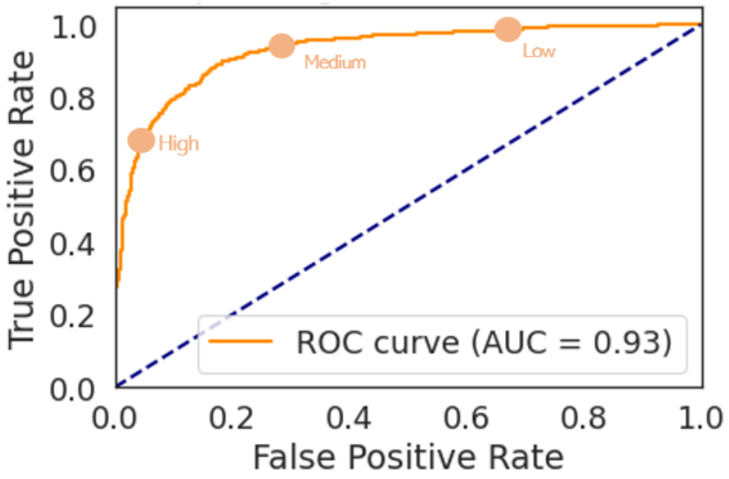

Figure 7: ROC_AUC curve of MAR-MELA-CNN

The graph above represents the Receiver Operating Characteristic curve and covers 93% Area Under the Curve, representing 93% of the total area of the rectangle. The curve assesses the comparison between the effectiveness of actual positive samples being correctly classified as positive (True Positive Rate or Recall) and the proportion of actual negative samples that are incorrectly classified as positive (False positive rate or 1 - Specificity). Following this, the model is very effective at discriminating between benign and melanoma classes, more specifically MAR-MELA-CNN has a high probability of accurately identifying melanoma (TP) while minimizing the misclassification of benign lesions as melanoma (FP).

To further understand the results, an extensive threshold comparison will be executed to comprehend the strengths and flaws at different probabilities. These thresholds are designated as High, Medium, and Low on the ROC curve plot, which are assigned to be around 80%, 50%, and 20%, respectively. These values will indicate the minimum probability of the model classifying an image as melanoma. To exemplify, for the High probability threshold, MAR-MELA-CNN will only classify an instance as melanoma if the image has a minimum of 80% predicted probability.

When considering a High threshold, the model is more conservative when making a positive prediction, thus reducing false positives, this inherently implies this threshold makes the model highly specific. Accordingly, this opens the door to misclassifying true melanoma patients as benign, considering the criteria for this identification is much more strict (i.e. if $p > 80\%$ there will be less FP implying specificity increases, albeit at the cost of reduced recall due to a reduction in melanoma classifications). In a real-world scenario, implications of this would mean that more melanoma cases would get unnoticed because of the rigorous criteria being employed, thus, not appropriating to this use case.

Upon examining the Medium threshold, it is visible that recall will not be as compromised as in the High threshold. This fact stems from the decreased difficulty of classifying melanoma that arises when the probability to classify is decreased, thus, implying this threshold is more relevant to the use case of classifying melanoma.

In the Low threshold setting the model has a high recall which directly implies it will very accurately identify melanoma injuries. However, this is simply not a realistic approach. Having a lower probability threshold will imply many cases will be identified as melanoma which will drive the false positive values up substantially, hence misclassifying healthy individuals as cancerous.

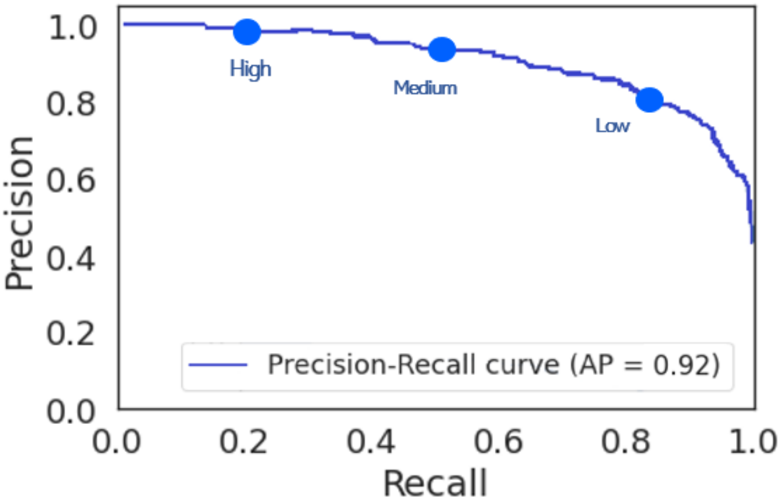

Figure 8: Precision-Recall curve of MAR-MELA-CNN

The plot above displays a graphical representation of precision and recall at different thresholds. The average precision (AP) statistic gives more importance to the precision values achieved at higher recall levels, or in other words, it represents a model's overall performance in identifying melanoma instances.

Average precision is calculated by averaging the precision values obtained at each recall level after being estimated, the formula for average precision is:

$$AP = \sum_{n} (R_n - R_{n-1}) \cdot P_n$$

The interpretation is that R_n and R_{n-1} are the recall values at points n and $n-1$ while P_n is the precision at point n.

Considering AP takes the value of 92%, it means that the model has a high level of precision (P_n) across different recall (R_n) levels, thus identifying melanoma very well. Furthermore, it is essential to understand that the model will reach high values of precision without excessively hindering recall, correctly identifying a large number of melanoma cases while keeping the false positive rate relatively low. Following the previous logic, we will further assess different levels of probability thresholds to understand the strengths and limitations of the algorithm. Moreover, maintaining the aforementioned threshold percentages of 80%, 50%, and 20% for the High, Medium, and Low thresholds.

A High probability threshold will mean that if the model predicts melanoma it is very probable to be a true classification. This is a very appealing result although this also implies that many melanoma cases with not as high a probability may go unnoticed as the criteria for identification is again stricter, inherently reducing recall and leading a very high precision (since FP are very low but FN are very high). Considering we are aiming to minimize exactly that issue, a conservative approach is not appropriate for melanoma detection.

A more balanced approach is the Medium classification level as it provides an even trade-off between both metrics, minimizing both FN and FP.

When assessing the Low probability level threshold there will be more melanoma classifications as the decision criteria will be more flexible, thus increasing the amount of TP's. Although this has the logical limitation of also increasing the false positives as precision goes down, thus once again not providing a realistic sense of the use case. However, this approach is still preferred to the higher threshold as although we have an increase in benign misclassifications, melanoma injuries will less likely go unnoticed.

Regardless, a key takeaway from both plots is the complexity that derives from establishing a universally correct threshold for melanoma classification, as all we can infer is the fact that recall matters more than precision. However, a good threshold to consider would be a probability within Medium to Low as it infers our valuation to recall without totally disregarding precision.

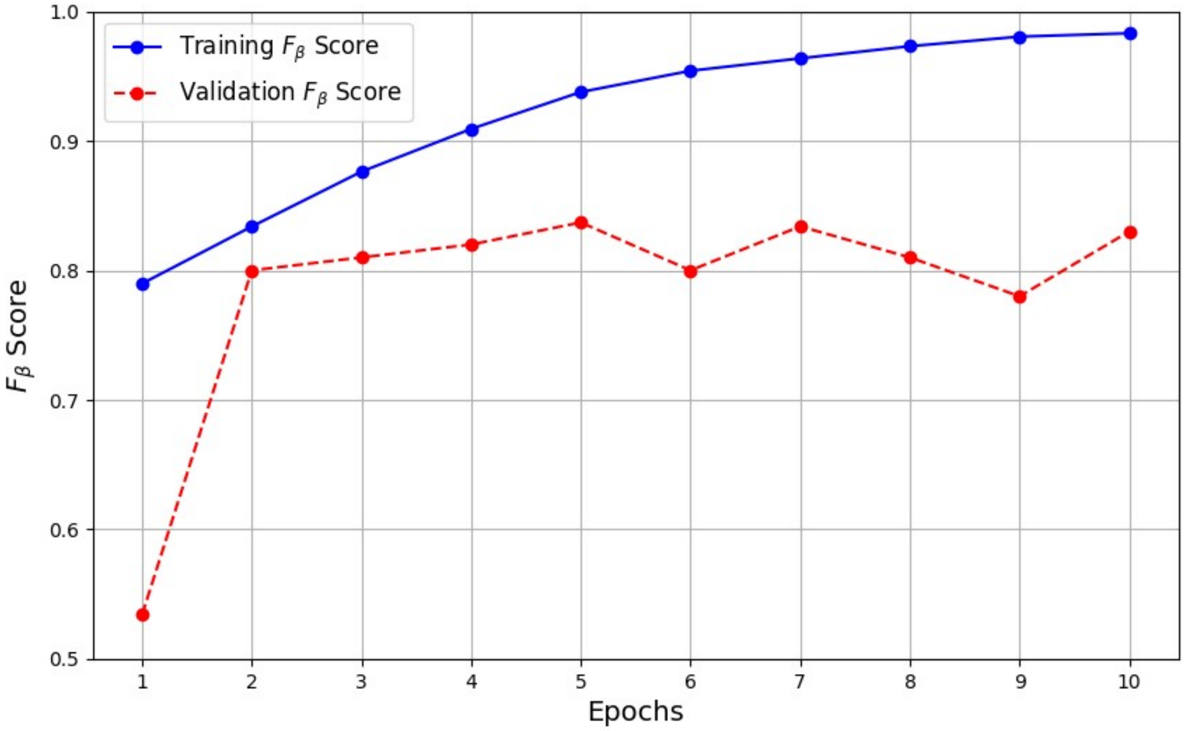

Figure 9: Training vs Validation F$_\beta$ score

The graph above displays the training and validation F$_\beta$ scores of MAR-MELA-CNN over ten epochs. The training phase displays a clear consistent increase going from 78.9% until peaking in the tenth epoch at an F$_\beta$ of 98% indicating the model is learning the training data at a constant rate. However, in contrast with the validation epoch history, the scores increase from 53,4% in the initial epoch until peaking in the third epoch at 83.7% and thereafter suffers some oscillations. Considering there is no discernible upward trend in the validation scores, along with the visible gap between training and validation scores, there is a clear indication of overfitting in the model. A method that could be implemented to mitigate this, would be the inclusion of more dropout/regularization methods.

6 Results Discussion

To provide a comprehensive understanding, discussion will be provided and a results comparison will be assessed with past literature. Before diving into results, it is important to mention that F_β score is a specific metric that has not yet been thoroughly explored in literature, thus limiting an exact comparison.

The transfer learning model-based research presented by Pacheco and Krohling (2021) provides us with slightly lower results in the balanced accuracy metric, reaching 83% on the final modeling of the proposed PAD-UFES-20 dataset. However, when comparing the ROC curve, Pacheco's model outperforms the proposed model by 2%, reaching 95%. Furthermore, to contrast with the techniques developed in this research, Pacheco's strategy of including patient data in the modeling phase provides a successful increase in the performance of the algorithm, which is not done in this study. This provides an impactful limitation as it could significantly increase MAR-MELA-CNN's performance had there been an option to include metadata. However, Pacheco fails to identify the importance of choosing the right metric for medical diagnosis, providing only the main focus on balanced accuracy and AUC without highlighting recall's impact. This limitation translates to an unrealistic assessment of the model as it fails to consider the impact of letting melanoma go unnoticed. Another significant difference is the way the modeling was conducted. Pacheco uses five transfer learning models and individually trains and assesses them, while MAR-MELA-CNN, is based on the ensembling technique which inherently impacts the results differently.

In the context of data augmentation techniques, Shoieb et al. (2016) uses the HSV method for the region of interest segmentation inherently providing good results. This was not the case when applied to MAR-MELA dataset as modeling performances were very low (18% in recall). Furthermore, it must be highlighted that the interpretation of the medical diagnosis is understood in this research. The application of the SVM in the DermIS and DermNet datasets has around 2% increased performance in the correct identification of patients with melanoma (recall), outperforming MAR-MELA-CNN in all datasets attempted. However, it is also important to state that these datasets had a significantly low number of images (around 250), opening the door to a biased performance, which is not the case in the proposed methodology.

To further compare data altering techniques, Premaladha and Ravichandran (2016) introduces the Contrast Limited Adaptive Histogram Equalization (CLAHE) pre-processing achieving 92.89% in the accuracy field and 94.83% as recall when using a DLNN. This methodology of image pre-processing shows significant enhancements in the model, as it outperforms MAR-MELA-CNN in both metrical categories by over 7%.

The research presented by Vocaturo et al. (2019), however, introduces a pre-trained model (GoogleNet) fine-tuned on a small melanoma dataset. It is significantly outperformed by MAR-MELA-CNN by 22% in the accuracy metric, further iterating the importance of ensembling and highlighting the success of the used architectures in this paper. Additionally, it also briefly introduces a similar study, using artificial neural networks, interestingly, the proposed algorithm outperforms this alternative approach by a margin of 3% which can be attributed to the difference in using pre-trained models.

7 Conclusion

7.1 Achievements

Given the vast successful applications of Convolutional Neural Networks applied to several medical areas, this study aims to develop an accurate CNN model for melanoma diagnosis, while providing a large mixed dataset as a basis and pillaring itself on an Azure infrastructure.

Through the art of trial and error, many models were attempted. The experimentation process comprised varying layers, filter sizes within the Conv2D layer, different image-altering techniques, and thorough hyper parameterization. Thus, through this extensive process, the final model was achieved, introducing the MAR-MELA-CNN: an ensemble of six pre-trained models (Xception, VGG16, ResNet50, NASNetMobile, MobileNetV2, and InceptionV3). The criteria of success for the trial and error was established through the maximization of the F_β score, having β equate to 2 which inherently implied an overall increased emphasis on recall.

Findings indicate a promising diagnostic performance when introduced to a dataset from various sources, as the proposed F_β score achieves 85%. This value is very important as in medical diagnosis reducing false negatives is of utmost importance, furthermore considering F_β prioritizes the importance of recall. Moreover, MAR-MELA-CNN proved very effective at discriminating between benign and melanoma classes through a 93% AUC score, and considering the 92% AP score, it is also accurate at identifying melanoma instances.

Ultimately, this research lays a foundation for potential efficient future modeling applications in the field of melanoma, providing a big dataset as the base. Furthermore, providing a methodology that could be picked up in the future, as cloud solutions have yet to be introduced in the field of detecting melanoma, and highlights the importance of ensembling in CNN modeling for melanoma diagnosis.

7.2 Limitations of this Study and Future Prospects

When attempting to model melanoma using deep learning, a number of limitations are given. This study acknowledges the fact that the odds of getting melanoma depend on how much contact a person has had with Ultra Violet light, as they are correlated. The images used for the research are obtained and do not consider the fact that the person had more or less contact with UV light based on where they live. Furthermore, a deeper study could be conducted using more Azure resources considering the one used for this research was run on a basic Azure student subscription, which does not allow for extensive use of Azure resources. Another note is the fact there was no monetary investment in the construction of the dataset considering it is only based on public data, and a point of its expansion could be purchasing available private datasets for future research.

Directions for future research could entail a different ensembling approach (i.e. weighed averaging), or even stacking the models, putting the ensemble's predictions through a second-level meta-learner model. Moreover, methods could be developed to identify exactly in which images the model classifies a false negative, as it could help to further understand and minimize misclassifications. Somehow attempting to get metadata by reaching out to the founders of each dataset to associate with the pictures of this dataset will further increase the model performance, as mentioned in literature by Chang et al. (2022) and Pacheco and Krohling (2021). On a more realistic stance, an agreement with a hospital could be crucial to the deployment of this model through Azure Kubernetes Service, allowing a hospital to use this modeling technique for real-time detection of Melanoma. This could also be deployed in a more capitalistic sense and perhaps be sold online to a retailer to be available to the average consumer, like a covid test, parallel to the concept previously mentioned by Shoieb et al. (2016). A prototype could be developed where its purpose would be to take a picture, connect to a specific server where the model is being employed and answer back with a diagnosis. An example prototype of this concept can be seen below:

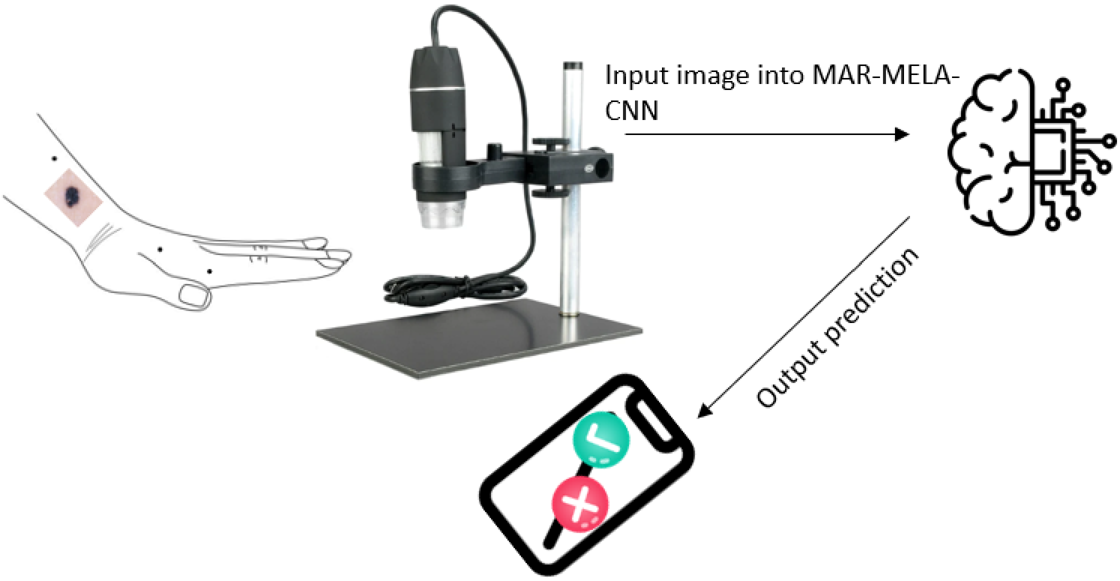

Figure 10: Example of implementable prototype for melanoma detection

References

Al-Hammouri, S., M. Fora, and M. Ibbini (2021, November). Extreme learning machine for melanoma classification. In *2021 IEEE Jordan International Joint Conference on Electrical Engineering and Information Technology (JEEIT)*, pp. 114–119.

American Cancer Society (2019). What is melanoma skin cancer? `https://www.cancer.org/cancer/melanoma-skin-cancer/about/what-is-melanoma.html`.

American Cancer Society (2023). Cancer facts and statistics. `https://www.cancer.org/content/dam/cancer-org/research/cancer-facts-and-statistics`.

Baccouche, A., B. Garcia-Zapirain, C. C. Olea, and A. Elmaghraby (2020, April). Ensemble deep learning models for heart disease classification: A case study from mexico. *Information 11*(4), 207.

Balkenende, L., J. Teuwen, and R. M. Mann (2022, September). Application of deep learning in breast cancer imaging. *Seminars in Nuclear Medicine 52*(5), 584–596.

Bosl, W., A. Tierney, H. Tager-Flusberg, and C. Nelson (2011, February). EEG complexity as a biomarker for autism spectrum disorder risk. *BMC Medicine 9*(1).

Centers for Disease Control and Prevention (2023a). Cancer facts and figures. `https://www.cancer.org/content/dam/cancer-org/research/cancer-facts-and-statistics/annual-cancer-facts-and-figures/2023/2023-cancer-facts-and-figures.pdf`.

Centers for Disease Control and Prevention (2023b). Deaths and mortality. `https://www.cdc.gov/nchs/fastats/deaths.htm`.

Chang, C.-C., Y.-Z. Li, H.-C. Wu, and M.-H. Tseng (2022, July). Melanoma detection using XGB classifier combined with feature extraction and k-means SMOTE techniques. *Diagnostics 12*(7), 1747.

Chekroud, A. M., R. J. Zotti, Z. Shehzad, R. Gueorguieva, M. K. Johnson, M. H. Trivedi, T. D. Cannon, J. H. Krystal, and P. R. Corlett (2016, March). Cross-trial prediction of treatment outcome in depression: a machine learning approach. *The Lancet Psychiatry 3*(3), 243–250.

Chollet, F. (2016). Xception: Deep learning with depthwise separable convolutions. *CoRR abs/1610.02357*.

Codella, N. C. F., D. Gutman, M. E. Celebi, B. Helba, M. A. Marchetti, S. W. Dusza, A. Kalloo, K. Liopyris, N. Mishra, H. Kittler, and A. Halpern (2017). 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).

Combalia, M., N. C. F. Codella, V. Rotemberg, B. Helba, V. Vilaplana, O. Reiter, C. Carrera, A. Barreiro, A. C. Halpern, S. Puig, and J. Malvehy (2019). Bcn20000: Dermoscopic lesions in the wild.

de Faria, S. M. M. et al. (2019). Light field image dataset of skin lesions. In *2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, Berlin, Germany, pp. 3905–3908. IEEE.

DermNet. Dermatology Pictures Skin Diseases.

Diepgen TL, Y. G. e. a. Dermatology online atlas.

Galatzer-Levy, I. R., S. Ma, A. Statnikov, R. Yehuda, and A. Y. Shalev (2017, March). Utilization of machine learning for prediction of post-traumatic stress: a re-examination of cortisol in the prediction and pathways to non-remitting PTSD. *Translational Psychiatry 7*(3), e1070–e1070.

Giotis, I., N. Molders, S. Land, M. Biehl, M. F. Jonkman, and N. Petkov (2015, November). MED-NODE: A computer-assisted melanoma diagnosis system using non-dermoscopic images. *Expert Systems with Applications 42*(19), 6578–6585.

Grady, D. (April 20, 2020). The pandemic's hidden victims: Sick or dying, but not from the virus. `https://www.nytimes.com/2020/04/20/health/treatment-delays-coronavirus.html`.

Groh, M., C. Harris, R. Daneshjou, O. Badri, and A. Koochek (2022). Towards transparency in dermatology image datasets with skin tone annotations by experts, crowds, and an algorithm. *arXiv preprint arXiv:2207.02942*.

Groh, M., C. Harris, L. Soenksen, F. Lau, R. Han, A. Kim, A. Koochek, and O. Badri (2021). Evaluating deep neural networks trained on clinical images in dermatology with the fitzpatrick 17k dataset. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1820–1828.

He, K., X. Zhang, S. Ren, and J. Sun (2015). Deep residual learning for image recognition. *CoRR abs/1512.03385*.

Holtel, M. R. (2022). Skin cancer - melanoma. `https://emedicine.medscape.com/article/846566-overview#a1`.

Kawahara, J., S. Daneshvar, G. Argenziano, and G. Hamarneh (2019). Seven-point checklist and skin lesion classification using multitask multimodal neural nets. *IEEE Journal of Biomedical and Health Informatics 23*(2), 538–546.

Khan, S., N. Islam, Z. Jan, I. U. Din, and J. J. P. C. Rodrigues (2019, July). A novel deep learning based framework for the detection and classification of breast cancer using transfer learning. *Pattern Recognition Letters 125*, 1–6.

Mayo Clinic Staff (2022). Melanoma. `https://www.mayoclinic.org/diseases-conditions/melanoma/symptoms-causes/syc-20374884#:~:text=Melanoma%2C%20the%20most%20serious%20type,in%20your%20nose%20or%20throat`.

Mendonça, T., P. M. Ferreira, J. Marques, A. R. Marcal, and J. Rozeira (2013). Ph^2 - a dermoscopic image database for research and benchmarking. In *2013 35th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pp. 5437–5440. IEEE.

Murugan, A., S. A. H. Nair, A. A. P. Preethi, and K. P. S. Kumar (2021). Diagnosis of skin cancer using machine learning techniques. *Microprocessors and Microsystems 81*, 103727.

Narasimman, P. (2023, March). Data science in healthcare: Applications, roles and benefits. `https://www.knowledgehut.com/blog/data-science/data-science-in-healthcare`.

National Cancer Institute (2022). Cancer staging. https://www.cancer.gov/about-cancer/diagnosis-staging/staging#:~:text=Localized%E2%80%94Cancer%20is%20limited%20to,to%20figure%20out%20the%20stage.

National Cancer Institute NCI (2018, November). Seer*stat databases: November 2018 submission. https://seer.cancer.gov/data-software/documentation/seerstat/nov2018/.

Omar, K. S., P. Mondal, N. S. Khan, M. R. K. Rizvi, and M. N. Islam (2019, February). A machine learning approach to predict autism spectrum disorder. In *2019 International Conference on Electrical, Computer and Communication Engineering (ECCE)*. IEEE.

Pacheco, A. G. C. and R. A. Krohling (2021). An attention-based mechanism to combine images and metadata in deep learning models applied to skin cancer classification. *IEEE Journal of Biomedical and Health Informatics 25*(9), 3554–3563.

Panwar, H., P. Gupta, M. K. Siddiqui, R. Morales-Menendez, and V. Singh (2020, September). Application of deep learning for fast detection of COVID-19 in x-rays using nCOVnet. *Chaos, Solitons &amp Fractals 138*, 109944.

Premaladha, J. and K. S. Ravichandran (2016, February). Novel approaches for diagnosing melanoma skin lesions through supervised and deep learning algorithms. *Journal of Medical Systems 40*(4).

Rotemberg, V., N. Kurtansky, B. Betz-Stablein, L. Caffery, E. Chousakos, N. Codella, M. Combalia, S. Dusza, P. Guitera, D. Gutman, A. Halpern, B. Helba, H. Kittler, K. Kose, S. Langer, K. Lioprys, J. Malvehy, S. Musthaq, J. Nanda, O. Reiter, G. Shih, A. Stratigos, P. Tschandl, J. Weber, and H. P. Soyer (2021, January). A patient-centric dataset of images and metadata for identifying melanomas using clinical context. *Scientific Data 8*(1).

Sacchetto, L., R. Zanetti, H. Comber, C. Bouchardy, D. Brewster, P. Broganelli, M. Chirlaque, D. Coza, J. Galceran, A. Gavin, M. Hackl, A. Katalinic, S. Larønningen, M. Louwman, E. Morgan, T. Robsahm, M. Sanchez, L. Tryggvadóttir, R. Tumino, E. V. Eycken, S. Vernon, V. Zadnik, and S. Rosso (2018, March). Trends in incidence of thick, thin and in situ melanoma in europe. *European Journal of Cancer 92*, 108–118.

Saez, A., J. Sanchez-Montero, A. Gutierrez, Pedro, and C. Hervas-Martinez (2016). Machine learning methods for binary and multiclass classification of melanoma thickness from dermoscopic images. *IEEE Transactions on Medical Imaging 35*(4), 1036–1045.

Sandler, M., A. G. Howard, M. Zhu, A. Zhmoginov, and L. Chen (2018). Inverted residuals and linear bottlenecks: Mobile networks for classification, detection and segmentation. *CoRR abs/1801.04381*.

Seebacher, N. A. (October 2022). Melanoma. `https://dermnetnz.org/topics/melanoma`.

Shoieb, D. A., , S. M. Youssef, and W. M. Aly (2016). Computer-aided model for skin diagnosis using deep learning. *Journal of Image and Graphics*, 122–129.

Simonyan, K. and A. Zisserman (2014). Very deep convolutional networks for large-scale image recognition.

Sondermann, W., L. Zimmer, D. Schadendorf, A. Roesch, J. Klode, and J. Dissemond (2016, July). Initial misdiagnosis of melanoma located on the foot is associated with poorer prognosis. *Medicine 95*(29), e4332.

Sun, X., J. Yang, M. Sun, and K. Wang (2016). A benchmark for automatic visual classification of clinical skin disease images. In *Computer Vision – ECCV 2016*, pp. 206–222. Springer International Publishing.

Szegedy, C., V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna (2015). Rethinking the inception architecture for computer vision. *CoRR abs/1512.00567*.

Tschandl, P., C. Rosendahl, and H. Kittler (2018, August). The HAM10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions. *Scientific Data 5*(1).

Vaishnavi, K., U. N. Kamath, B. A. Rao, and N. V. S. Reddy (2022, January). Predicting mental health illness using machine learning algorithms. *Journal of Physics: Conference Series 2161*(1), 012021.

Vision, I.P. Lab, U. o. W. (2021). Skin cancer database. `https://uwaterloo.ca/vision-image-processing-lab/research-demos/skin-cancer-detection`.

Vocaturo, E., D. Perna, and E. Zumpano (2019). Machine learning techniques for automated melanoma detection. In *2019 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pp. 2310–2317.

Wallert, J., J. Boberg, V. Kaldo, D. Mataix-Cols, O. Flygare, J. J. Crowley, M. Halvorsen, F. B. Abdesslem, M. Boman, E. Andersson, N. H. Isacsson, E. Ivanova, and C. Rück (2022, September). Predicting remission after internet-delivered psychotherapy in patients with depression using machine learning and multi-modal data. *Translational Psychiatry 12*(1).

Wang, X., W. Yang, J. Weinreb, J. Han, Q. Li, X. Kong, Y. Yan, Z. Ke, B. Luo, T. Liu, and L. Wang (2017, November). Searching for prostate cancer by fully automated magnetic resonance imaging classification: deep learning versus non-deep learning. *Scientific Reports 7*(1).

Yang, J., X. Wu, J. Liang, X. Sun, M.-M. Cheng, P. L. Rosin, and L. Wang (2020, August). Self-paced balance learning for clinical skin disease recognition. *IEEE Transactions on Neural Networks and Learning Systems 31*(8), 2832–2846.

Zoph, B., V. Vasudevan, J. Shlens, and Q. V. Le (2017). Learning transferable architectures for scalable image recognition. *CoRR abs/1707.07012*.

A Appendix

A.1 List of Tables

Table 3: Datasets listed in Literature

Name	Number of Images	Reference
DERMQUEST	137	Vision (2021)
DermNet	19500	DermNet (DermNet)
OLE	-	Vocaturo et al. (2019)
ISIC 2019	25331	Tschandl et al. (2018) Codella et al. (2017) Combalia et al. (2019)
ISIC 2020	33126	Rotemberg et al. (2021)
MED-Node	170	Giotis et al. (2015)
DermIS	1000	Diepgen TL (Diepgen TL)
PAD-UFES-20	2298	Pacheco and Krohling (2021)

Table 4: Transfer Learning Models

Model	Details
VGG-13	Visual Geometry Group - 13 layers
VGG16	Visual Geometry Group - 16 layers
VGG19	Visual Geometry Group - 19 layers
GoogLeNet	Inception architecture - 22 layers
EfficientNet-B4	EfficientNet, 4th variant - 52 layers
DenseNet-121	Dense connectivity - 121 layers
MobileNet-v2	Mobile-friendly - 53 layers
InceptionV3	Inception, improved version - 48 layers
InceptionResNetV2	Inception + ResNet - 164 layers
ResNet-50	Residual Networks - 50 layers

Table 5: Datasets Used in Research

Name	Number of Images	Reference
ISIC 2017	2000	Codella et al. (2017)
ISIC 2019	25331	Tschandl et al. (2018) Codella et al. (2017) Combalia et al. (2019)
ISIC 2020	33126	Rotemberg et al. (2021)
PH^2	200	Mendonça et al. (2013)
SKINL2	376	de Faria et al. (2019)
7-Point Criteria Evaluation	1011	Kawahara et al. (2019)
DermNet	19500	DermNet (DermNet)
DERMQUEST	137	Vision (2021)
DermIS	1000	Diepgen TL (Diepgen TL)
MED-Node	170	Giotis et al. (2015)
FitzPatrick17k	16577	Groh et al. (2021) Groh et al. (2022)
SD-198	6584	Sun et al. (2016)
SD-260	20600	Yang et al. (2020)

A.2 MAR-MELA & MAR-MELA-CNN Code

The Python code and dataset for this research are available on my GitHub repository: https://github.com/404MiguelMarinhoNotFound/Thesis_Code_MAR-MELA.