A Comparative Deep Learning Approach to Accurate Skin Cancer Diagnosis

Susana Alvarez Zuluaga[1]

Advisor(s):
Daniel Otero Gomez[2]
Francisco Ivan Zuluaga[3]

Research practice III
Mathematical Engineering
School of Applied Sciences and Engineering
Universidad EAFIT

November 2023

[1]Mathematical Engineering Student (salvarezz1@eafit.edu.co)
[2]Mathematical Engineering Graduate (doterog@eafit.edu.co)
[3]Professor in School of Applied Science and Engineering (fzuluag2@eafit.edu.co)

1 Introduction

Skin cancer is a global health concern, with increasing incidence rates and severe consequences for public health systems. Over the last years, melanoma rates in the US have been growing rapidly, doubling from 1982 to 2011 (American Academy of Dermatology, 2022). According to Global Cancer Observatory (2020), approximately 1.2 million non-melanoma skin cancers and more than 300,000 melanoma skin cancers were diagnosed in 2020. Furthermore, nearly 9,500 people in the U.S. are diagnosed with skin cancer every single day (American Academy of Dermatology, 2022).

Detecting skin cancer can be challenging for doctors due to various factors. The wide range of presentations that skin cancers can exhibit, including multiple colors, shapes, and sizes, can lead to difficulties in distinguishing them. Additionally, certain skin tumors may mimic non-cancerous growths or typical skin characteristics, making accurate diagnosis even more difficult (Gururaj *et al.*, 2023). According to Sondermann *et al.* (2016), approximately 30% of skin cancers are misdiagnosed. In the case of children, misdiagnosis rates are elevated constituting 40 % of the occurrences (The Skin Cancer Foundation, 2023). In addition to the difficulty of detecting skin cancer, traditional diagnostic approaches involve time-consuming and invasive procedures, causing delays in the diagnosis and the initiation of treatment. According to Cancer Research UK (2023), it takes about two to three weeks to get the results of a biopsy making the whole diagnosis process take more than one month. These challenges raise substantial concerns since diagnosing skin cancer at an early stage is crucial in order to be able to treat it (Gururaj *et al.*, 2023).

The use of artificial intelligence (AI), specifically deep learning (DL), in this domain offers a revolutionary solution. Utilizing advanced DL algorithms that can learn autonomously and extract features from images can lead to fast and accurate diagnosis (Garg *et al.*, 2021). Numerous researchers, including Zhang *et al.* (2020), Cengil *et al.* (2021) and Shete *et al.* (2021), have conducted studies implementing deep learning techniques to develop models capable of diagnosing skin cancer. After performing a thorough literature review, it's evident that although several deep learning structures have been proposed and applied to dermatological datasets a systematic performance comparison is necessary. Few authors such as (Arshed *et al.*, 2023) have presented comparative studies.

This research project addresses this problem by conducting a comparative analysis of diverse deep learning models applied to skin cancer diagnosis. This comparative study will highlight the strengths and weaknesses of each model, enabling researchers to make informed decisions about model selection based on specific diagnostic needs. Furthermore, this exploration will contribute to the enrichment of the field of medical imaging research. The importance of this investigation extends beyond its technological implications, impacting human lives and healthcare systems worldwide. Efficient and accurate skin cancer diagnosis can decrease mortality rates and improve patient outcomes (Garg *et al.*, 2021). Moreover, incorporating deep learning algorithms into dermatology opens avenues for a broader integration

of AI into other medical diagnostics, transforming how various diseases are detected and managed.

2 Statement of the problem

2.1 Statement of the problem

Skin cancer is a prevalent and increasingly concerning health issue. According to The Skin Cancer Foundation (2023), one in every three cancers diagnosed is skin cancer, and one in every five Americans will develop skin cancer in their lifetime. Skin cancer affects diverse population groups across various geographical areas and operates globally; however, almost 80% of the cases are presented in people from North America and Europe (Global Cancer Observatory, 2020). Skin cancer was first diagnosed in the 1800s but was not very common. Over the last 30 years, melanoma rates in the US have doubled, making it a concerning health issue nowadays (American Academy of Dermatology, 2022).

Skin cancer originates from skin cells. These cells grow and divide to generate new cells. While older skin cells naturally die and are replaced, sometimes this process malfunctions, excess cells are produced when not needed, and old cells die when they should not. This surplus of cells aggregates into a mass known as a tumor (Queen, 2017). Melanoma is the most aggressive and severe type of skin cancer. In some cases, melanoma spreads via the lymphatic or circulatory systems, and it can cause death. The exact cause of melanoma remains unidentified; however, several factors, such as genetics and exposure to ultraviolet radiation, can influence (Yaiza *et al.*, 2019).

Research indicates that early and correct skin cancer detection substantially reduces mortality rates (Zhang *et al.*, 2020). Diagnosing skin cancer can be challenging for doctors due to the wide range of skin cancer presentations and the time-consuming nature of conventional diagnostic methods. Because of this, different AI approaches have been used to detect skin cancer using images of pigmented skin lesions (Garg *et al.*, 2021). This problem is commonly known in the literature as Image Classification. Image classification problems can be solved using a wide variety of different AI algorithms and techniques. However, this research project will focus on deep learning algorithms, specifically convolutional neural networks (CNNs).

Image classification is a problem that traverses many domains. From agriculture and environmental conservation to retail, security, and healthcare, it is vital in optimizing processes and improving decision-making. Specifically, image classification algorithms in healthcare have been used to identify lymph nodes in breast cancer patients to categorize and diagnose various medical conditions from X-rays and medical images and others (Pai & Giridharan, 2019). Particularly in the field of dermatology Rosas-Lara *et al.* (2022), A. *et al.* (2018) and Esteva *et al.* (2017) have all performed studies on skin cancer diagnosis using image classification.

2.2 Formalization of the problem

Convolutional networks, also known as CNNs, are a type of neural network used for processing data with grid-like topology, such as image data. They employ convolutions which are a mathematical operation that combines two functions to produce a third, reflecting how one function modifies the other. In the context of CNNs, convolution involves sliding a filter (also known as a kernel) over the input data and computing the dot product at each position. Mathematically, the convolution of two functions f and g, is denoted as $(f * g)(t)$ and is defined as:

$$(f * g)(t) = \int_{-\infty}^{\infty} f(\tau)\, g(t - \tau)\, d\tau$$

In the case of CNNs, the functions f and g are the input data and the convolutional filter, respectively(Goodfellow *et al.*, 2016).

The success of CNNs in image processing lies in their ability to capture hierarchical features. The human visual system is excellent at recognizing patterns, and CNNs mimic this process. Convolutional layers are designed to detect low-level features like edges and gradually combine them to form complex high-level features. The key intuition behind using convolutions for images is that they exploit spatial hierarchies. Images have hierarchical structures, where low-level features combine to form more abstract representations. Convolutional layers identify local patterns, and by stacking these layers, the network learns to recognize increasingly complex features (Goodfellow *et al.*, 2016).

In the 1990s, Lecun *et al.* (1998) introduced the LeNet-5 architecture, representing one of the initial applications of Convolutional Neural Networks (CNNs) to real-world problems. Since then, the landscape of CNNs has witnessed a remarkable evolution with numerous advancements. Over the past 15 years, contributions have played a significant role in shaping and advancing the field. Particularly noteworthy is the introduction of AlexNet by Krizhevsky *et al.* (2012). This work marked a revolutionary moment by demonstrating a significant performance improvement over conventional methods in the ImageNet Recognition Challenge, thereby underscoring the potential of deep learning. Following this, the concept of residual learning was introduced in He *et al.* (2016a) were residual networks (ResNets) addressed the challenge of vanishing gradients, enabling the training of extremely deep networks.

A diverse range of architectures exists today, each presenting unique strengths and applications. Within this study, the primary challenge involves choosing architectures specifically suited to skin cancer diagnosis. The objective is to implement and train various models and subsequently compare the results obtained from each. This comparative analysis aims to contribute to a deeper understanding of the most effective approach for our research goals.

3 Objectives

3.1 General objective

To conduct a comparative analysis of existing deep learning techniques in the domain of skin cancer diagnosis highlighting the strengths and weaknesses of each one.

3.2 Specific objectives

- To identify how different deep learning techniques have been used by other authors in this same field through an exhaustive literature review.
- To obtain a preprocessed dataset by using different techniques such as data cleaning, data transformation, data augmentation, and others.
- To enable a comprehensive understanding of the capabilities of distinct deep learning models by training and fine-tuning them with dermatoscopic images of skin lesions.
- To asses which models are the most adequate for this problem by comparing their results.

4 Justification

Skin cancer is a prevalent global health concern, accounting for many cancer cases worldwide. Conventional methods of skin cancer diagnosis depend on human visual assessment, which is subjective and prone to variability (Khater *et al.*, 2023). This limitation highlights the need to explore technological solutions, such as deep learning, to improve the reliability of skin cancer diagnosis. By creating algorithms to recognize patterns in images that are frequently invisible to the human eye, advances in deep learning have completely changed how medical image analysis is done.

Previous studies have shown that deep learning improves medical image interpretation, making diagnoses more reliable and accurate. For instance, Esteva *et al.* (2017) demonstrated the viability of deep neural networks in detecting skin cancer by achieving a performance comparable to dermatologists. Furthermore, A. *et al.* (2018) presented a study of convolutional neural networks vs. dermatologists. The CNNs outperformed most dermatologists. It is evident that dermatologists can benefit from assistance from CNNs' image classification. These results demonstrate the importance of researching deep learning methods for skin cancer diagnostics.

This study's value comes from its ability to accelerate the diagnosis process, reduce human error, and facilitate early intervention. Early skin cancer detection is fundamental for successful treatment, as delayed diagnosis can decrease survival rates. By providing an

accurate deep learning model, dermatologists and healthcare professionals can make informed decisions, improving patient care. Furthermore, given the global shortage of dermatologists in underserved regions, automated diagnosis with deep learning models can help close the breach in access to medical expertise (Schlessinger *et al.*, 2019). The significance of this study extends beyond its impact on clinical outcomes, encompassing a technical contribution to the field of artificial intelligence. Through the presentation of a comparative analysis of different deep learning techniques, this research provides a solid foundation for other researchers to initiate their research, providing them with valuable insights and methodologies applicable to their own investigations and applications.

5 Scope

The scope of this research project includes a comparative study of deep learning methods for skin cancer diagnosis. While the research aims to provide valuable insights into optimal deep learning models, it's important to acknowledge certain limitations that may impact the execution of the project. A significant obstacle lies in the nature of the dataset we will use, which is highly unbalanced. This issue can have an impact on how well deep learning models perform. Additionally, the computational complexity associated with training and optimizing deep learning models demands powerful computational resources, which might impact the feasibility of exploring a wide array of architectures and hyper parameters. Another barrier lies in the explainability of deep learning models. As they often operate as black-box systems, understanding the rationale behind their predictions becomes challenging, raising questions about their integration into clinical practice and ethical considerations.

Python will serve as the primary programming language for this research. The models' implementation, training, and evaluation will be conducted using a deep learning framework named PyTorch Lightning (Falcon & The PyTorch Lightning team, 2019). For accelerating model training, powerful GPU resources will also be needed. Access to academic databases and medical literature will provide the necessary knowledge for understanding skin cancer diagnosis and deep learning methodologies.

6 State of the art

As mentioned before, deep learning algorithms are widely used in the field of image classification. Because of this, deep learning has been used by different authors to address the problem of skin cancer detection and classification. For example Cengil *et al.* (2021) propose a hybrid methodology in which they combine machine learning and deep learning algorithms. These authors used CNNs to perform feature extraction from the images, and then they used 3 machine learning algorithms to increase accuracy. When dealing with classification problems, CNNs have a classification function in their last layer which is usually a Softmax function. However, these authors used two common neural network architectures: Alexnet and Resnet and instead of having a Softmax classifier in the last layer they used Decision

Tree (DT), K- Nearest Neighbors (KNN) and Support Vector Machine (SVM) structures in order to experiment an see which one yielded better results. After experimenting, the best results were obtained when a SVM was used in the last layer. Results were very satisfactory. When training an Alexnet combined with a SVM an accuracy of 0.7780 was obtained and when training a Resnet18 combined with a SVM an accuracy of 0.7623 was obtained. These authors used the MNIST HAM-10000 dataset.

Shete *et al.* (2021) reveal another intent of using deep learning to diagnose skin cancer. Byusing CNN and transfer learning models, the authors were able to increase classification accuracy. The authors used deep learning models pre-trained on the ImageNet dataset for Transfer learning. Then these pre-trained models were further trained on the HAM10000 dataset. The authors encountered a problem with the dataset used because it was heavily imbalanced and one category in the dataset accounted for more than 50% of the overall dataset. To solve this problem Shete *et al.* (2021) used Data Augmentation to prevent data from being overfit. By varying the translation, rotation, and zooming of the files, they were able to make multiple copies of the existing dataset.

In a similar way, Garg *et al.* (2021) used the MNIST HAM-10000 dataset to implement a system designed to identify instances of skin cancer and categorize them into distinct classes, using CNNs. The proposed approach integrates image processing and a deep learning model for the diagnostic process. The dermoscopy images undergo a series of techniques in order to eliminate noise and enhance image resolution. Moreover, image augmentation techniques are employed to expand the dataset. Ultimately, Garg *et al.* (2021) employ Transfer Learning to enhance the accuracy of image classification. The CNN model utilized achieved a weighted average Precision score of 0.88. The application of Transfer Learning through the ResNet model resulted in an accuracy rate of 0.91.

Pai & Giridharan (2019) also performed a study in which they use deep learning to predict and classify seven different types of skin lesions. The observations and results where obtained using the MNIST:HAM10000 dataset. The authors split the image data in the ratio of 80:20, for training and testing data respectively. 15% of the training dataset was used for validation. The original dimensions of the images which were 450x600x3 were reduced to 224X224X3 as part of the pre-processing. VGGNet was selected as the CNN architecture to train this model. The architecture implemented consisted of 16 convolutional layers, 3X3 kernels, categorical cross entropy function and Adams optimizer. The model yielded an accuracy of 78% on test data. After training the model, the authors developed a website for the real time usage of the model, which can be used to predict the three most probable types of skin lesions for a given image.

Lastly, Alam *et al.* (2022) introduce an approach grounded in deep learning principles, specifically designed to tackle the challenge of an imbalanced dataset. To address this issue, data augmentation techniques were employed, effectively balancing the representation of

different skin cancer classes within the dataset. Alam *et al.* (2022) used the Skin Cancer MNIST: HAM10000 dataset, encompassing seven distinct categories of skin lesions. In this study, the classification of skin cancer was accomplished using deep learning-based models, specifically AlexNet, InceptionV3, and RegNetY-320. The findings demonstrate that RegNetY-320 surpassed both InceptionV3 and AlexNet in terms of accuracy, F1-score, and ROC curve analysis, across both imbalanced and balanced datasets. The performance of the proposed framework consistently outperformed conventional methodologies. The proposed framework achieved an accuracy of 91%.

7 Methods

When addressing this types of research questions authors usually use the CRISP-DM (Cross-Industry Standard Process for Data Mining) methodology. CRISP-DM is a widely recognized methodology which was presented in 1996. This methodology provides a structured approach for solving complex data driven problems through iterative phases. The selection of the CRISP-DM methodology is derived by its established success in guiding data-driven research projects. For example Rosas-Lara *et al.* (2022) present a study where they use Convolutional Neural Networks to support melanoma skin cancer detection in which they effectively use the CRISP-DM methodology.

The methodology contains six stages: Business Understanding, Data Understanding, Data Preparation, Modeling, Evaluation, and Deployment. The Business Understanding phase is the phase in which a clear understanding of the goals, objectives, and requirements of the research project are established. The Data Understanding phase involves selecting or collecting the data, exploring it and gaining insights into its structure, quality, and potential limitations. The Data Preparation phase focuses on data preprocessing techniques such as data cleaning, transformation, and augmentation. The Modeling phase focuses on the creation, evaluation, and optimization of the predictive models using the preprocessed dataset. The Deployment phase involves deploying the selected and optimized model into a real-world environment for operational use (Wirth & Hipp, 2000). It is worth noting that this phase will not be covered in this research project. We will now go into detail on how each stage was conducted.

7.1 Business Understanding

This phase was the first one conducted. We started by understanding the problem and the insights gained are summarized in Section 2. Subsequently, an exhaustive literature review was conducted to understand how previous researchers addressed this same problem. This review is presented in Section 6. These preliminary investigations served as a foundation to establish the objectives and scope of the of this research project which are presented in Sections 3 and 5 respectively.

7.2 Data Understanding

As mentioned earlier, this stage entails the selection or collection of the data, conducting an exploration, and deriving insights into its structure, quality, and potential limitations. The dataset selected was the MNIST HAM-10000 Skin Cancer dataset (Mader, 2018) which contains 10013 dermatoscopic images of skin lesions. These images are of size 600×450 pixels and they are divided in the following seven classes:

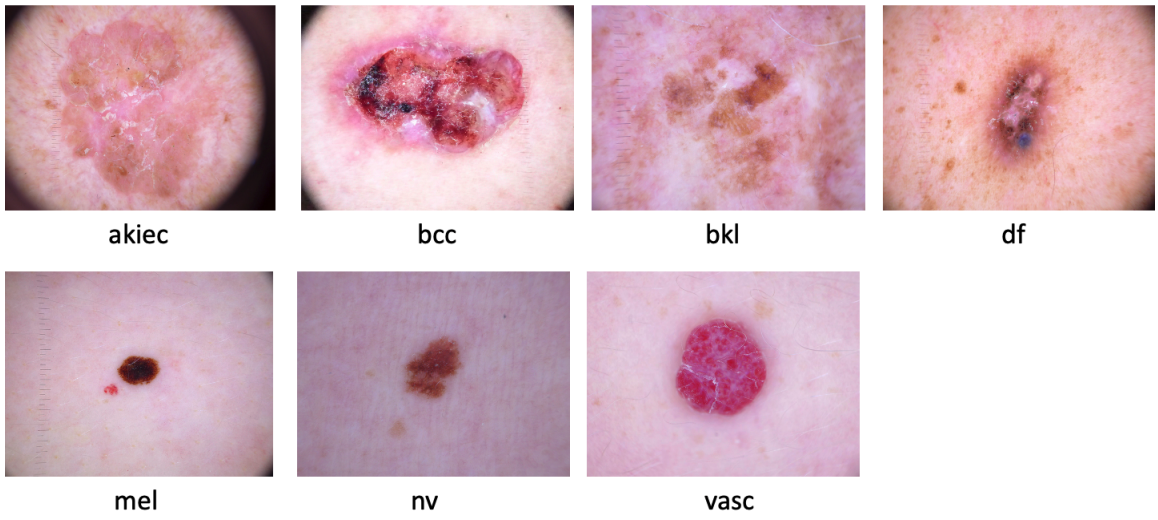

Figure 1: Skin Lesion Classes

- Actinic keratoses (akiec): Precancerous skin lesions resulting from prolonged sun exposure. They typically appear as dry, scaly patches and, if left untreated, may progress to squamous cell carcinoma.
- Basal cell carcinoma (bcc): Common type of skin cancer that arises from the basal cells in the skin's outermost layer. It often appears as a pearly or waxy bump and is typically slow-growing and rarely metastasizes.
- Benign keratosis-like (bkl): Non-cancerous skin growths characterized by thickening of the outer layer (epidermis). These growths, such as seborrheic keratosis, are usually tan, brown, or black, and have a waxy, scaly, or crusty surface.
- Dermatofibroma(df): Benign skin tumors characterized by an overgrowth of fibrous tissue. They often appear as small, brownish nodules on the skin and are typically harmless.
- Melanoma (mel): Melanoma is a malignant type of skin cancer that originates in melanocytes, the pigment-producing cells. It is known for its potential to metastasize

and poses a significant risk if not detected and treated early. Melanomas often exhibit irregular borders, uneven color, and changes in size, making regular skin examinations crucial for early diagnosis.

- Melanocytic nevi (nv): Are benign growths on the skin composed of melanocytes, the pigment-producing cells. They are typically harmless, but monitoring changes in size, shape, or color is essential for early detection of potential malignancy.

- Vascular lesiones (vasc): Vascular lesions encompass a range of skin abnormalities related to blood vessels. These can include birthmarks, hemangiomas, and other vascular malformations, which vary in appearance and potential complications.

(Mayo Clinic, 2022)

Table 1 shows the number of images and proportions of each class in the dataset. It is evident that the dataset is highly unbalanced were the class nv accounts for 67% of all the data and the class vasc accounts for only 1% of all the data. This is a limitation that the dataset has and that will be taken into account in the data preparation phase.

Class	Number of images	Proportion
nv	6705	0.67
mel	1113	0.11
bkl	1097	0.11
bcc	514	0.05
akiec	327	0.03
vasc	142	0.01
df	115	0.01

Table 1: Skin lesion images for each class

7.3 Data Preparation

As previously stated, this phase is centered on the preparation of the dataset through the application of various data preprocessing techniques, such as data cleaning, transformations, and augmentation. Due to the inherent imbalance within the dataset, data augmentation techniques were implemented. Random transformations were applied to the images during each sampling instance, thereby introducing variability in the images. The different transformations applied to the images are shown in Table 2. These transformations were applied using Torchvision (maintainers & contributors, 2016). An exmaple of a transformed image is shown in Figure 2.

Transformation	Configuration
Rotation (Random)	[-180°, 180]
Affine (Random)	[-90°, 90]
Horizontal Flip (Random)	
Vertical Flip (Random)	
Resize	224x224 px

Table 2: Image Transformations

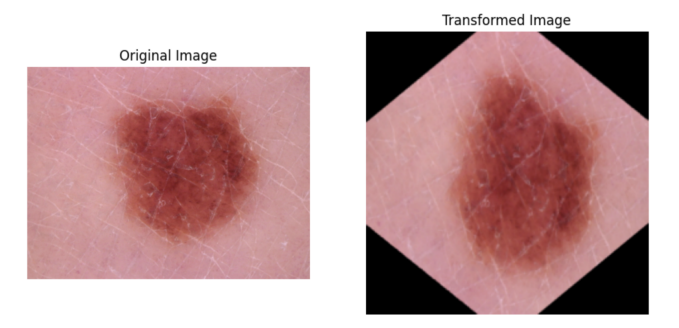

Figure 2: Original Image vs Transformed Image

7.4 Modeling and Evaluation

This phase consists on creating, evaluating, and optimizing different models. In this section we will go in depth in the CNN architectures that were trained and also the different techniques applied for dealing with the problem of an imbalanced dataset.

7.5 Models

Four different models were trained to detect skin cancer. For three of these models, we used transfer learning. This involved starting with pre-trained architectures—ResNet18, ResNet50, and RegYNet400—that had originally learned from the ImageNet dataset. Afterward, we fine-tuned these pre-trained models to better suit the characteristics of our skin cancer dataset. In addition to transfer learning, we created a custom Convolutional Neural Network (CNN) from scratch, providing a distinct approach. The different CNN architectures trained are presented below in detail.

7.5.1 ResNet18 and ResNet50

ResNet18 is a CNN architecture introduced by He *et al.* (2016b). It is part of the Residual Network (ResNet) family and is specifically designed for image classification tasks. ResNet18 employs residual blocks to mitigate the vanishing gradient problem, enabling the training of very deep networks. The ResNet18 architecture consists of 18 layers, including convolutional

layers, batch normalization, and shortcut connections. The design of residual connections allows for the successful training of deep networks, making ResNet18 a popular choice in computer vision applications. The ResNet50 architecture is another extension of the ResNet architecture which is designed for more complex tasks. It comprises 50 layers and includes bottleneck blocks, which use 1x1 convolutions to reduce the computational complexity while preserving representational power. ResNet50 has demonstrated superior performance in various image recognition tasks, owing to its ability to capture intricate features through its deep structure (He *et al.*, 2016b).

7.5.2 RegYNet400

Regularization network (RegNet), first introduced by Radosavovic *et al.* (2020), is an innovative neural network architecture designed for balancing model efficiency and performance across a broad spectrum of tasks. RegNet employs a systematic design space exploration strategy to discover optimal architectures. The key feature of RegNet is its modular structure, which consists of repeated multi-path building blocks to facilitate efficient scaling. This modular design allows for flexibility in adapting the architecture to different computational constraints and task requirements. RegNet's exploration of the design space has led to models that consistently outperform other state-of-the-art architectures in terms of accuracy while requiring fewer computational resources (Radosavovic *et al.*, 2020).

7.5.3 Self made CNN

A self made CNN was implemented. This CNN's architecture contains a stack of convolutional layers. The convolutional layers are characterized by 3x3 filters, a stride of 1, and a padding of 1, aiming to preserve spatial information in the input. Each convolutional layer is followed by a Rectified Linear Unit (ReLU) activation function and a max-pooling layer with a 2x2 kernel. After the convolutional block, the model includes a flattening operation and a fully connected layer (linear layer) serving as a classifier.

7.6 Methodologies implemented for dealing with class imbalance

7.6.1 Weighted Loss

For this project we used the cross entropy loss function. This loss is very common when dealing with classification problems. We implemented a weighted loss which consists in introducing weights in the loss function in order to handle class imbalance in a dataset. The idea is to weight the loss computed for different samples differently based on whether they belong to the majority or the minority classes. A higher weight was assigned to the loss encountered by the samples associated with minor classes. There are different weighting schemes that can be used to compute the weight for each sample. We used the following formula to calculate the weight for a sample n in class c. With that being said, the weights

for each class are presented in Table 3

$$W_{n,c} = \frac{\text{Maximum class proportion}}{\text{Proportion of samples in class c}}$$

Class	Weights
df	58.30
vasc	47.22
akiec	20.50
bcc	13.04
bkl	6.11
mel	6.02
nv	1.00

Table 3: Weights for a sample in each class in weighted sampling

7.6.2 Weighted Sampling

Weighted sampling is another a technique commonly employed to address class imbalance issues. Weighted sampling aims to mitigate this bias by assigning higher weights to instances from minority classes during the sampling process. In weighted sampling, the sampling probability for each instance is determined based on the assigned weights. The idea is to give more importance to instances from underrepresented classes, ensuring that the model is exposed to a more balanced set of examples. The weights assigned to samples in each class are the same ones used in the weighted loss implementation which are shown in Table 3.

7.6.3 Adaptive Loss

Adaptive loss was another technique which was implemented to deal with class imbalance. This technique consists in changing the weights assigned to each class in the loss function on every epoch depending on the f1 validation score obtained for the given class. A higher weight is given to those classes which have lower f1 scores and lower weights are given to those classes which have higher f1 scores. Figure 3 shows the evolution of the validation f1 scores obtained for each class and the weights assigned to each class during 5 epochs. For example it can be seen that in epoch 0 the class with the lowest validation f1 score is class mel (yellow) so the highest weight in this epoch is assigned to this class. Another example can be seen in epoch 1 were the class with the highest validation f1 score is bcc (orange) so the weight assigned to this class in this epoch is the lowest of all.

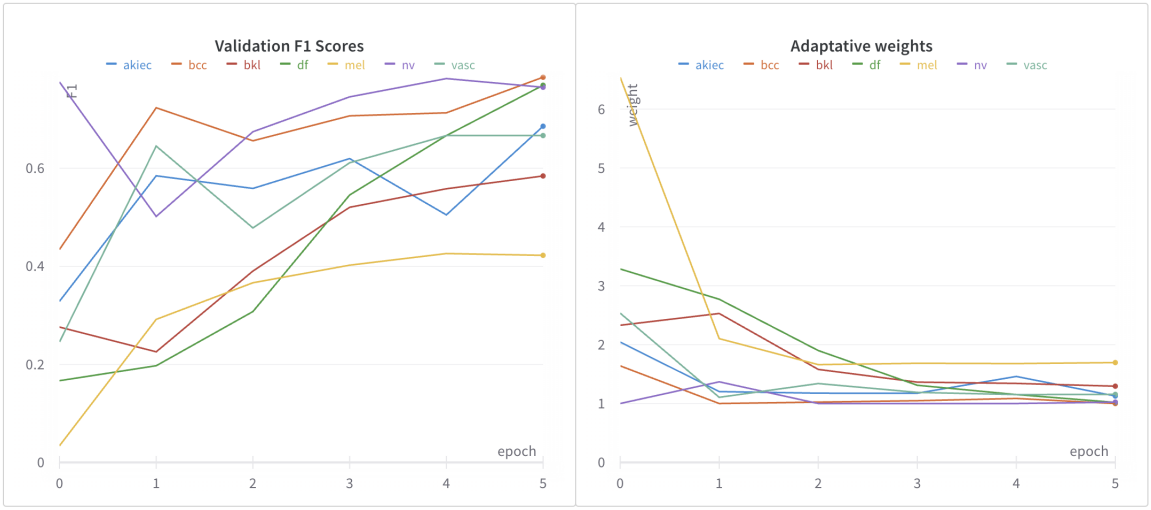

Figure 3: Weights vs Validation F1 Scores

8 Results

The deep learning models were trained on a MacBook Pro M1. The dataset was partitioned, allocating 80% for training and 10% each for validation and testing. Stochastic gradient descent was employed as the optimizer, with the Cross Entropy Loss serving as the chosen loss function. Various hyperparameters, including learning rate, momentum, and weight decay, were varied through short experiments. During these experiments, the different implemented techniques such as adaptive loss, weighted loss, and weighted sampling were combined. The experimentation involved fine-tuning a ResNet18 model for 5 epochs. Results from this experimentation are presented in Table 4 and Figure 4.

Notably, Experiment 6, employing a learning rate of 0.001, momentum of 0.9, weight decay of 0.01, and incorporating both weighted sampling and adaptive loss, demonstrates superior results across all the experiments. This configuration achieved the highest train accuracy of 0.8616, validating its efficacy in capturing intricate patterns within the training data in just 5 epochs. Furthermore, it yielded validation and test accuracies of 0.7832 and 0.8013, respectively. Conversely, Experiment 2, were adaptive loss was applied but without weighted sampling, illustrates the nuanced impact of individual techniques. These results demonstrate the importance of carefully selecting hyperparameters and considering different techniques to deal with class imbalance.

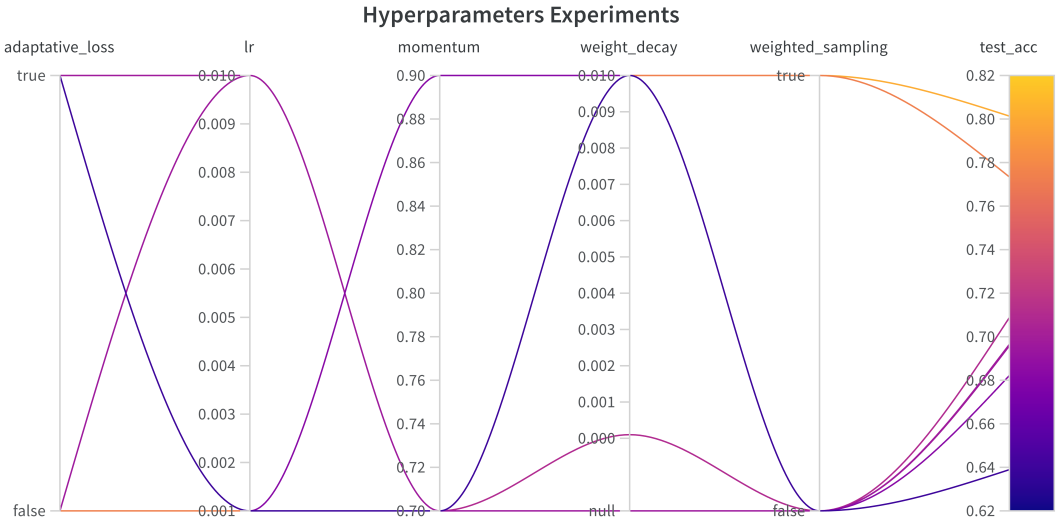

Figure 4: Test accuracy obtained with different combinations of hyperparameters and methods.

Exp	Learning Rate	Momentum	Weight decay	Weighted Sampling	Adaptative Loss	Train Accuracy	Validation Accuracy	Test Accuracy
1	0.01	0.7	0.0001	false	false	0.6568	0.6243	0.6958
2	0.01	0.7	0.0001	false	true	0.6277	0.5568	0.7088
3	0.01	0.7	0.01	false	true	0.5635	0.6342	0.6965
4	0.001	0.7	0.01	false	true	0.6507	0.6215	0.6389
5	0.001	0.9	0.01	false	true	0.7025	0.6624	0.6822
6	**0.001**	**0.9**	**0.01**	**true**	**true**	**0.8616**	**0.7832**	**0.8013**
7	0.001	0.9	0.01	true	false	0.8755	0.767	0.7734

Table 4: Results obtained with different hyperparameters and methods

After conducting the experimentation, architectures ResNet18, ResNet50, RegYNet400 were fine tuned for 150 epochs and the self made CNN was also trained for 150 epochs. The hyperparameters and techniques chosen were those used in the experiment 6 in Table 4. The results obtained after training this models are presented in Table 5 and Figure 5. These results reveal distinctive performance characteristics across different models. The ResNet18 model achieved a test accuracy of 0.8611, indicating its ability to generalize well to new, unseen data. However, a noticeable performance gap exists between the training accuracy (0.9680) and the validation accuracy (0.7904), suggesting a potential risk of overfitting. The F1 scores for individual classes vary, with some classes showing strong performance (e.g., 'df' with 0.8571), while others could benefit from improvement (e.g., 'mel' with 0.6410). The ResNet50 model outperforms others with a test accuracy of 0.8791, demonstrating superior generalization

capabilities. Moreover, the differences between training, testing and validation accuracies also suggest signs of overfitting. The F1 scores for all classes are notably high, reflecting the model's proficiency in distinguishing between different skin cancer types. RegNetY400 while achieving a test accuracy of 0.8452, exhibits a similar overfitting pattern. On the other hand the self-made CNN while achieving a lower overall test accuracy of 0.7203, demonstrates a consistent performance pattern between training and validation. However, the relatively lower F1 scores for most classes suggest the need for improvement in class-wise predictions, especially for 'akiec', 'bcc', and 'mel'.

	Accuracy			Test F1 Scores						
Model	Test	Valid	Train	akiec	bcc	bkl	df	mel	nv	vasc
ResNet18	0.8611	0.7904	0.9680	0.8060	0.7500	0.7521	0.8571	0.6410	0.9296	1.0000
ResNet50	**0.8791**	**0.8762**	**0.9828**	**0.8529**	**0.8224**	**0.7600**	**0.9565**	**0.6577**	**0.9358**	**0.9630**
RegNetY 400	0.8452	0.8353	0.9827	0.6512	0.7523	0.6897	0.6667	0.6518	0.9252	0.8889
Self made CNN	0.7203	0.6437	0.7308	0.4483	0.4681	0.3819	0.3529	0.2930	0.8734	0.5143

Table 5: Results obtained when training different CNN architectures for 150 epochs

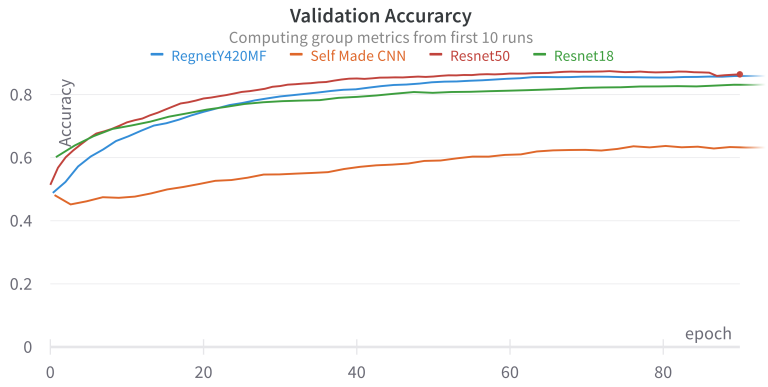

Figure 5: Validation Accuracy Curves for different CNNs

It's evident that the ResNet50 architecture produced the best results among all the models. Across all models, there was a consistent observation of the 'mel' class obtaining the lowest F1 scores. Figure 6 displays the test confusion matrix derived from the ResNet50 model. There is a significant confusion between the 'mel' and 'nv' classes, which is a concerning issue considering melanoma is a form of skin cancer, while 'nv' represents Melanocytic nevi—benign growths of skin. This confusion raises the risk of misdiagnosing melanomas and yielding false positives. Strikingly, this confusion pattern persisted across all the models studied.

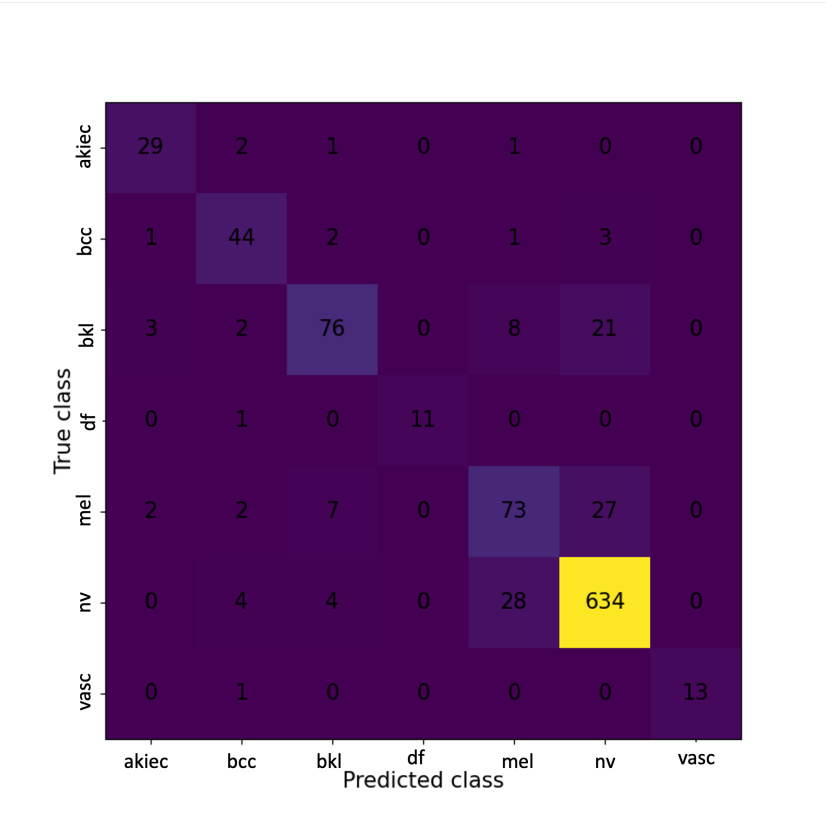

Figure 6: ResNet50 Confusion Matrix

9 Conclusions and Further Works

The achieved results and learnings obtained were very satisfactory when compared to other papers. Notably, ResNet50 emerged as the top-performing architecture in this study. Nevertheless, the outcomes from ResNet18 are highly commendable as well. It's worth noting that ResNet50, being a deeper architecture with substantially more parameters, demands twice the training time compared to ResNet18. Additionally, during the experimentation phase, it was evident that the weighted sampling and adaptive loss techniques produced highly effective results, showcasing the potential of combining these strategies for improved model performance.

Given the observed challenge of the misdiagnosis for the melanoma class, a critical focus for future work involves addressing the issue of false negatives in melanoma detection. To enhance the precision of the models, exploring advanced techniques such as ensemble learning or leveraging more sophisticated architectures specifically designed for imbalanced datasets could be beneficial. Additionally, fine-tuning hyperparameters and incorporating class-specific adjustments might help mitigate the misclasification of the the melanoma class. Integrating additional clinical data or dermatological features into the model training process

could further refine the distinction between benign and malignant cases. Furthermore, the prevalent observation of overfitting in our models, as highlighted in the results, shows the importance of incorporating effective regularization techniques. To mitigate overfitting, we plan to implement different techniques such as dropout. Finally we would also like to explore using other optimizers such as Adam.

10 Intellectual property

According to the internal regulation on intellectual property within Universidad EAFIT, the results of this research practice are product of Susana Alvarez Zuluaga and Daniel Otero Gomez.

In case further products, beside academic articles, that could be generated from this work, the intellectual property distribution related to them will be directed under the current regulation of this matter determined by Universidad EAFIT (2017).

References

A., Haenssle H., Fink, C., Schneiderbauer, R., Toberer, F., Buhl, T., Blum, A., Kalloo, A., Ben Hadj Hassen, A., Thomas, L., Enk, A., & Uhlmann, L. 2018. Man against Machine: Diagnostic performance of a deep learning convolutional neural network for dermoscopic melanoma recognition in comparison to 58 dermatologists. *Annals of Oncology*, **29**(8), 1836–1842.

Alam, Talha Mahboob, Shaukat, Kamran, Khan, Waseem Ahmad, Hameed, Ibrahim A., Almuqren, Latifah Abd, Raza, Muhammad Ahsan, Aslam, Memoona, & Luo, Suhuai. 2022. An Efficient Deep Learning-Based Skin Cancer Classifier for an Imbalanced Dataset. *Diagnostics*, **12**(9).

American Academy of Dermatology. 2022 (Apr).

Arshed, Muhammad Asad, Mumtaz, Shahzad, Ibrahim, Muhammad, Ahmed, Saeed, Tahir, Muhammad, & Shafi, Muhammad. 2023. Multi-Class Skin Cancer Classification Using Vision Transformer Networks and Convolutional Neural Network-Based Pre-Trained Models. *Information*, **14**.

Cancer Research UK, ". 2023 (Sep). *Tests for skin cancer*.

Cengil, Emine, ÇINAR, Ahmet, & YILDIRIM, Muhammed. 2021. Hybrid Convolutional Neural Network Architectures for Skin Cancer Classification. *European Journal of Science and Technology*, 694–701.

Esteva, Andre, Kuprel, Brett, Novoa, Roberto A, Ko, Justin, Swetter, Susan M, Blau, Helen M, & Thrun, Sebastian. 2017. Dermatologist-level classification of skin cancer with deep neural networks. *Nature*, **542**(7639), 115–118.

Falcon, William, & The PyTorch Lightning team. 2019 (Mar.). *PyTorch Lightning*.

Garg, Rishu, Maheshwari, Saumil, & Shukla, Anupam. 2021. Decision Support System for Detection and Classification of Skin Cancer Using CNN. *Advances in Intelligent Systems and Computing*, **1189**, 578–586.

Global Cancer Observatory. 2020.

Goodfellow, Ian, Bengio, Yoshua, & Courville, Aaron. 2016. *Deep Learning*.

Gururaj, H. L., Manju, N., Nagarjun, A., Manjunath Aradhya, V. N., & Flammini, Francesco. 2023. DeepSkin: A Deep Learning Approach for Skin Cancer Classification. *IEEE Access*, **11**(May), 50205–50214.

He, Kaiming, Zhang, Xiangyu, Ren, Shaoqing, & Sun, Jian. 2016a. Deep Residual Learning for Image Recognition. *Pages 770–778 of: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

He, Kaiming, Zhang, Xiangyu, Ren, Shaoqing, & Sun, Jian. 2016b. Deep Residual Learning for Image Recognition. *In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Khater, Tarek, Ansari, Sam, Mahmoud, Soliman, Hussain, Abir, & Tawfik, Hissam. 2023. Skin cancer classification using explainable artificial intelligence on pre-extracted image features. *Intelligent Systems with Applications*, **20**, 200275.

Krizhevsky, Alex, Sutskever, Ilya, & Hinton, Geoffrey E. 2012. ImageNet Classification with Deep Convolutional Neural Networks. vol. 25. Curran Associates, Inc.

Lecun, Y., Bottou, L., Bengio, Y., & Haffner, P. 1998. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, **86**(11), 2278–2324.

Mader, K Scott. 2018 (Sep). *Skin cancer MNIST: Ham10000*.

maintainers, TorchVision, & contributors. 2016 (Nov.). *TorchVision: PyTorch's Computer Vision library*.

Mayo Clinic, ". 2022 (Dec). *Skin cancer*.

Pai, Kiran, & Giridharan, Anandi. 2019. Convolutional Neural Networks for classifying skin lesions. *IEEE Region 10 Annual International Conference, Proceedings/TENCON*, **2019-October**, 1794–1796.

Queen, Lauren. 2017. Skin Cancer: Causes, Prevention, and Treatment. *Senior Honors Theses*, 1–29.

Radosavovic, Ilija, Johnson, Justin, Xie, Saining, Lo, Wan-Yen, & Dollár, Piotr. 2020. RegNet: A Highly Efficient Neural Network for Image Classification. *arXiv preprint arXiv:2003.13678*.

Rosas-Lara, Mauro, Mendoza-Tello, Julio C., Flores, Aldrin, & Zumba-Acosta, Gema. 2022. A Convolutional Neural Network-Based Web Prototype to Support Melanoma Skin Cancer Detection. *Pages 1–7 of: 2022 Third International Conference on Information Systems and Software Technologies (ICI2ST)*.

Schlessinger, Daniel I., Chhor, Guillaume, Gevaert, Olivier, Swetter, Susan M., Ko, Justin, & Novoa, Roberto A. 2019. Artificial intelligence and dermatology: Opportunities, challenges, and future directions. *Seminars in Cutaneous Medicine and Surgery*, **38**(1), E31–E37.

Shete, Amit Sanjay, Rane, Aniket Sanjay, Gaikwad, Prajakta Sanjay, & Patil, Manasi Hanumantrao. 2021. Detection of Skin Cancer Using CNN Algorithm. *INTERNATIONAL JOURNAL OF ADVANCE SCIENTIFIC RESEARCH AND ENGINEERING TRENDS*, **6**(5), 1–5.

Sondermann, Wiebke, Zimmer, Lisa, Schadendorf, Dirk, Roesch, Alexander, Klode, Joachim, & Dissemond, Joachim. 2016. Initial misdiagnosis of melanoma located on the foot is associated with poorer prognosis. *Medicine (United States)*, **95**(7).

The Skin Cancer Foundation. 2023 (Mar).

Universidad EAFIT. 2017. *Reglamento de propiedad intelectual*.

Wirth, R., & Hipp, Jochen. 2000. CRISP-DM: Towards a standard process model for data mining. *Proceedings of the 4th International Conference on the Practical Applications of Knowledge Discovery and Data Mining*, 01.

Yaiza, Jiménez Martínez, Gloria, Ruiz Alcalá, María Belén, García Ortega, Elena, López Ruiz, Gema, Jiménez, Juan Antonio, Marchal, María Ángel, García Chaves, & Houria, Boulaiz. 2019. Melanoma cancer stem-like cells: Optimization method for culture, enrichment and maintenance. *Tissue and Cell*, **60**(June), 48–59.

Zhang, Ni, Cai, Yi Xin, Wang, Yong Yong, Tian, Yi Tao, Wang, Xiao Li, & Badami, Benjamin. 2020. Skin cancer diagnosis based on optimized convolutional neural network. *Artificial Intelligence in Medicine*, **102**(March 2019), 101756.