Arab American University – Palestine

Faculty of Graduate Studies

Skin Cancer Classification Using Deep Learning Techniques

By
Helal Abd-Alraouf Assaf

Supervisor
Prof. Mohammed Awad

This thesis was submitted in partial fulfillment of the requirements for the Master`s degree in

Data Science and Business Analytics

2022-2023

Skin Cancer Classification Using Deep Learning Techniques

By
Helal Abd-Alraouf Assaf

This thesis was defended successfully on 20/02/2023 and approved by:

Committee members	Signature
1. Supervisor: **Prof. Mohammed Awad**	
2. Internal Examiner: **Dr. Rami Yusif**	
3. External Examiner**: Dr. Rami Hadrob**	

Declaration

I**, Helal Assaf**, one of the students of the Faculty of Graduate Studies at the Arab American University, hereby declare that this thesis entitled "**Skin Cancer Classification Using Deep Learning Techniques**", is all my work and the resources that are used in this thesis (including the internet resources) have been referred to and properly acknowledged as required. I declare that I have fully understood the concept of plagiarism and I acknowledge that my thesis will be immediately rejected in case of including any type of plagiarism.

Name: Helal Assaf

Student ID: 201912657

Signature: 

Date: 20-2- 2023

Acknowledgments

First and foremost, I'd like to express my gratitude to my supervisor, Prof. Mohammed Awad, for allowing me to work with him on my thesis. Without his help, this project would not have been possible. He was always willing to give me advice on my work and recommendations when I was stuck on something. He was always open to meeting with me on short notice if I needed assistance. Next, I would like to thank both my parents and my wife for their support and love throughout my life. And finally, I would like to thank my university for opening the Data science major, which was a dream that came true to be one of the first students who enroll in this major.

Abstract

One of the most prevalent types of cancer in the world is skin cancer and Melanoma is widely regarded as the most common and potentially deadly skin cancer. Similar to all other types of cancer, early detection increases the likelihood of successful treatment; fortunately, with the advancement of AI in image processing and the availability of skin cancer historical datasets, it is possible to develop a classification model using deep learning algorithms to facilitate and have a more accurate skin cancer detection, this helps lab teams and simultaneously positively affects the length of the period until the disease is detected, in addition to its cost. By the usage of this system, early detection will save effort, time, and human lives. In this work, the CNN model and various transfer learning-based models, such as; ResNet50, VGG16, VGG19, EfficientNetB7, DenseNet169, Xception, and InceptionV3 are proposed. In addition, averaging ensemble approach is performed on various combinations of the models. This helps to increase the performance of cancer detection. The SIIM-ISIC Melanoma Classification Challenge dataset is used to train and test the proposed model. In this work, data augmentation is applied to the datasets to improve the model training process. We will implement augmentation in two phases; first, applying multiple techniques to the malignant class of images to increase the number of images in this class and solve the imbalance problem. Second, conducting random left and right flipping and up and down flipping augmentation methods on both categories of data to make a diverse set of images from existent images. The proposed method's performance is evaluated using well-known quantitative measures, such as; accuracy, precision, recall, and f1-score. The ensemble modeling of five of the best

models, which are; CNN, EfficientNetB7, VGG16, VGG19, and DenseNet169 outperforms all other models, with an accuracy of 93.79%, a precision of 94.18%, recall of 93.79%, f1 score of 94.77% and AUC score 93.8% on the unseen test dataset. The findings show that the proposed model is more efficient and reliable for automating skin cancer classification in comparison to previous works, shortens the recognition processes, and is valuable in its contribution to saving patients' lives.

Key Words: CNN, Ensemble Learning, Melanoma, Skin Cancer, Transfer Learning.

Contents

List of Tables

List of Figures

List of Abbreviation

Abbreviation	Explanation
DL	Deep Learning
ML	Machine Learning
ANN	Artificial Neural Network
CNN	Convolutional Neural Network
TL	Transfer Learning
VGG	Visual Geometry Group
ResNet	Residual Neural Network
VGG16	Visual Geometry Group with 16 layers
VGG19	Visual Geometry Group with 19 layers
ANN	Dense Convolutional Network
ROC	Receiver Operating Characteristic
AUC	Area Under Receiver
TP	True Positive
TN	True Negative
FP	False Positive
FN	False Negative
AUC	Area Under Receiver

Chapter 1

Introduction

1.1 Introduction

The skin is the largest organ in humans as it covers the outer surface of the human body functions as a protective barrier from the external environment, and constitutes the first line of defense against infections-causing microorganisms. In addition, it fulfills the function of regulating the body's temperature, providing sensitive information from the surrounding environment, excreting and absorbing substances, and synthesizing vitamin D [1]. Epidermal cells are primarily responsible for the production of a pigment called melanin. This pigment contributes to imparting color to the skin and absorbing harmful ultraviolet (UV) rays. The melanocytes are dendritic cells whose projections extend between the keratinocytes and transfer them to melanin granules. Once inside the granules, they clump together to form a veil protector over the core, towards the surface of the skin, which protects it from damage caused by different lights; from UV light to nuclear DNA. However, the melanocytes themselves are highly susceptible to getting damaged by UV radiation [1]. The increase in cancer rate is a major health problem around the world, among which one of the most common types of cancer is skin cancer. This increase in the rate of skin cancer patients is due to various factors, one of which is prolonged exposure to Ultraviolet (UV) radiation, which is a form of non-ionizing radiation that is emitted by the sun and artificial sources. Although this kind of radiation has some benefits, like stimulating the production of Vitamin D, it can cause health risks. Skin cancer is characterized by the development of abnormal cells, which divide, grow and spread uncontrollably in

any part of the body. They are different from normal cells because the normal ones divide and die over a period of time. Skin cancer can develop in any individual, both in men and women of any age and regardless of their skin color. The incidence of melanoma reported worldwide varies widely depending on regions. To prevent skin cancer, it is important to know the risk factors that are common and lead to its development. Some of the factors that could lead to the development of skin cancer are photo type, skin, and eye color, number of melanocytic nevi, presence of dysplastic nevi, and individual or family history of skin cancer. Moreover, exogenous factors type, degree of cumulative sun exposure, history of sunburn, and sun protection behavior are also part of the factors that could lead to skin cancer [2]. Skin cancer is one of the most common causes of cancer. Despite being the least common skin cancer, melanoma is liable for 75% of skin cancer deaths. Tumors can be one of two types: benign or noncancerous and malignant or cancerous as shown in figure 1.1.

Benign tumors grow slowly and rarely spread. Malignant tumors can spread throughout the body, infiltrate and decimate nearby healthy tissue, and grow rapidly. As tumors rapidly grow and spread throughout the body, they pose a threat to healthy tissues and have the potential to destroy them. In terms of mortality rates, skin cancer ranks among the highest in the world of having a high mortality rate [3].

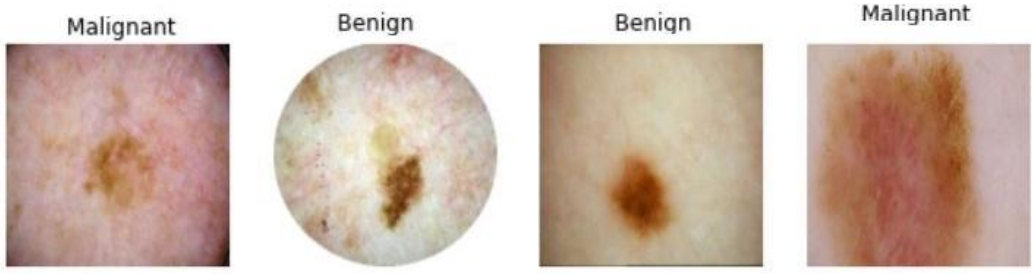

Figure 1.1: Skin Cancer types images

There are two types of skin cancer: melanoma and non-melanoma [4], where Nonmelanoma skin cancer can refer to any cancer that forms in the basal, squamous or Merkel cells of the skin and It rarely spreads to other parts of the skin and very slowly, the Melanoma is a cancer that develops in the skin's melanocytes and spread quickly.

If the lesions are caught early enough, the cure rate could reach 90%. Visual examination is difficult and leads to an improper investigation because of the high similarity between different forms of skin lesions. Therefore, the classification of skin lesions needs to be performed automatically. This group made use of computer vision and image processing methods [5], to upgrade a tool that is capable of detecting this skin disease in time, and with a high accuracy level.

In recent years, a multitude of skin cancer classification systems that make use of deep learning has been developed to classify images of skin tumors, such as Malignant-Melanoma (MM) and other forms of skin cancer [6]. According to some accounts, their precision was comparable to or even superior to that of dermatologists. In earlier reports, the targeted detection range encompassed everything from just malignant melanoma to the entirety of skin cancer. Clinical images and Dermoscopy images were the types of images utilized as data for machine learning. So far, there haven't been any published studies on the effectiveness of using clinical image data of pigmented skin lesions to train a neural network for the classification of skin cancers like MM and basal cell carcinoma. This is because there has been no such training done (BCC). While developing a system, it is essential to find out the proper endpoints as per the type of skin tumor that is going to be targeted along with the proper method of imaging that will be used. When new patients present themselves at a medical facility with skin

lesions as their primary complaint, in most of those cases, the patient's primary concern is not whether or not the lesions are malignant melanomas, but rather whether or not the lesions are skin cancers. Therefore, it is necessary to develop a system that can detect other skin tumors that have an appearance pigmented in a similar way to that of malignant melanoma [7]. Deep Learning (DL) is a specialized subset of Machine Learning (ML), which in turn is a subfield of Artificial Intelligence (AI). The main goal of AI is to provide a set of algorithms and techniques that can be used to solve problems that humans perform by themselves; intuitively and almost automatically, but are challenging for computers to solve. ML is applied for pattern recognition and Deep Learning is a class of ML algorithms inspired by the structure and functioning of the brain. Deep Learning tries to simulate the perception of objects by connecting artificial neurons or nodes within layers that could extract features from objects. This means that when DL is applied to image classification, it seeks to take a set of images and identify patterns that can be used to discriminate between various classes or objects. DL can perform the extraction of characteristics of an image automatically during the network training process [8]. For physicians, the main result used to identify skin cancer is the cancer tissue tests, which process will consume time. In this thesis, we proposed a DL methodology applying different DL models and various transfer learning-based models, such as; ResNet50, VGG16, VGG19, EfficientNetB7, DenseNet169, Xception, and InceptionV3, applied on SIIM-ISIC Melanoma dataset to train and test the models. On the other hand, an augmentation process of two phases; applying multiple techniques to the malignant class of images to increase the number of images in this class and solve the imbalance problem, and conducting random left and right flipping

and up and down flipping augmentation methods on both categories of data to make a diverse set of images from existent images to improve the model training process.

1.1 Research Problem

There were two main issues with previous computer-aided systems for dermatological picture classification. The lack of data is the first obstacle, and the imaging procedure is the second, in which dermoscopy is used to acquire images of the skin, and microscopes and biopsies are used to obtain images of other medical conditions. In the past, skin image classification needed substantial pre-processing, segmentation, and extraction procedures [9]. Effective skin cancer detection can be beneficial to millions of people. Early treatment can avoid severe problems caused by malignant tumors, but currently, there is no objective methodology that is used to accurately identify and diagnose such tumors. Applications for the diagnosis of malignant melanomas and other forms of skin cancer have already been made available for mobile phones in several nations [10]. On the other hand, skin cancer self-assessment apps have provided some trouble because there is not enough evidence to support the accuracy of their diagnosis. In addition, there is also a problem with their ability to recognize images, as there is a wide variety of angles, distances, as smartphones have become more widely used, they have become an appealing tool for health-related functions. For dermatology alone, hundreds of applications are available to download for both patients and providers. Many of the diverse capabilities of smartphones come from the different applications available. The number of applications available has been increasing with advancements in the image quality. and the quality of the Smartphone’s camera which needs to be taken into consideration. In

addition, the more difficult the shooting conditions are, the less accurate the shots will be. Therefore, it is required to carry out these steps and to consider them if you intend to take photographs of individuals using the camera on your cell phone.

Although deep learning allows automatic feature extraction and classification a, main challenge lies in data availability and legal issues. Deep learning performs better with abundance of data and currently, it is difficult to obtain abundant of medical data. The biggest impediments are the issues of data availability and the need to address legal issues related to patient's privacy protection.

Data scarcity, data protection, confidentiality and other legal issues are barrier to the training of an efficient deep learning model.

To address the problem of limited data, different data augmentation methods are available. These augmentation methods enclose different techniques to enhance the size and quality of data to facilitate training of deep learning algorithms [10]. Traditional data augmentation techniques in medical imaging use transformation like scaling, rotation, flipping, translation, etc., to augment samples.

These days, huge advancements have been innovated in artificial intelligence, especially in machine learning and deep learning fields. Researchers find a lot of innovative methods for classifying images more precisely and this has diverted their approach toward the medical field which has image data types, especially cancer detection. As data on skin cancer is prepared in image type, this domain has become one of the favorable domains for researchers. Considering to importance of skin cancer detection in the early stage in the lives of people. This thesis takes more attention to

finding the best CNN architecture for skin cancer classification and also use ensemble technique to cover their weakness, then the problem of this thesis defines as follows:

1.2.1 Variability of models

CNN architectures are the most powerful models for convolving through images and extracting their features and classifying them. CNN need huge amounts of data to perform accurately and also quantity of data impacts their results. Researchers did many developments in CNNs to structures, residual and inception networks are two of them. They apply their model to a huge ImageNet dataset with millions of images and thousands of classes to verify their proposed idea. Nowadays, in image classification, it is possible to develop a CNN architecture from scratch and train it on specific data which needs more powerful hardware and time, or used previous state-of-the-art pre-trained architectures. As different data in the case of quality and quantity affect the results of both groups, selecting the best one for skin cancer detection is the first problem that should come into consideration.

1.2.2 Variability of architectures in feature extraction

CNN architectures same as all deep learning algorithms initiate with random weights and then try to improve these weights during training, then in different pieces of training, it is possible to get different results to regard to performance. Fortunately, this problem can be reduced with the help of transfer learning when we train our data on pre-trained architectures and use their weights which are learned on huge datasets. on the other hand, each architecture utilizes a different method for extracting images and

highlighting a different part of the image while training. Extracting features is the most of image classification and each architecture has a different capability for it. Thus, the other problem is finding a method to combine the strength of all architecture and improve final results.

1.3 Research Objectives

The research aims to create an image classification pipeline that will enable early objective detection of benign and malignant skin lesions. In recent times, Convolutional Neural Network (CNN) has widely been applied with different classification architectures to classify skin cancer lesions and they significantly exceeded skilled healthcare professionals' performance. Many techniques have improved the performance of those models, like increasing the deepness of the network and utilizing transfer learning convolutional networks such as VGG-16 [11]. Here by implementing various state-of-the-art architectures and ensemble approaches, we will work on finding the best model for skin cancer detection in terms of speed, cost, and accuracy to help specialists during the detection process and save the lives of people respectively.

This work tries to develop a precise model for detecting skin cancer from microscopy images, which aim is to shorten the duration of disease detection. In this regard, the summary of the objectives is as follows:

1. To propose the best DL architecture for skin cancer classification.
2. To compare the performance of CNN with other pre-trained architectures.
3. To assess the effect of the ensemble approach on performance.

1.4 Research Significance

Skin cancer is the most prevalent cause of cancer. Even after being the least common type of skin cancer, melanoma accounts for 75% of skin-cancer deaths. Over 100,000 new cases of melanoma were identified in 2020, as per the American Cancer Society, and it was expected to kill nearly 7,000 people in the next year. Early and accurate detection, similar to other cancers, can improve treatment efficiency, and possibly, it can be done with the help of data science. Melanoma is a deadly disease, but if detected early enough, it has a high chance of recovery by performing minor surgery. Dermatologists' diagnostic accuracy will improve with image analysis methods that automate melanoma diagnosis [12]. Millions of people could benefit from improved melanoma detection. Thus, the expected outcome of the proposed research project is a fully reliable and efficient system that can classify skin diseases as benign or malignant. Melanoma is not well known because it is one of the rarest skin cancers, so patients do not see dermatologists in the early stages. The average time from the first awareness of having an issue to the first visit is from two months to nine and a half months. Because of the long period between being aware of having an issue the and first visit, the rate of late-stage cancer is high [12]. To overcome the prognosis of skin cancer, shortening the time from awareness to a visit is essential. Quick detection of skin cancer aided by a reliable detection system is an option that dermatologists may have missed.

1.5 Thesis Outline

The remaining sections of the thesis are organized as follows:

- Chapter 2 presents the background of the research topic, where the work related to the general search topic and the applied dataset are presented, in addition to a deep learning algorithms overview and an introduction of the intended work.
- Chapter 3 presents the methodology, starting with the dataset description, preprocessing of the data, the applied DL models, and finally the improved transfer DL model, lastly, it will introduce the model development and building.
- Chapter 4 presents the experiments and the results of the dataset with different models. The best model will be shown through a comparison between the different models. Additionally, this chapter will discuss the work challenge and limitations.
- Chapter 5 includes the conclusion and further work recommendations.

1.6 Conclusion

In this chapter, we discussed the main idea of the thesis, and our aim to explore the possible options to create an image classification pipeline that will enable early objective detection of benign and malignant skin lesions. Moreover, the proposed solution uses Deep Learning Techniques by using Convolutional Neural Networks (CNN) and Transfer Learning Algorithms. We also discussed how this project would help in saving people’s life by early detection of skin cancer.

Chapter 2

Background

2.1 Background

One of the leading causes of human death is cancer. World Health Organization projections show that by 2030, cancer will have killed 13.1 million people worldwide [13]. The United States has the highest rate of skin cancer cases worldwide. Experts estimate that 20% of Americans will be diagnosed with skin cancer at some point in their lives [14], in some cases, skin cancer can be treated and cured, which is why early detection is crucial in saving lives. There is a need for further research on human skin and various skin malignancies to gain further insight and to increase the chance of earlier detection and throw diagnosis methods.

In the last ten years, skin cancer has become one of the fastest-growing cancers. Because the skin is the largest organ in the body, it seems that this contributes to it being one of the most commonly affected human parts by cancer. Melanoma, the skin cancer that accounts for only 1% of all occurrences of cancer, has a much higher mortality rate. Melanoma is dangerous, uncommon, and ultimately fatal. Automatic melanoma recognition from dermoscopy images remains challenging despite dermoscopy's ability to improve the visual impression of a skin lesion. To begin, it is challenging to check segment lesion areas accurately due to the poor contrast between skin lesions and normal skin regions. Second, it might be challenging to tell a melanoma lesion from a non-melanoma lesion since they can look very similar. Third, melanoma can occur in a wide range of colors, textures, and shapes because of

individual differences in skin characteristics like pigmentation, hair, and veins. Undertreating a malignant skin lesion owing to a false-negative diagnosis is much more dangerous than over-treating a benign lesion due to a false-positive diagnosis (unnecessarily costly). UV exposure is a contributing factor in the rising incidence rate of melanoma, which now ranks as the 19th most common form of cancer with an annual increase of 53%. This type of cancer is more dangerous than squamous and basal because it metastasizes rapidly. Therefore, saving the patient's life requires a therapy approach that detects skin cancer quickly and accurately. The importance of early detection of skin cancer in terms of therapy cannot be overstated. The normal method for diagnosing skin cancer that doctors follow is to do a biopsy to determine if a suspicious mole on the skin is malignant or not. This is a difficult, slow, and lengthy procedure. Diagnosing skin cancer with the help of a computer is more convenient, less costly, and much faster than ever before. Imaging modalities, such as x-rays, MRI, CT scans, ultrasound, and CAT scans, are all examples of medical imaging techniques that provide visual representations of the body's inside and exterior for diagnostic purposes [15]. Computer tomography (CT) scans, X-rays, and ‘magnetic resonance imaging’ (MRI) are commonly used to investigate anatomical structures deep within a living organism. Digital chemiluminescence dermatoscopic is used in dermatoscopy; an imaging modality that captures images of the skin using polarized light. Dermatoscopy improved the diagnostic accuracy of melanoma detection by 10-27 percentage points, according to studies, compared to that of human experts who previously relied solely on their specialized training and experience when making diagnoses.

In a study undertaken by researchers from Germany, the United States, and France [16], that follows the process that Machine Learning describes, the researchers used an AI system to detect and categories malignant and benign skin lesions by feeding more than 100,000 photos of skin lesions into a Convolutional Neural Network (CNN). The AI system's classification accuracy was 95%, whereas that of human specialists was only 86%. One area of artificial intelligence where this occurs automatically is computer vision, which uses visual representations of data as input.

The AI system employs a suite of methods for algorithmic learning through repetition. The researchers developed a system for automatically recognizing melanoma [17]; this system is called Complete Convolutional Residual Network (FCRN) and it was meant to be used in the ISBI (2016) dataset to classify lesions according to their unique characteristics. After identifying the input images, the researchers utilized a Deep Residual Network (DRN) to distinguish between melanoma and non-melanoma lesions in the segregated lesion areas. In contrast, many DCNN methods have been tried and tested such as VGG-16, Google Net, FCRN-38, RCRN-50, and FCRN-101. Keeping this context in mind, the goal of this chapter is to apply transfer learning to the classification and identification of skin cancer. To simplify feature extraction, detection, and classification from raw data, the deep learning method incorporated representation learning. In the field of computer vision, Convolutional Neural Networks (CNNs) are the most widely used deep learning technology. CNNs are capable of autonomous learning and feature extraction with sufficient and appropriate training. CNN is used in computer vision applications because it can classify images based on the priority it assigns to various aspects in the input image by updating the

weights and biases associated with those features [18]. However, we will propose a (CNN) model using transfer learning throughout the research paper. The proposed model will be beneficial for classifying skin cancer in humans. In the future, it will help decrease the rate of death cases due to this disease in the whole world.

2.2 Literature Review

In this section of the research paper, we will present and discuss the literature review related to our research subject; skin cancer. The researcher has reviewed different articles on this specific topic from different resources. Simon Kalouche published a paper titled Vision-Based Classification of Skin Cancer Using Deep Learning in 2016 [19]. His study proposed detecting the presence of skin cancer, Melanoma in specific, by the use of deep learning algorithms. He suggested capturing images of skin lesions with a standard camera, to lose them like what happens each year due to late-stage diagnoses. Therefore, it is necessary to make melanoma screening and diagnosis methods less expensive, faster, easier, and more accessible. The goal of this study was the development of a computer-vision-based machine learning tool that has low cost and works fast to help doctors and patients track and classify suspicious skin lesions as benign or malignant with adequate accuracy, by using only a cell phone camera. The data set was trained on three different learning models, each with increasing classification accuracy. One of the three models used was logistic regression, which is a deep neural network, and a fine-tuned, pre-trained VGG16 Convolutional Neural Network (CNN). The initial results show that, by using fine-tuned VGG-16 CNN, the developed algorithm could, with an accuracy of 70%, segment moles from images and classify skin lesions as melanoma with 78% balanced

accuracy. In 2019, a paper called Real-Time Recognition of Malignant Skin Lesions using Ensemble Modeling [20] was published. This paper, to correctly identify malignant skin lesions, proposes a hybrid method that uses an ensemble of deep predictive models (Feedforward Neural Network, Image Histogram Classification using KNN, LeNet-5 CNN architecture, and VGG-11 CNN architecture). From the ISIC archive (International Skin Imaging Collaboration), which is an open-source organization dedicated to furthering research into skin cancer detection, the images dataset was extracted which contains a variety of images of benign (nevus, Seborrheic Keratosis) and malignant (melanoma, carcinoma) skin lesions. The challenge is to correctly segment the skin lesion in the early stages, even with the presence of hairy images. For Ensemble modeling to be useful the classifiers used must be uncorrelated; therefore, the researchers created an ensemble of three unrelated models and built the final model using stacking. They measured each model's accuracy, sensitivity, and specificity using 10-fold cross-validation. In 2018[21], Kartikeya Agarwal and Tismeet Singh performed training on multiple transfer learning models such as DenseNet, ResNet, XceptionNet, and MobileNet using the ISIC dataset. The highest accuracy was 86.5% for ResNet followed by 86% for DenseNet, 83% for XceptionNet, and the least accurate one was MobileNet with 80.8%. In 2020, a scientific paper was published in the Journal of Scientific & Industrial Research Vol. 79 by five researchers, and it is called Skin Cancer Classification using Convolutional Capsule Network (CapsNet) [22].

The main problem of the current skin cancer classification models is that they do not consider the important spatial relations between features. They are only able to

effectively classify when there are certain features in the test data, and they ignore the relative spatial relationship with one another, which results in false negatives. Moreover, another reason that results in getting false positives is that they lack rotational invariance, which means that the same legion can be assigned to different classes when viewed from different angles. To address the aforementioned issue, the Capsule Network (CapsNet) is used. The Capsule Network architecture consists of many convolution layers with one capsule layer as the final layer. Therefore, in the proposed work, CapsNet architecture is performed for skin cancer classification, because it works well with high-dimensional hyperspectral skin images. However, if the experimental validation datasets are small, the training will not be effective. Precision, recall, f1-score, and support are used to measure the proposed network's performance. This network outperformed all existing skin cancer classification techniques with 92% precision, 91% recall, 91% f1-score, and 93.8% support. In 2022, Bill Cassidy et al. summarized the usage of the ISIC data set images and presented an analysis of early release throughout 2016 – 2020. The proposed benchmark results analysis shows good levels of melanoma prediction with an AUC of 0.80 for the best-performance model [23]. In 2021, Yiming Zhang and Chong Wang used ISIC 2020 data set in their research. Their implemented method has achieved superior performance over the other deep learning approaches. The implemented DenseNet model gained 0.925 with the AUC metric, which is higher than the previous approaches that applied VGG and ResNet backbone [24]. In 2020, Qishan Ha et al. used the SIIM-ISIC Melanoma Classification Challenge dataset. They proposed an ensemble of Convolutional neural network models with different backbones and input

sizes. Their implementation and submission scores are 0.9600 AUC on cross-validation and 0.9490 AUC on the private leader board [25]. In 2021, Ismail Elansary et al. used ISIC 2020. In this suggested model for melanoma classification EfficientNet –B6 has been used, which is its deep Convolutional Neural networks CNN for classifying a patient's skin lesion into malignant or benign with a competitive system accuracy ratio of 97.84% in comparison to other models [26], in this paper, the author used the SIIC-ISIC 2020 Challenge data set and took a two-step two-pronged approach to a binary classification of images of skin lesions as either melanoma or not melanoma. The first step involves using a U-Net trained on the HAM10000 dataset to produce masks of the images of the ISIC2020 dataset before CNN. They implement a VGG model, a Siamese model adopted from Messina et al, and an Efficient Net model that was used in the winning solution to achieve an AUC of 0.9442 [20]. They concatenate one-hot-encoded metadata and the output of the CNN model and put it into a final two-layer dense network. Our best CNN architecture is the Siamese model, combined with the rest of the pipeline achieved a test AUC score of 0.831.

In 2021, Satin Jain et al. used transfer learning models such as VGG19, InceptionV3, InceptionResNetV2, ResNet50, Xception, and MobileNet on the HAM10000 dataset. It is inferred that Xception Net outperforms the rest of the transfer learning nets used for the study, with an accuracy of 90.48% and VGG19 produced the least accuracy of 66.36% [27]. In 2021, Shubhajoy Das and Debashis Das used SIIM ISIC 2020 data set in their research. They implemented a feed-forward neural network with fully connected dense layers. A Gray Level Co-occurrence matrix is used as feature

extraction, which gives an AUC score of 0.63 on the private leader of the SIIM ISIC 2020 Challenge [28].

In 2021, Giraldo et al. proposed the SIIM-ISIC melanoma classification challenge. They implemented a multi-class, multi-modal classification model using images and metadata. They tested both binary and multi-class image-only models and a binary multi-modal model [29]. Also, with the usage of the available metadata, they achieved AUC values of 0.95 and F1 of 0.71 for the validation data. In 2021, ten researchers introduced a paper called Skin Cancer Detection: A Review Using Deep Learning Techniques [30], This paper presents a detailed systematic review of deep learning techniques for the early detection of skin cancer. This study reviewed various research papers from reputable journals on related works. To ease the comprehension of the research findings, they are presented in the form of tools, graphs, tables, techniques, and frameworks. To identify and categorize the best available approaches to skin cancer detection using neural networks (NNs) this systematic review was done. They have collected and analyzed existing studies using predefined evaluation criteria as part of their systematic review. Such reviews help in determining what is already known in the relevant domain of study and in defining the review framework. The study started with an overall plan and was followed by a systematic literature review. The plan consisted of three layers: a planning layer; data selection, an evaluation layer; a results-generation, and a conclusion layer. To extract information that is relevant to NN techniques for skin cancer detection, they conducted their initial search on well-reputed search engines such as IEEE Xplore, ACM, Springer as well as Google Scholar. In 2020, S. Sasikala et al. used a dataset of 3700 clinical images and its performance was

tested with 660 images to train CNN models such as RestNet34, ResNet50, VGG16, and AlexNet. ResNet34 produced the highest accuracy of 88.4% and the lowest accuracy being AlexNet with 65.3% [31]. In 2022, Khalil Aljohani et al. published research named 'Automatic Classification of Melanoma Skin Cancer with Deep Convolutional Neural Networks'. They used various CNN architectures, including DenseNet201, MobileNetV2, ResNet50V2, ResNet152V2, Xception, VGG16, VGG19, and GoogleNet on a dataset consisting of 7146 images. The experimental results showed that GoogleNet can obtain the highest performance accuracy on both the training and test sets (74.91% and 76.08%, respectively) [32].

2.3 Related Works for the Applied Data Set

The dataset includes 33,126 dermoscopy's training images of distinct benign and malignant skin lesions from more than 2,000 patients. A unique patient identifier is used to link each patient with his/her image. To confirm the malignant diagnoses, histopathology was used, and to confirm the benign diagnoses; expert agreement, longitudinal follow-up, and histopathology were used. In 2022, Ranpreet Kaur et al. used a dataset related to malignant vs. benign melanoma. The proposed LCNet model was inspired by a deep convolutional network for skin classification, and the implemented LCNet achieved average accuracy, precession, and recall of 81.41%, 81.88%, and 81.30% on ISIC 2016, 88.23%, and 87.86% on ISIC 2017 and of 90.48%, 90.39% and 90.48% on ISIC 2020 [33].

In 2022, Ayushi Kumar, and Avimanyou Vatsa used a dataset for two classes (Malignant and Benign). They compared the performance of CNN, RNN, and XG-Boost, and the results were that the accuracy of VGG16 architecture was 89.6%, which

is the best architecture, among the seven architectures, for CNN. Furthermore, the RNN's bidirectional architecture outperformed the other four RNN architectures with an accuracy of 95.96%. And the XG-Boost method had an accuracy of 97.22% [34]. In 2021**,** Mohammed Rakeibul Hasan et al. used a two-class data set (Malignant and Benign) and implemented different approaches. The researchers gained accurate results for the VGG16 model with an accuracy rate of 93.18%, SVM 83.48%, ResNet 84.39%, sequential model_1 74.24%, sequential model_2 77.00%, and sequential model_3 84.09%. VGG16 is the one that had the highest accuracy rate; 93.18% [35]. In 2021, Sagar Karkai et al. proposed a method that has been evaluated on the SIIM-ISIC Melanoma Classification 2020, and the best ensemble model achieved 0.9411 areas under the ROC Curve on hold out test data [36].

In 2018, Md Ashraful Alam Milton employed models like PNASNet-5-Large, InceptionResNetV2, SENet154, and InceptionV4 on the ISIC dataset. The InceptionV4 model reached 67% accuracy and the highest accuracy was achieved by PNASNet-5-Large with 76% [37]. In 2021, a paper named 'Convolutional Neural Network Architecture for Skin Cancer Diagnosis' was published. They used CNN architecture on two dataset repositories for this work, the first set with the ISBI nomenclature and the second with the IMD nomenclature. They gained an accuracy of about 88.87% with the ISIC dataset. [38]. In 2022, a research work called 'Skin lesion classification of dermoscopic images using machine learning and convolutional neural network' used the HAM10000 dataset. The research compares the performance of machine learning models such as Decision Tree (DT), Random Forest (RF), Support Vector Machine (SVM), K-Nearest Neighbor (KNN), Logistic Regression (LR), Gaussian Naïve Bayes

(NB), and Linear Discriminant Analysis (LDA) with Convolutional Neural Networks. The highest accuracy among the machine learning models was given by random forest with a 58.57% while CNN produced a much higher accuracy of 95.18% [39]. In 2020, a paper named 'Convolutional Neural Networks for Classifying Melanoma Images' was published. They used the ISIC dataset with transfer learning models such as ResNet50, InceptionV3, Inception ResNet V2, MobileNet, and Densenet169. The highest accuracy by the transfer learning models was given by ResNet5 wh a 93.5% [40].

In 2022, Jaisakthi SM et al. proposed a digital diagnosis of skin lesions using EfficientNet architecture on the ISIC dataset with an AUC score of 91.74%, EfficientNet-B6 was concluded as the most suitable classifier for classifying melanoma and non-melanoma [41]. In 2021, Marud Hossain et al. researched skin cancer classification using ensembles of multi-scale EfficientNet models. The ISIC2019 dataset was used. The ensemble of different backbones of EfficientNet such as B0, B3, B4, B5, and B6 produced an accuracy of 89.30% [42]. In 2022, Flavia et al. conducted a review of machine-learning approaches for skin cancer classification from dermoscopic images. Various public datasets such as the HAM10000, ISIC archive, PH2, and MedNode were studied. Machine learning approaches such as Decision trees, SVM, KNN, and ANN and Deep learning approaches such as custom and pre-trained CNNs were reviewed. It was seen that the most used dataset was the ISIC archive, the most widely used machine learning algorithm was SVM, and pre-trained was widely used in deep learning models [43].

In 2022, Atheer Bassel et al. proposed a hybrid deep-learning approach for classifying malignant and benign skin cancers. The ISIC dataset was used on a stacked CV system

and compared its performance with models such as SVM, KNN, Random Forest, and Naive Bayes. For feature selection, they used ResNet50, Xception, and VGG16 models. It was seen that the proposed Stacking CV approach showed the highest accuracy of 90.9% among all the models with the Xception feature selector [44]. In 2017, Romero Lopez et al. performed skin lesion classification on the ISIC archive using deep learning techniques. They proposed three methods: Training from scratch, ConvNet as a feature extractor, and Fine-tuning using ConvNet. During the testing evaluation, ConvNet as a feature extractor method gave the highest accuracy of 81.33% [45].

2.4 Definitions of Deep Learning Algorithms

In the struggle against skin cancer, deep neural networks are a crucial tool. They are made up of a group of nodes that communicate with one another. In terms of the connectivity of their neurons and their architecture, they are very comparable to that of the human brain. The term "deep learning" is used to describe a specific type of machine learning based on the idea that higher-level concepts are defined by lower-lever ones, and the same lower-lever notions can serve to define multiple higher-lever concepts. Understanding visual, auditory, and textual input is made possible via deep learning, which involves learning at numerous representational and abstractive levels. Multilayer perceptron, which has more hidden layers than the average neural network, is an example of a Deep Learning structure [35]. As opposed to traditional machine learning methods, deep learning may directly extract relevant characteristics from images, text, and sound in supervised and/or unsupervised approaches.

2.4.1 Convolutional Neural Network (CNN)

The Convolutional Neural Network algorithm is a multilayer perceptron tailored for the recognition of two-dimensional visual data. Input, convolution, sample, and output layers are always present. Moreover, in deep network architecture, both the convolution layer and the sample layer may have more than one iteration. Most commonly, CNN is employed to detect invariants of distortion in two-dimensional graphics, such as displacement and zoom. When we utilize CNN, the feature detection layer learns from the training data without the need for any explicit feature extraction. CNN algorithm is also used to classify skin cancer. Along with this, transfer learning models are built for classifying skin cancer images containing CNN layers. These models help to decide whether the skin cancer is malignant or benign.

For the CNN algorithm, convolution and sampling are the two most important steps.

i. Convolution process:

De-convolution of the input picture using a trainable filter Fx (the first stage is the input image, the input of the post-convolution is the image features of the current layer and is named; Feature Map), followed by the addition of a bias bx, yields a convolution layer Cx.

ii. Sampling process:

After each neighborhood's **n** pixels have been pooled into a single pixel, the resulting pixel is then subjected to a scalar weighting of Wx + 1, given a bias of bx + 1, and activated to generate a narrow n-times feature map Sx + 1. CNN is excellent at solving facial recognition issues, however training such a network requires a large number of annotated photos.

2.4.2 Transfer Learning

In machine learning, the term "***Transfer Learning***" refers to the process of adapting pre-existing models to address an unanticipated problem [46]. Contrary to popular belief, transfer learning is not an algorithm but, rather a method or strategy employed throughout the training process. Whenever training a system to tackle a new task would require a large investment of time and energy, researchers and developers frequently turn to transfer learning for machine learning. The procedure involves adapting a portion of an existing machine learning model to address an unrelated problem. Generalization is an essential component of transfer learning. A good option for transfer learning is a machine learning model that can pick out one specific subject from among several photos. There are many ways in which the process of creating machine learning models might benefit from transfer learning [19]. The primary advantages of transfer learning are the time and money it saves when training new models. Furthermore, it can aid in model training when only unlabeled datasets are available, as the majority of the model will have already been pre-trained. By combining insights from multiple models, developers can better hone in on an optimal approach to a problem. An improved model can be created by combining the information from two or more models. The method supports iterative model construction “C SKIN LESION/CANCER DETECTION USING DEEP LEARNING” published in 2020 [47]. Deep learning is an artificial intelligence operation that creates patterns for decision-making and organizes data by mimicking the way the human brain does that. Convolutional neural networks are the artificial neural networks that

most of the modern deep learning models are built on. In this paper, we developed a unique deep-learning architecture that focuses on the timely evaluation of skin cancer.

2.5 Discussion

In previous papers, researchers try to implement various CNN architectures, either from scratch or by employing some pre-trained architectures to find the best model with high performance for skin cancer classification. For instance, in 2022, Ranpreet Kaur et al. proposed LCNet architecture and assessed it on different versions of ISIC. Their best accuracy was on ISIC2020 which has more images as compared to the previous version. In 2021, Satin Jain et al. used pre-trained architectures such as VGG19, InceptionV3, InceptionResNetV2, ResNet50, Xception, and MobileNet and compared them together; they could obtain the highest accuracy with VGG19. Moreover, in 2018, Md Milton tried PNASNet-5-Large, InceptionResNetV2, SENet154, and InceptionV4 on the ISIC dataset and the highest result belongs to PNASNet-5-Large. In our research, we try to select the most common architectures and propose the best one regard to their accuracy on the last version of ISIC. Also, we implemented CNN architecture from scratch, thus it is possible to compare the performance of CNN with pre-trained architectures. Initiate random weights of CNN and also the existence of many parameters for tuning while training the model, then applying the ensemble approach sometimes can be helpful as it combines the capabilities of architectures to increase the performance. There are different methods in ensemble learning which divide into 3 major groups; boosting, stacking, and a mixture of them. However, the ensemble approach was rarely used in previous papers but, at any time it was used; it improved the classification. In 2020, Qishan Ha et al. used a

stacking ensemble for combining some CNN with non-related architectures and they could increase their performance. Dr. Ameri 2020 also combined the output of three different models which increased the certainty of their final outputs.

2.6 Intended Work

As per the objective of this research paper; classifying Skin Cancer, first, we have to collect a dataset. The methodology begins with the collection of datasets. The publicly available skin cancer classification dataset was collected from Kaggle. The dataset contains 33126 images of benign and malignant classes. The dataset contains 32542 benign images and 584 malignant images.

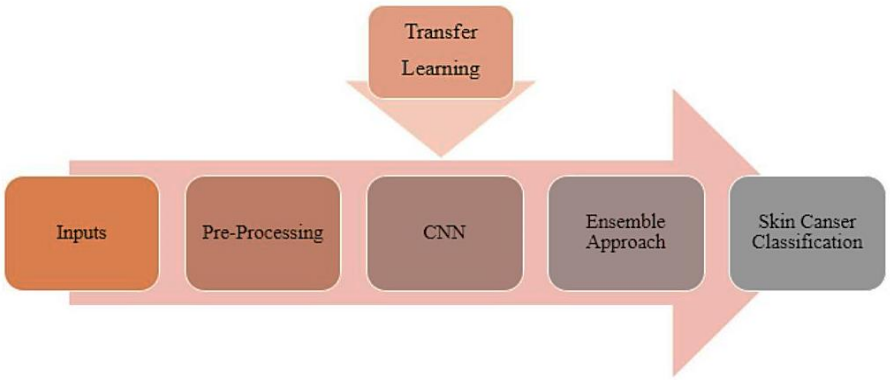

Figure 2.1: Overall System Architecture

From this figure (Overall Architecture Approach), we can see different steps to complete the report of skin cancer classification. According to the figure, we first input data and then start pre-processing. After that, we will build a CNN model to classify skin cancer and use transfer learning architecture. Finally, we will apply the ensemble technique to increase the performance of all implemented models.

2.7 Conclusion

In this chapter, we introduced previous work of multiple studies that used Deep Learning techniques in building classification models for the early detection of skin lesions. We also introduced definitions for two deep learning techniques; CNN and Transfer Learning Algorithms. Finally, we introduced the intended work that uses CNN, transfer learning, and ensemble approach in building a classification model to classify skin cancer in an early stage. The proposed system depends on training the CNN from the scratch, as well as using pre-trained models, and then using the technique of the ensample approach between them, as we try through this combination of the three techniques used to increase the accuracy of diagnosis, which helps in early detection of the disease, which facilitates the work of doctors in providing treatment required at the right time.

Chapter 3

Methodology

3.1 Introduction

Skin cancer is the out-of-control growth of abnormal cells in the epidermis, the outermost skin layer, caused by unrepaired DNA damage that triggers mutations. These mutations lead the skin cells to multiply rapidly and form malignant tumors. there are two main types of skin cancer, non-melanoma which a group of cancers that slowly develop in the upper layers of the skin. and the second one is melanoma which can appear in any part of the skin such as under fingernails and palms. Fortunately, as an early diagnosis rises the chance of successful treatment, and considering the development of AI in the field of image processing and the existence of skin cancer historical datasets, it is possible to develop a prediction application with the help of deep learning algorithms to facilitate and make the cancer detection process more accurate. In this research, we will compare various types of image processing models such as CNN and transfer learning architecture to find the best one.

The structure of this chapter will be in the following order: first, we will address the design and the procedure of this work which includes an explanation of the dataset and its preparation, as well as, the model selection process. Secondly, we will introduce the dataset, its features, and the weaknesses that need more attention before using it in modeling. Finally, we will explain the proposed methodology which includes: first, preparing data by common image preprocessing methods which will help us in obtaining high-quality input. Second, presenting the augmentation techniques that will

be used to handle the imbalance issue of the dataset. Third, we will apply CNN architecture. Fourth, we will test the transfer learning which we plan during our investigation, and finally, we will ensemble the approach that will be used to t a higher performance.

3.2 Research Design and Procedures

Skin cancers appear when mutations occur in the skin cells' DNA and cells start growing fast and uncontrollably. It begins from the epidermis which is a protective cover of the skin cells and reproduces continuously. If any abnormal changes emerge, three major techniques for checkups are done, which include: clinical images, dermoscopy images, and histopathological images. Clinical images are mostly taken by mobile and are used for remote diagnosis, while dermoscopy and histopathological images are taken by medical professionals and most of the time is used for identifying the level of cancer. There are various public datasets related to each method and some of the common and most used ones are PH, MED-NODE, HAM10000, Derm7pt, BCN20000, and ISIC [47]. Referring to table 3.1 which presents details the ISIC dataset has more than 32000 images and is bigger as compared to others; therefore, we chose it for this study. After verifying the dataset, we recognized that it is imbalanced and it has fewer samples in the malignant group. This weakness exists not only in this dataset but also for all other types of cancer with the same type of data. Therefore, conducting one of the handlings of imbalanced data is mandatory. Fortunately, we have vast techniques of augmentation to apply, such as; rotating, flipping, zooming, etc. and we can use some of them for tackling this problem and make data with similar features with the help of previous data, thus, we can start the modeling phase

Table 0.1: Skin Cancer Datasets

Dataset	No. of images	Modality of images	No. of lesion types	Image format	Published year
PH^2	200	Dermoscopic	3	.bmp	2013
MED-NODE	170	Macroscopic	2	.jpg	2015
HAM10000	10,015	Dermoscopic	8	.jpg	2018
Derm7pt	2,000	Dermoscopic Structured data	15	.jpg	2018
BCN20000	19,424	Dermoscopic	9	.jpg	2019
ISIC Archive	>13,000	Dermoscopic	9	.jpg, DICOM	2016–2020

Generally, medical image verification is done by lab specialists who take a long time until they come up with results [48]. The level of experience and accuracy of the lab experts will directly impact the correct diagnosis. It is clear, in this domain that any mistake is dangerous because it may threaten the life of a human, however, humans do make mistakes and mistakes' occurrence probability will never be zero, which is why developing a CAD system will be helpful. After developing deep learning, especially convolutional neuron networks (CNN) [49], researchers started utilizing them for image-based disease detection such as; skin cancer. Today we have many state-of-the-art CNN-based architectures with high performance like ResNet and inception in image processing cases. In this study, we will develop a CNN-based network from scratch and employ some of those pre-trained architectures to compare their results on the dataset and suggest an accurate CAD system that can predict dermoscopy image class which will lead to the diagnosis of skin cancer. For evaluation of our model's various metrics, such as; accuracy, precision, recall, and F1 scores will be calculated.

Furthermore, in medical problems confusion matrix, getting the number of wrong predictions will be vital. Finally, to control the training process in case of overfitting, we will compute and plot AUC/ROC curve. Due to the implementation of a few architectures and training all of them with this data, performing ensemble techniques on the result of those models will be effortless and we decide to apply it to models' predictions to cover the disadvantages of the models and improve performance.

3.3 Data Set

The dataset contains 33126 images which are collected from 2000 patients. The metadata of these images are in an SCV file and include information about patient ID, sex, age, and general anatomic site per image. This dataset is divided into two main parts: benign tumor with 32542 samples and malignant tumor with 584 samples. Benign tumors grow slowly and remain localized while malignant tumors are invasive and can destroy adjacent structures, spread to other areas, and could cause metastases which usually lead to the death of the patient. The dataset was generated by the International Skin Imaging Collaboration (ISIC) and images are from the following sources: Hospital Clinic de Barcelona, Medical University of Vienna, Memorial Sloan Kettering Cancer Center, Melanoma Institute Australia, University of Queensland, and the University of Athens Medical School. The data set was curated for the SIIM-ISIC Melanoma Classification Challenge hosted by Kaggle during the summer of 2020 which is accessible from this link: https://challenge2020.isic-archive.com/

3.4 Proposed Model

3.4.1 Data Set Exploration

First of all, we did data exploration to show the features of the data set and prepare it for the proposed system; firstly, we set up the environment using Google Colab and Google Driver. We uploaded all the images and metadata files to google driver and set up the GOOGLE COLAB to be mounted with Google Drive. After removing the duplicate and null values by applying NumPy library techniques via python environment, we got 32531 images for 2051 patients,

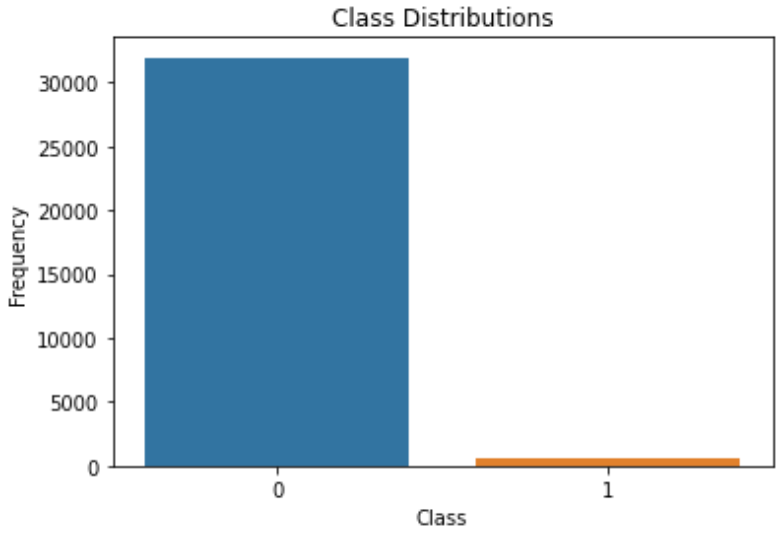

Figure 0.1: Class Distributions (Benign & Malignant)

As shown in the above plot, we see that the data is heavily imbalanced with 31956 samples containing benign data (class 0) and 575 samples containing malignant data (class 1). The proposed model is based on deep learning techniques. Our data set should be a balanced label to get better model performance. This obliged us to use data augmentation to balance the data set label, which will be discussed in more detail in the next section. For an imbalanced dataset, mostly various resampling methods such as over-sampling, under-sampling or hybrid are used to balance data, in image

classification selecting the suitable method is important and augmentation is common, especially in the case of the CNN algorithm which the dimension of data is important, it makes a more robust model that can generalize better in prediction. The cost-sensitive method focuses on changing the composition of training data which can be changing the distribution of them or using weighting cost proportionate for calculating loss, changing the distribution of data for images is not a good idea because may create noise in images, and impact negatively of performance of the model. SMOTE choose a random minority observation, then by k nearest neighbors try to find the next sample which is near to the first one and finally creates a new sample by randomly interpolating these two samples and generating a value between 0 and 1. For implementation, the image shape should change to two dimensions. The problem is during random interpolation, it is possible some noise adds to the created data and this noise reduces the final performance of the model. Also, SMOTE is sensitive to the quality and content of data and it will not perform well on all images, here the problem statement is related to the medical field then selecting the most stable and trustable method is vital.

3.4.2 Data Preprocessing

When we are talking about image data, different types of preprocessing are compared to structural data and are considered. These techniques help to increase either quality or quantity of the dataset while decreasing the cost of the modeling by scaling data and lessening computational complexity. Since utilizing deep learning has

expanded, scientists have worked on this issue and nowadays we have many methods to handle this part. In this work, preprocessing is divided into two parts:

- Data augmentation, as mentioned before, the dataset is imbalanced and data augmentation helps us to overcome this issue easily. Generally, creating new images by simple techniques can be exerted, such as; cropping, rotating, flipping, zooming, etc. In this work we will implement augmentation in two phases: firstly, we will apply multiple techniques to a malignant class of images to increase the number of images in this class and solve the imbalance problem. Techniques like rotating left and right, zooming, random distortion, random contrast, random brightness, random crop, and flipping left and right are applied to malignant classes for train, validation, and test sets. Secondly, random left and right flipping and random up and down flipping augmentation methods will be conducted on both classes of data to make a diverse set of images from existent images, as this will help the proposed models see different data while training and extracting most features that will be helpful for the proposed model's generalization.
- The CNN architecture we are going to use can accept any size of the image but, it is better for transfer learning architectures that we prepare the same image size as the ones that have been trained before, Generally, pre-trained models take images in 224*224 dimension. therefore, we will reduce the size of the image to 224*224, this way we will not only be able to compare the CNN result with pre-trained models but also the amount of computation will

reduce due to smaller image. After that, we will perform normalization which will decrease computation complexity more.

The dataset is highly imbalanced and in the first step in preprocessing, we had 31,956 samples if we use all of them for training, it will increase the cost of modeling. Therefore, we applied random sample for majority class, so we selected 8,126 images of random benign samples according to the patient id, age, and anatomy site variables. And we used all 575 images of the malignant class. So, we have 8701 samples for both classes. Then we split the dataset into 70:20:10 as train, validation, and test respectively. Afterward, the dataset which we want to use in this work was ready for preprocessing and further actions. Firstly, we loaded images from the original location and created a data frame with the address of each image and its corresponding label as the metadata. We have also moved images into separate folders of benign and malignant classes for train, validation, and test sets to facilitate future access during implementation. To enhance the number of images in the malignant data, we have used the augmenter library for image augmentation. Finally, we have defined a pipeline to process images, this pipeline, load each image and resize it into 224*224 pixels then normalize them by dividing all the data value into 255, converting labels to tensors, and augmenting images.

3.4.3 Modeling

For modeling, we can use one of two approaches, either implement a CNN architecture from scratch or use transfer learning. The first one is mostly used when there are enough data to train the model and it is certain that the overfitting issue will not occur, and the second one is used when you face a lack of data or you want to

reduce the cost of modeling. Since transfer learning architectures have been trained on numerous data that belong to 1000s of classes, thus, their calculated weights can be used in the modeling to reduce the number of epochs while training.

3.4.3.1 Convolutional Neural Networks (CNN)

After the introduction of AlexNet in 2012, which achieved high accuracy with far distance as compared to the traditional machine learning algorithm [50], Convolutional Neural Networks have been used immensely. This neural network is a Deep feed-forward architecture that has been widely used in three main groups of problems, image classification, image segmentation, and object detection. When we are talking about image datasets especially in the medical field with results that can save the life of humans, general machine learning algorithms cannot perform accurately as needed, fortunately, after the launch of AlexNet, researchers are attracted to doing immense innovational investigations and implementation. While reviewing previous works, we witnessed exerted CNNs with greater performance even when compared to AlexNet. Every simple CNN model is constructed from one or few convolution and pooling layers that are used for extracting features, at the end, one or more fully connected layers are used to create classified outputs. The convolutional layer has a learnable kernel, convolving through each image and extracting features. This layer is a matrix with discrete numbers as weights, these numbers are selected randomly at the start of the training and will update after each epoch to reduce errors. Another part is the stride that controls the speed of the kernel when convolving the image. In each layer, the kernel moves through the image, and values of that part of the image will be updated by computing special formula, this step will continue

until the image ends and creates a features map, which is the output of this layer and will be sent to the next layer as input.

The output size of a convolutional layer in a CNN can be computed using the following formula:

$$h\_new = (h - f + 2p)/s + 1$$

$$w\_new = (w - f + 2p)/s + 1$$

where:

h and w are the height and width of the input image, respectively

f is the size of the filter (or kernel)

p is the amount of zero-padding applied to the input image (if any)

s is the stride of the convolution

The output of a convolutional layer will have dimensions h_new x w_new x d, where d is the number of filters used in the layer.

Pooling layers have to down sample created data by replacing the values of one shrink with its representation of it. Max-pooling, min-pooling, and average-pooling are some of the down-sampling methods for selecting the representor of that shrink [51]. Another part of CNN is the non-linear activation function that tries to map inputs to outputs and decide if a neuron is useful for solving the problems or not in the following steps. There are many activation functions, and the most common are Sigmoid, Tanh, and ReLUs, which are selected based on data and special parts of the architecture. While we are going deeper into CNN, we will be able to extract more details. The greater the number of layers is, the more accuracy we will have, but with more necessary computation process to be done, which adds up more cost as well. On the other hand,

sometimes adding more hidden layers would not be effective, and if we compare the result with the less deep architecture, we might find less performance efficiency in the first one in comparison to the latter. Respectively, selecting the number of layers is one of the main things that should be considered during hyphae next operation in neural networks is calculating errors after each epoch; after creating outputs, the loss function will be used to compare the estimated labels with the true labels and then calculate the errors to evaluate the model. Common loss functions are Cross Entropy, Euclidean, and Hinge. After getting the performance of the model after one epoch, we should update the weight of the next epoch with the help of the optimizer function, and one of the best optimizers is Adam. All of these processes will also be peated in the next epochs. Also, we can compare the errors of the current epoch with the best previous one, and save the improvements if they occur as the best model. The following figure illustrates a CNN with two hidden layers; in the first layer kernel size is (9*9*1) and in the second layer its size is (5*5*4).

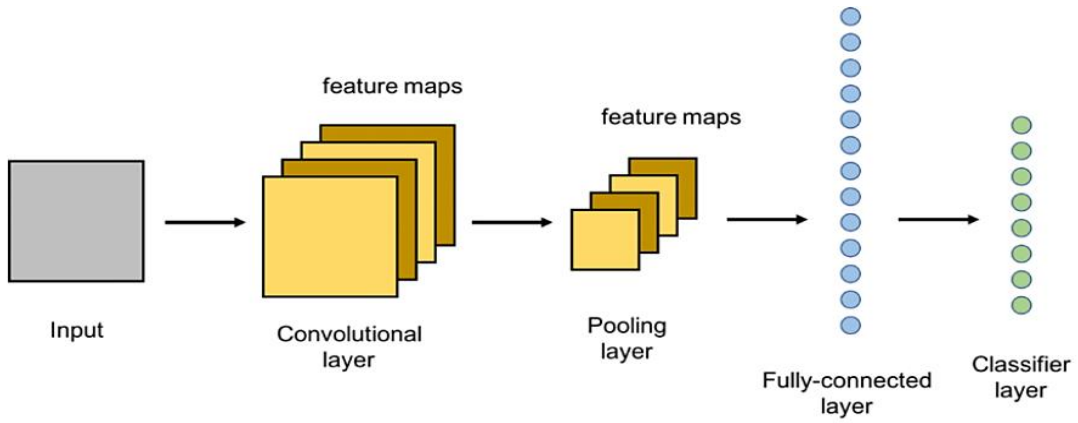

Figure 3.2: Sample Convolutional Neuron Network

3.4.3.2 Transfer Learning

Deep learning algorithms such as CNN can work great when the number of images in each class is high. Also, when we use enough layers to extract data, a huge amount of computing will happen during the training and it may take days or weeks for modern hardware to finish, which is surely not possible for every problem, especially medical ones. Therefore, following the transfer learning approach via reusing calculated weights of a developed model on similar data will save us costs. We can reuse the knowledge which has been collected by an algorithm in previous training on similar data. For example, if we trained a model for the diagnosis of cancer on a histopathology image, we can use it for skin cancer, too, by transferring the weights of that model to our new work. However, it is only worth it because the previous task had vast labeled data and the model has collected enough knowledge. This technique is used in image processing, NLP, sound, and video processing.

In CNN, training starts with recognizing the edge and border of images, then it identifies the shape, and as it goes deeper, more detail will be collected, and at the end, it will try classifying them based on given labels. Afterward, and by following the transfer learning approach, we will exert the first parts, which collect the features, and for the latest layers, we will re-train the model based on the new data. There are many state-of-the-art transfer learning architectures, most of them are trained on the ImageNet dataset, which contains a huge number of images belonging to 1000 classes, such as; ResNet and Inception, and some of those classes were the winners of ILSVRC in different years. Partially customization is the positive point of these architectures. We can customize the input and output layers of these models based on our inputs and

expected outputs. For instance, if we have the same size of data, then no change is required on the input layer, however, in case of different input sizes, we can change the input layer based on our will, or we can replace it with other types of the powerful classification algorithm, such as; XGBoost or SVM, instead of their default classification algorithm, and by doing so the architecture will be used only as a feature extractor. The classifications that we plan to use in our work:

3.4.3.2.1 Resnet50

In deep learning, when we face complex input, such as; an image that has countless details to extract, we add several hidden layers as the first solution because in the first layer, the frame and boundary of the objects in the image will be recognized and the deeper we travel, the more features, such as; color, different objects, etc. will be identified. However, researchers found this is not true, and if the number of layers passes a special limit, the error of the model will increase, although, the performance increases, too. In 2015 Kaiming He et al. introduced a new neural network and they won ILSVRC 2015 classification competition, which is known as Residual Network or ResNet [27]. This architecture is divided into two blocks and each block has a skip connection. The summation of the previous block output and the current block is calculated and the activation function will run on f(x) + x instead of f(x) and create the output of the block, which is sent to the next block. In ResNet 50, instead of using two Conv layers, they used three Conv layers with a specific setting as presented in figure 3.3.

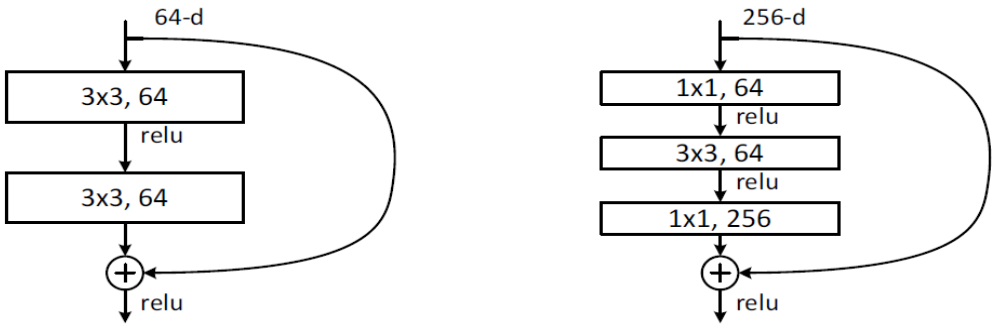

Figure 03.03: Residual Block [52]

3.4.3.2.2 VGG16

The VGG is an abbreviation for Visual Geometry Group Net that was used in CNN, and it has approximately 143 million parameters. These parameters are learned using the ImageNet dataset comprising 1.2 million images, which contain thousands of classes to be used for training. It is a very good architecture to be used for benchmarking. The VGG-16 model architecture has 13 convolutional layers, 2 fully connected layers, and 1 SoftMax classifier.

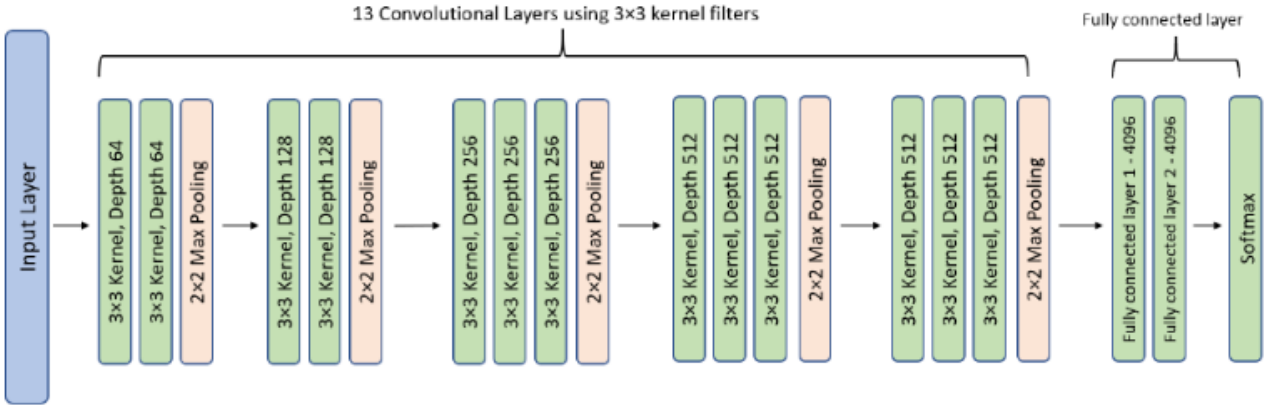

Figure 0.4: VGG16 Architecture [53]

The first and second convolutional layers are made up of 64 3x3 sized feature kernel filters. When passed through the first and second convolutional layers, the input image's dimensions (RGB image with depth 3) changed to 224x224x64. The output is

then passed to the max pooling layer with a stride of 2. The third and fourth convolutional layers consist of 124 3x3 sized feature kernel filters. Following these two layers is a max pooling layer with stride 2, and the resulting output size is 56x56x128. Convolutional layers with kernel size 3x3 are used in the fifth, sixth, and seventh layers. All three make use of 256 feature maps. Following these layers is a max pooling layer with stride 2. Layers eight through thirteen are two sets of convolutional layers with kernel sizes of 3x3 and 512. Following these layers is a max pooling layer with a stride of 1. The following two layers are fully connected hidden layers with 4096 units each, followed by a SoftMax output layer (sixteenth layer) with 1000 units.

3.4.3.2.3 VGG19

The VGG-19 Neural Network consists of 19 layers of deep neural network and has more weights. The size of the "VGG-19" network in terms of fully connected nodes is 574 MB. As the number of layers increases, the accuracy of DNN improves, too. The VGG-19 model is comprised of 19 trainable layers performing convolution, which is fully connected with max pooling and dropout layers.

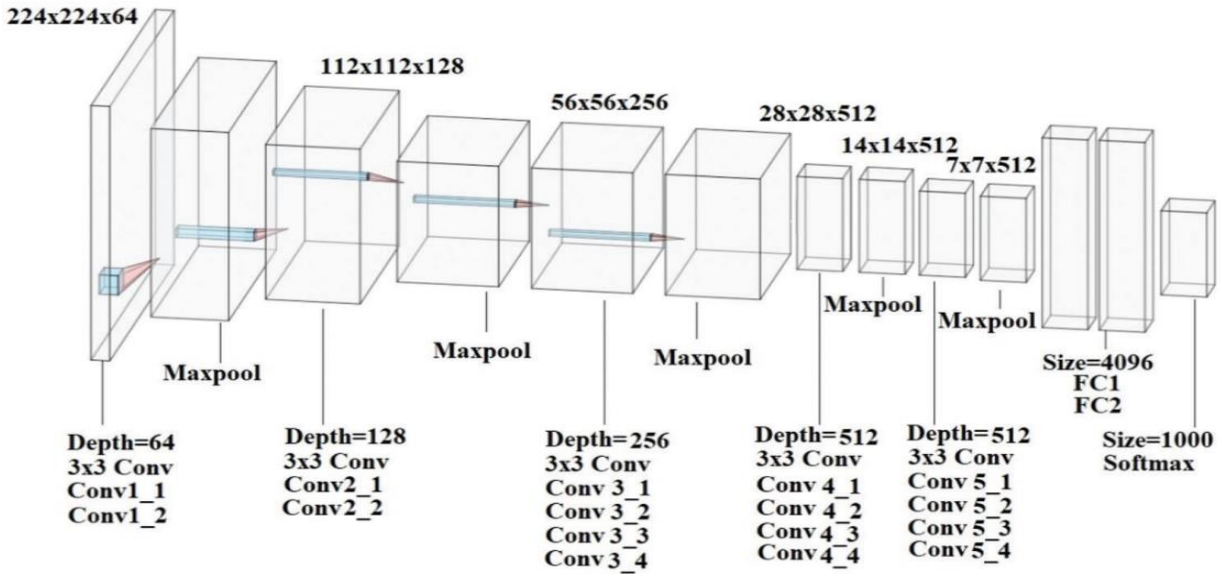

Figure 0.5: VGG19 Architecture [54]

VGG-19 is so beneficial and it simply uses 3 × 3 ConvNet, arranged as above to extend the depth. To decrease the size, max-pooling layers are applied as a handler. FCN layers are two in number and have 4096 neurons applied. VGG19 is trained based on individual lesions, and for the testing, all types of lesions were considered so that we reduce the number of false positives. Convolution layers perform convolving process over images at every pixel, allowing the outcome to pass through the subsequent layer. Filters used in the convolution layer are of 3 × 3 dimensions and are trained for feature extraction. Every stacked convolution layer is subsequently added with Rectified Linear Unit (ReLU) activation function and max-pooling layer. ReLU is presently the best-known non-linear activation function, which allows only the positive portion of the input.

$$\boldsymbol{f(x) = max(0, x)} \qquad \textbf{(2)}$$

3.4.3.2.4 EfficientNetB7

The EfficientNet models are based on fundamental and highly efficient compounded scaling approaches. With the help of this model, we can increase the ConvNet baseline to any target with constrained resources while maintaining the model's usefulness for transfer learning. With significantly fewer parameters, the EfficientNet models generate useful results. EfficientNet versions generally outperform current CNNs like AlexNet, GoogleNet, and MobileNetV2 in terms of accuracy and performance. From EfficientNetB0 to EfficientNetB7, the model's depth, width, resolution, and size all continue to expand while the accuracy also rises. The best model, EfficientNetB7, performs better than earlier CNNs in terms of ImageNet accuracy. It is also 8.4 times smaller and 6.1 times faster than the best existing CNN. For segmentation-related problems, transfer learning might be effective, but all other brand-new problems necessitate completely separate classes. As shown in figure 3.6, the network architecture of EfficientNetB7 is divided into seven blocks based on the filter's size, striding, and several channels.

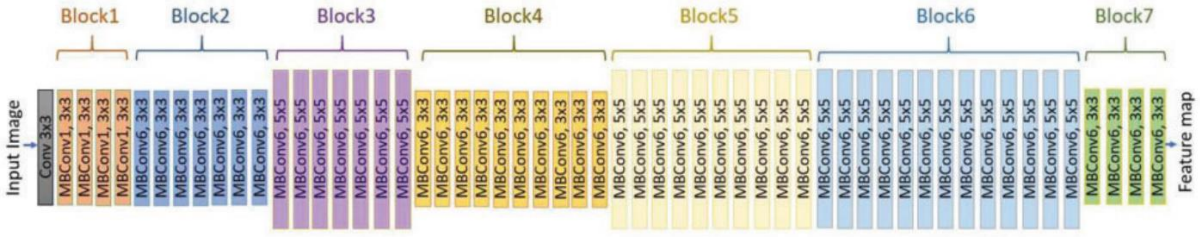

Figure 0.6: EfficientNetB7 Architecture [55]

3.4.3.2.5 DenseNet 169

DenseNet has been built from some Dense blocks, they are connected by a transition block that includes a set of convolution and pooling layers. The structure of DenseNet with three Dense blocks is presented in the upcoming figure 3.7.

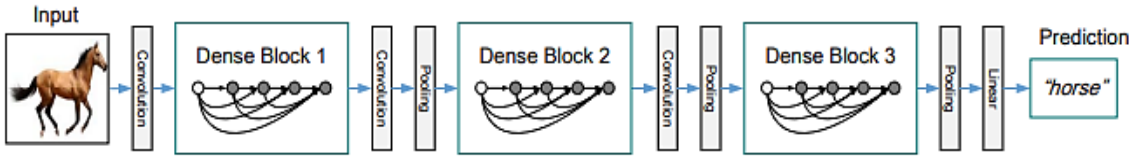

Figure 3.7: DenseNet Architecture [56]

Each Dense block is made from a set of blocks that contain some Conv layers, and each subset of the conv block receives outputs of all previous blocks with the same feature map size. Resnet does summation on outputs, and DenseNet concatenates outputs. In the next figure 3.8, the structure of a Dense block with a growth rate of k = 4 is presented.

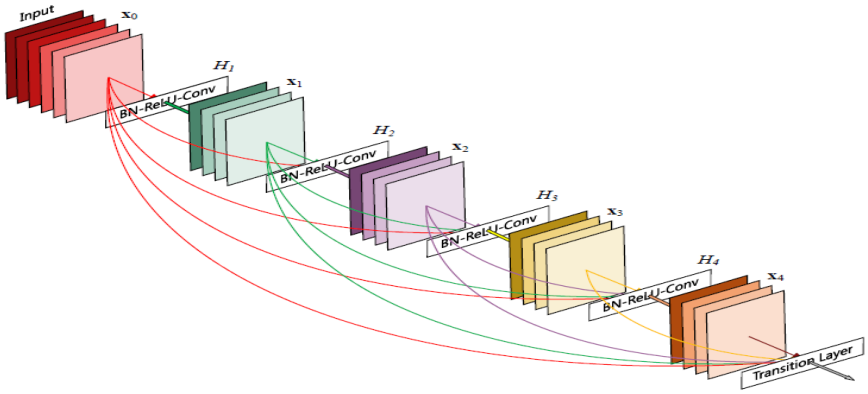

Figure 0.8: DenseNet block with a growth rate of k =4 [56]

Here 'k' is used to set the number of feature maps that will be created by the H function, then if each H function creates k feature maps, the layer n has $k_0 + k \times (n-1)$ input feature maps while k_0 is the number of channels in the input layer. The architecture used for DenseNet 169 on the ImageNet dataset contains four Dense blocks, each

Dense block has 6,12,32,32 sets of Conv blocks, which include two Conv layers with 1*1 and 3*3 kernel size respectively, and each Conv layer corresponds to the sequence BN-ReLu-Conv, in this architecture the growth rate k=32 was selected.

3.4.3.2.6 Inception V3

We can say Inception is inherited from Inception V2, but with some improvements that make it work in a greater way in comparison to the previous version. As is presented in table 3.2 below, this architecture has a stem block that contains six Conv layers with kernel size 3*3, and after the third one, there is a pooling layer. The padding of Conv layers is set to zero when they are mentioned, and for other Conv layers, there is not any padding setting. To reduce the size of the grid, a reduction method is utilized, which decrease decreases between Inception blocks. The deepness of this architecture is 42 and its cost is 2.5 higher than GoogleNet but, it is still much less than VGG architecture. Its structure is presented in figure 3.9. The main difference between Inception v2 and v3 is factorizing convolutions, which reduce the number of connections or parameters without the effect of the operation of the network, for example replacing a Conv layer with kernel size 5*5 with two Conv layers with kernel size 3*3, this makes the network extract features explicit same as before but, the computational part and the cost of the model are reduced simultaneously.

Table 0.2: Architecture of Inception V3

type	patch size/stride or remarks	input size
conv	3×3/2	299×299×3
conv	3×3/1	149×149×32
conv padded	3×3/1	147×147×32
pool	3×3/2	147×147×64
conv	3×3/1	73×73×64
conv	3×3/2	71×71×80
conv	3×3/1	35×35×192
3×Inception	Inception Block A	35×35×288
5×Inception	Inception Block B	17×17×768
2×Inception	Inception Block C	8×8×1280
pool	8 × 8	8 × 8 × 2048
linear	logits	1 × 1 × 2048
softmax	classifier	1 × 1 × 1000

There are various types of factorizations, like; converting a 5*5 Conv layer to two 3*3 layers one after another, which is used in Inception Block A, the second one, a 3*3 converting Conv layer into two 1*3 and 3*1 layer one after another, which is used in Inception block B, and the last on is adding parallel 1*3 and 3*1 Conv layers on top of the 3*3 and 1*1 layers to increase dimensional representations that are used in Inception block C.

3.4.3.2.7 Xception

This architecture has 36 convolutional layers with a kernel size of 3*3, which are used for extracting features, these layers are organized in 13 separable Conv sets with a linear residual connection, however, the first and the last layers do not have that residual connection. After extracting features, logistic regression is involved to classify images, and if it is needed in some problems a fully connected layer can add before logistic regression.

As is clear in the next figure 3.9, the Xception architecture is divided into three main flows; the entry flow starts with two Conv layers with kernel size 3*3 which are followed by a batch normalization layer, after those, three separable blocks are defined and each block is constructed with two separable Conv layers with kernel size 3*3, with an activation function, Relu, and a Maxpooling layer. The output of this block will sum with the output of the Conv layer (residual connection) with kernel size 2*2, and prepare the input of the next block, and the next residual connection. Both separable Conv layers are also followed with a batch normalization layer. The middle flow has a block containing three separable Conv layers with kernel size 3*3 and the activation function, Relu, the output of the block sum with the input creates the output of this flow.

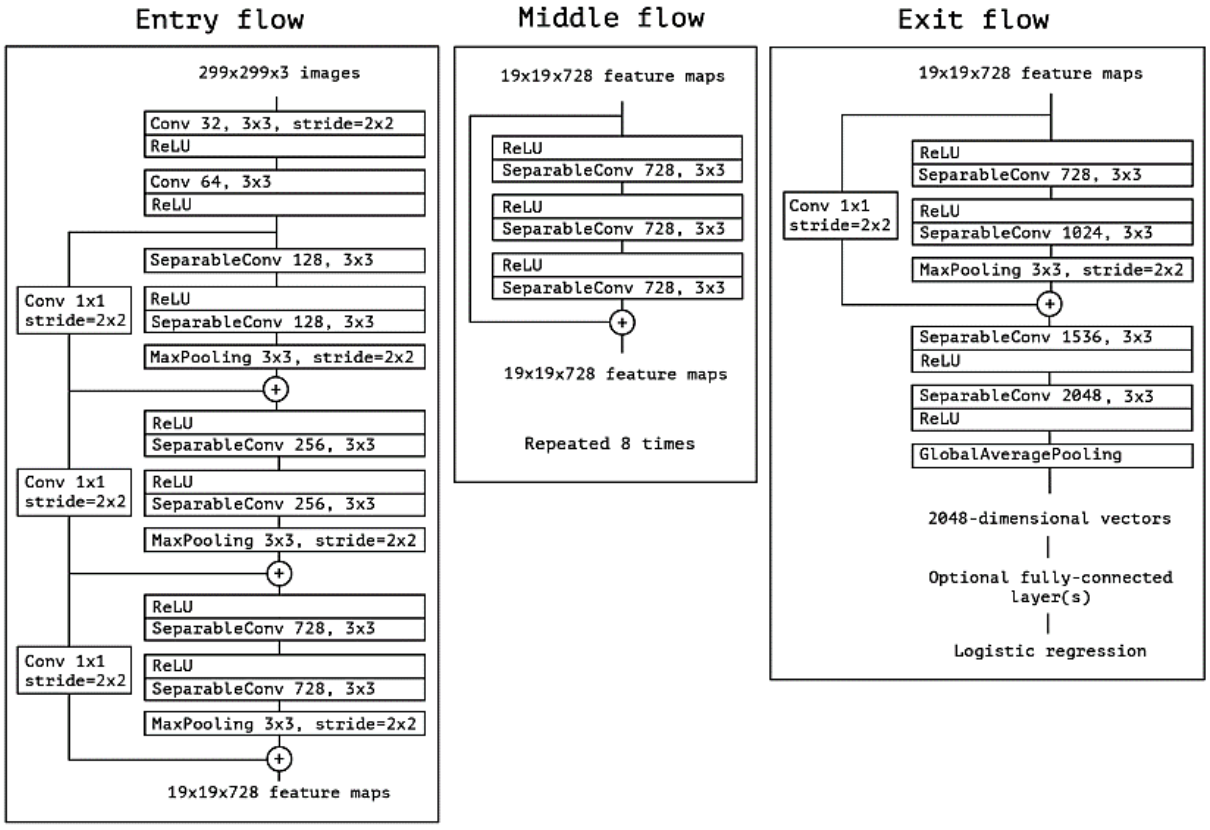

Figure 0.9: Xception Architecture blocks

Finally, in the exit flow, we have one separable block, which was explained before, that is followed by two separable Conv layers and a Global Average Pooling, which prepares data for the logistic regression algorithm. This structure is easy to create with a library, such as; Karas. The complete detail of Xception architecture is illustrated in figure 3.10. This architecture did not perform more accurately on the ImageNet dataset as compared to Inception v3 but, on the JFT dataset, there was a significant difference.

3.4.3.3 Ensemble Approach

This approach tries to improve the result by combining various results which are collected beforehand from diverse sources. Generally, we use a similar attitude when we need to decide about something, for example, before visiting a specialist doctor, we ask others to introduce some professionals they know, then we will select the most suggested one. This is also used when some treatment is needed, we will visit some specialists and choose the most suggested prescribe. Thus, we can say this method is based on our method of deciding daily life and we are using this technique because there is some uncertainty. As it is mentioned during the CNN explanation, neural networks use random weights in the first epoch, after finishing the first epoch the errors are calculated, and with the optimizer function, calculated error weights will be updated for achieving a better result in the next epoch. If we run our model again the accuracy of the model may change to use different random weights at the starting phase, then it will be the same as our approach in real life so, ensemble learning comes into the picture to make the results more trustworthy and accurate.

For implementing this approach, we can train one model with different data which is called bagging, or we can train our data with various models, which is called boosting,

or a combination of them, which is stacking. In the case of K-Fold cross-validation, it is also possible to perform bagging instead of selecting the best model, but in this work, boosting technique will be implemented. We plan to run some transfer learning architectures, such as; Inception v3, DenseNet, and a CNN architecture that will be designed and trained from scratch, each of them has its way of feature extraction and classification, hence, they have varied weaknesses and strengths. So, by utilizing ensemble learning and merging predictions of all models, at the end of the day, we can be sure the data is verified from different aspects and our result will be more explicit.

In ensemble approach the results obtained from multiple models combined to develop higher accuracy and performance. In this work, various pre-trained architectures were trained on skin cancer dataset with different strengths and weakness, they had different parameters for learning and even methods of extracting features. Then it may cause each one cultivates knowledge from divers' aspect of input data during training. Therefore, ensemble approach can help to combine output of models and select the most frequent answer predicted by each model as the final prediction, then the output is more accurate and reliable.

Complexity: Here max voting was applied to the predictions of models when there was the minimum data for making the final decision as compared to the complexity of images, thus the hardware which is used for model training can handle it easily.

Diversity: stacking ensemble approach was used to create diversity, and various types of pre-trained architectures were used for predicting the class of images. This

architecture, trained on the huge amount of data, and then reduced the need for hyperparameter tuning for CNN models.

Ensemble Weighting: In this paper, the weighting technique did not implement, the top four high-performance models were selected with equal weight to predict final outputs because the performance of the selected models was so close and weighting would not helpful.

3.5 Model Development

According to the American Cancer Society, over 108,000 new melanoma cases are expected to be diagnosed in 2022. It is also expected that nearly 9,000 people will die as a result of skin cancer disease[1]. Similar to other cancers, early detection can improve treatment effectiveness. Dermatologists currently examine all of the patient's moles to identify outlier lesions that could show that there is melanoma. The number of evidences to verify might be high and take a long time, which is the root of many human mistakes if it is done by one person, and is also related completely to the experience of that lab member. With fewer sources in the lab and perfect specialists, a medical test is expensive and fewer people can afford its cost, and normally, they would postpone it, till it becomes severe and the chance of healing becomes so low. Skin cancer detection applications can be fed with all moles' images of a patient to determine whether any image represents a melanoma which makes the process faster, cheaper, and more accurate, in addition to dropping human mistakes at the same time.

[1] https://www.cancer.org/

3.5.1 System Structure

As per what is explained in the previous chapter, the structure of the system divides into four major sections as shown in figure 3.10.

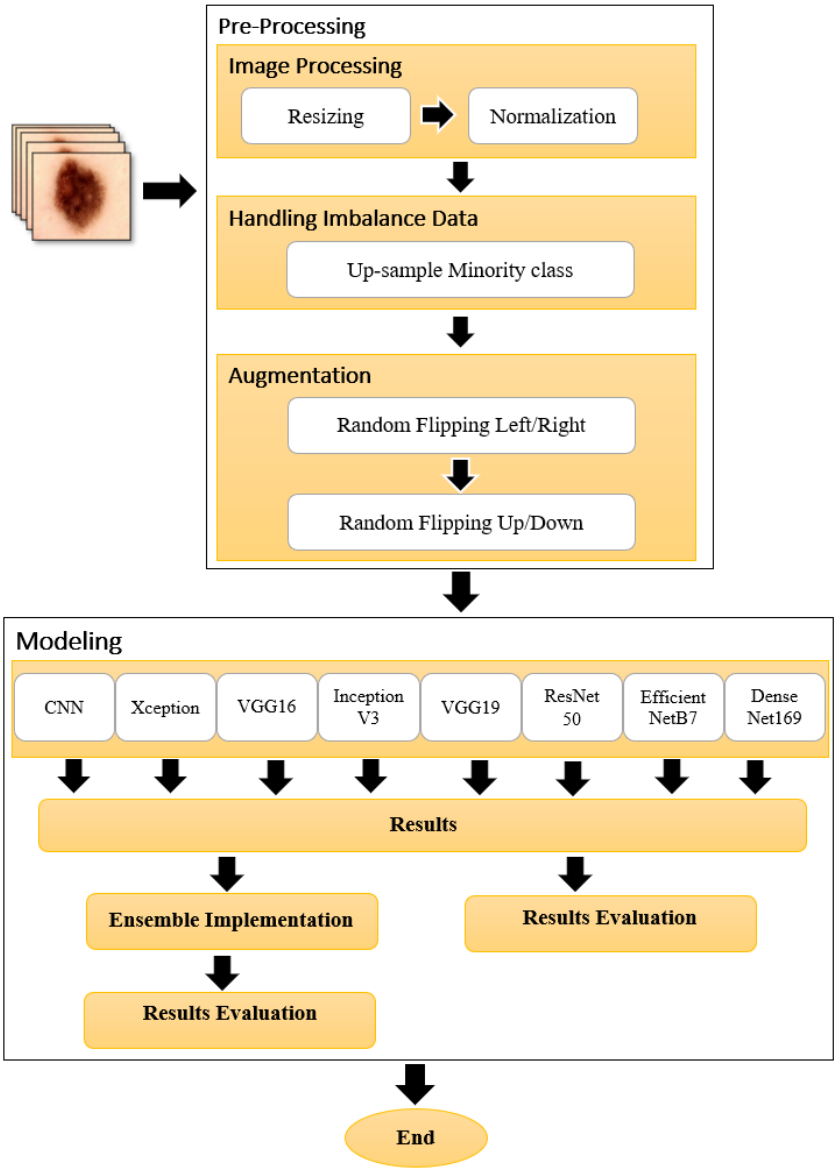

Figure 0.10: Process Flow

The first one is preprocessing, which includes preparing data for models and augmentation, the next section is modeling with the help of CNN and some transfer learning architectures after that ensemble learning section is added to combine the

strengths of models, and the last section is the evaluation. The evaluation was done per architecture on validation and test datasets separately and after implementing ensemble learning. Afterward, we could easily compare every result and select the best solution for the problem. The following chart will give an overall idea about the structure in graphic format:

3.5.2 Preprocessing

Now we can make our input pipeline. This pipeline creates a Dataset tensor slice for the training, validation, and test dataset separately, for each of these sets. The first preprocessing function will be the transformation of each element of the dataset, and then the elements are cached. The first-time dataset is iterated, its elements will be cached either in the specific file or in the memory and subsequent iterations will use these cached data. Then we combine consecutive elements of this dataset into batches and create batches in size of 16. Most dataset input pipelines should end with a call to prefetch. This allows later elements to be prepared while the current element is being processed. Prefetch doesn't allow the CPU to stand idle when the model is training and GPU is busy, and then the delay between the training of every batch will be zero. Therefore, with the same loading and transferring time we can train more batches at the same time. In addition to the above pipeline processes, data shuffling and online augmentation are applied only on the train set to give some appearance variation to data such as flipping around the x or y axis, this impacts the training performance.

3.5.3 Models Building

For modeling, different structures that are a mix of the designed CNN model and transfer learning models are selected. Transfer learning architectures include ResNet50, VGG16, VGG19, EffientNetB7, DenseNet169, Xception and Inception v3. We will also execute an ensemble of the mentioned models. The final goal is to classify tumor images into benign and malignant, which is a binary classification. In the coming sections, we will discuss the setting of each network.

3.5.3.1 CNN Model Training

The CNN model has 3 convolutional layers with (3,3) kernel size, the first layer has 256 filters and it reduces to 128 and 64 in the second and third Conv layers respectively, and the Relu activation function is used in all Conv layers. Each Conv layer is followed by a max pooling layer with (2,2) pool size. So, by selecting the maximum value, the down samples of the calculated data shrink. This will help us in controlling the computing complexity. Afterward, batch normalization is used to improve the efficiency of CNN training significantly. It helps avoid vanishing or exploding gradient by keeping the activations more stable during training, which in turn allows using larger learning rates, also a dropout layer is added with the size of 0.3, this layer will drop neurons randomly based on the given value in each epoch to balance the process and reduce the probability of overfitting. The feature extraction is done through these Conv layers and the model will be ready for classification. The classification part starts with a fully connected layer, which is a Flatten layer, then, two Dense layers are used. The filter size of the first one is 2,048 and the second is 1,028.

The formula for the loss function with L2 regularization in a CNN can be written as:

Loss with L2 regularization = Loss + (lambda / 2m) * sum(square(W)) where:

Loss is the original loss function

lambda is the regularization parameter, which controls the strength of the penalty

m is the number of training examples

W is the matrix of kernel weights for the convolutional layers

The L2 regularization term penalizes the model for having large values in the kernel weights, which can help prevent overfitting to the training data. By setting the kernel_regularizer to L2 in the CNN layers, the regularization term is automatically applied to the kernel weights during training.

The impact of the L2 penalty on the loss is the additional cost incurred by the model for having larger kernel weights. This additional cost can help balance the trade-off between fitting the training data well and avoiding overfitting, and can lead to a more generalizable model with better performance on unseen data.

The selected value of l2 is 0.001 and 0.001. Relu is set as the Dense layers' activation function. Finally, a Dense layer, which is set by filter size 2, and an activation function, SoftMax, which duty is to do the final binary classification, are added to the Sequential architecture. In this way, one epoch of CNN is finished.

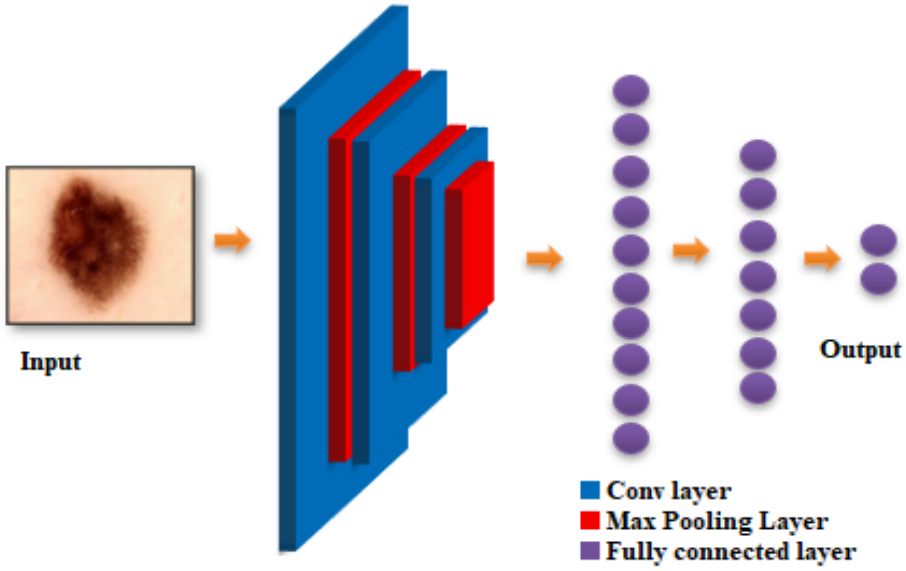

Figure 0.11: CNN Model Structure

In this model, the Adam optimizer with a learning rate of 0.00001 and binary-cross-entropy as loss is utilized to update weights for the next epoch, and accuracy is selected for evaluation of the model after each epoch. During the modeling phase, we will use early stopping to stop the train when the validation loss does not decrease for 7 epochs. And the last step, if the new calculated validation accuracy increases as compared to the previous best-saved model, the checkpoint will be saved and the best model will be updated until that point in time. Then we will save all the results, such as; accuracy and loss, after each epoch in a CSV file. With the help of this file, we can plot and follow the process of training, while concluding the number of required epochs. Overfitting recognition has a possibility of happening. CNN Model is proposed to train on 150 epochs with a 16-batch size.

3.5.3.2 Transfer Learning Models

Transfer learning is a technique where we can use pre-trained architectures in the modeling. The key point is that those models should be trained on similar data.

Generally, deep learning architecture faces a huge amount of data such as image, video, sound, or natural language text, thus, extracting features needs precisely powerful hardware, and even in that case their training will take days or months and most of the time it is not affordable, in the financial and time manners. Furthermore, deep learning algorithms start their training with random weights and then try to improve performance by updating them after each epoch, so their result is completely related to those random weights and it may differ if we repeat the training. To solve these problems, the concept of transfer learning is introduced and it has been approved by many researchers in different fields new technique is helpful and even may work better. We also plan to reap the benefit of it in this work. For transfer learning, some high-performance architectures are utilized, they include ResNet50, VGG16, VGG19, EfficientNetB7, DenseNet169, Inception v3, and Xception. All models are added as a base model into the architecture and Include is set as the top parameter into False, this ensures that we can add the custom input and output layers based on the dataset. We also selected weights that each based model calculated while training on the ImageNet dataset by setting weights parameters into ImageNet.

To change the output layers, first Global Average pooling is added as a fully connected layer, then, two Dense layers with 512 and 256 filter sizes are added to the sequential, ‘he-uniform’ is chosen as kernel initializer, which draws samples from a uniform distribution, and Relu is used as their activation function. We also added a Batch Normalization layer which normalizes the output of the previous layer to reduce the computational process. Another action we selected, to reduce cost and overcome overfitting at the same time, is adding a dropout layer with a size of 0.5. Then, we

added three more dense layers with 2,048, 1,024, and 512 as filter sizes and L2 regularization as Kerner regularizer which removes a small percentage of weights at all of the iterations. The Relu activation function is selected here as well. Finally, a Dense layer with filter size 2 for binary classification, which is the goal, and activation function SoftMax are added to the architecture as the last layer.

For the transfer learning architecture, 0.00001 is chosen as the learning rate and Adam is selected as the optimizer. The Beta_1 and Beta_2 parameters of the Adam optimizer are set to 0.9 and 0.999 respectively and the 'amsgrad' parameter is set to True to accelerate the optimization process. The target is binary classification, then we config the loss parameter of the model to 'Binary Cross Entropy' and matrices to 'accuracy'. The batch size that is selected is 16 as we used in CNN Model where the epochs are 75 for all transfer learning models. Same to the CNN model, we used early stopping to prevent model overfitting and saved the model if the current validation accuracy is higher than a previously saved model, have the best-fitted model after finishing training for further use. Because we want to compare the results of the different transfer learning architectures together, we kept all the explained setting equal during the training of each one.

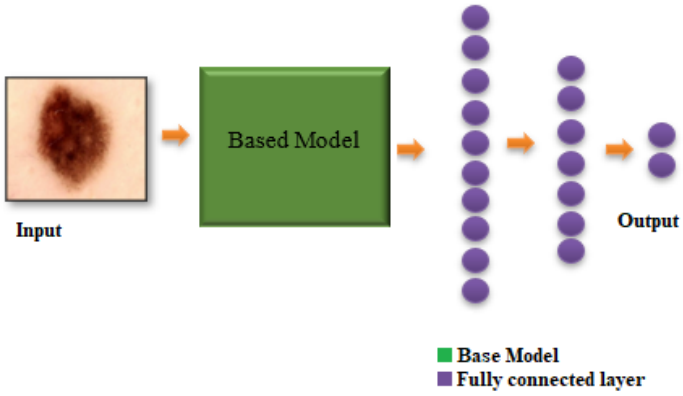

Figure 0.12: Transfer Learning Architecture

3.5.3.3 Ensemble Approach

Ensemble learning is the process of combining the result of multiple models; they can be the same or different algorithms. There are a lot of techniques in this domain which are divided into three main groups, Bagging, Boosting, and Stacking. The first one is performed when we want to combine the result of one model when it trains with various sets of data multiple times. The second one combines the result of some models and the last one is a mix of both. In this work, we will implement the second one, then after training the models, which are CNN, ResNet50, VGG16, VGG19, EfficientNetB7, DenseNet169, Inception v3, and Xception, we will use them to predict the class of images in validation and test dataset, then we will run max voting technique with the aim of sum function, because result belongs to [0, 1] class, and select the class which gets the highest vote. This solution reduces variance as well as the generalization of errors.

Figure 0.13: Ensemble Approach

3.5.4 Evaluation

There are two different ways for evaluating the model when we face supervised problems. First, one is breaking data into training and testing, then training the model with training data and using this model for predicting test dataset labels. By using this way, we can compare predicted results with true values and compute the errors. The second one is K-Fold cross-validation, in this case, the data will be divided into K

subsets and the model will be trained with all subsets, except for one which we keep out for evaluation. After each training round, we calculate matrices, and finally, the best model will be selected. This method is better than the first one but, it brings much extra computation and cost than in deep learning, because of the huge amount of computation in networks, it is not recommended unless it is affordable.

In this work, we used the ISIC skin cancer dataset which includes 33,126 images, 32,542 benign tumors, and 584 malignant tumors. After augmentation, we will select more than 8,000 images from each class, to have sufficient data for assessing the prediction. We will split the data into 75:20:10 as train, validation, and test datasets respectively, we will keep some unseen data as a test dataset to check the functionality of models and be sure there are not any bias issues. Several metrics such as accuracy, precision, recall, f1-score, and AUC score are generally used for evaluating classification models, thus, we will calculate them too. In addition, the problem that belongs to the medical domain, and confusion matrix, as well as AUC/ROC, is mandatory. The confusion matrix checks how many wrong classifications occurred and AUC/ROC presents the capability of models in predicting true classes graphically. We then can compare Sensitivity and Specificity together.

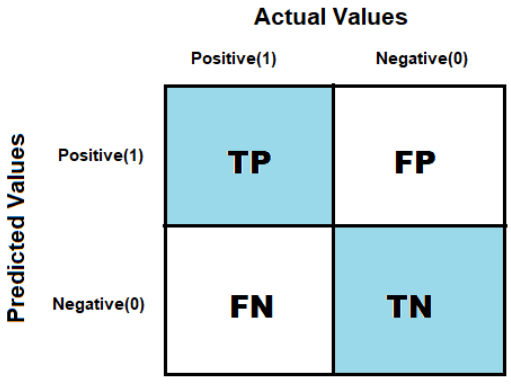

Figure 0.14: Confusion Matrix

Where

- ✓ True positive (TP): The true class of the sample is a positive example, and the result that the model classified correctly is also a positive example.
- ✓ True Negative (TN): The negative class of the sample is a negative example, and the result that is correctly classified by the model is also a negative example.
- ✓ False Positive (FP): The true class of the sample is negative, but the model classifies it as positive.
- ✓ False-negative (FN): The true class of the sample is a positive example, but the model classifies it as negative.

The formulas for calculating metrics are:

Accuracy: this is a performance metric to measure how often the algorithm classifies a data point correctly. It is calculated as the ratio between the number of correct predictions to the total number of predictions. More formally, it is defined as the number of true positives and true negatives divided by the number of true positives, true negatives, false positives, and false negatives.

$$Accuracy = \frac{(TP + TN)}{(TP + TN + FP + FN)} \quad (3)$$

Sensitivity or Recall: Measures how capable the system is in classifying Skin Cancer as malignant and benign; the recall is calculated as the ratio between the number of Positive samples correctly classified as Positive to the total number of Positive samples. The recall measures the model's ability to detect positive samples. The higher the recall, the more positive samples are detected.

$$Sensitivity = \frac{TP}{TP+FN} \quad (4)$$

Specificity: Specificity measures the proportion of true negatives that are correctly identified by the model. It is also known as the True Negative Rate (TNR) of the model.

$$Specificity = \frac{TN}{TN+FP} \quad (5)$$

Precision: The precision reflects how reliable the model is in classifying samples as Positive. Precision is the ratio of the number of correctly Positive Samples (TP) to the total number of Samples that are classified as Positive (TP + FP). The Precision increases as the FP decreases.

$$Precision = \frac{TP}{TP+FP} \quad (6)$$

F1-Score: The F1-score combines the precision and recall of a classifier into a single metric by taking their harmonic mean. It is primarily used to compare the performance of two classifiers.

F-Score is characterized as:

$$\mathrm{F} - \mathrm{Score} = 2 * \frac{\mathrm{P\,recision} * \mathrm{Recall}}{\mathrm{Precision} + \mathrm{Recall}} = \frac{2\mathrm{TP}}{2\mathrm{T\,P} + \mathrm{F\,P} + \mathrm{FN}} \quad (7)$$

Receiver Operating Characteristic Curve (ROC):

It is a graph that shows the performance of a classification model at all classification thresholds. Essentially, it separates the “signal” from the “noise” by plotting the TPR against the FPR at different threshold levels. Each point on the roc curve represents sensitivity to the same signal stimulus, making it a complete indicator reflecting the continuous sensitivity and specificity variables.

True Positive Rate (**TPR**) is a synonym for recall and is therefore defined as follows:

$$TPR = \frac{TP}{TP+FN} \quad (8)$$

And **False Positive Rate** (**FPR**) is defined as follows:

$$FPR = \frac{FP}{FP+TN} \quad (9)$$

In figure (3.5-2): “TPR” on the y-axis and False Positive Rate “FPR” on the x-axis fig

where:

The Area under Curve (AUC): AUC provides an aggregate measure of performance across all possible classification thresholds. The area under the ROC may be used as a gauge to assess the classification method since it is obvious that when the FPR is relatively low and the TPR is relatively high, there will be a better model and a bigger area under the corresponding curve. As a result, the AUC stands for the areas under the ROC curve, or the model's capacity to categorize classes. Figure 3.15 shows AUC with the blue area and ROC curve.

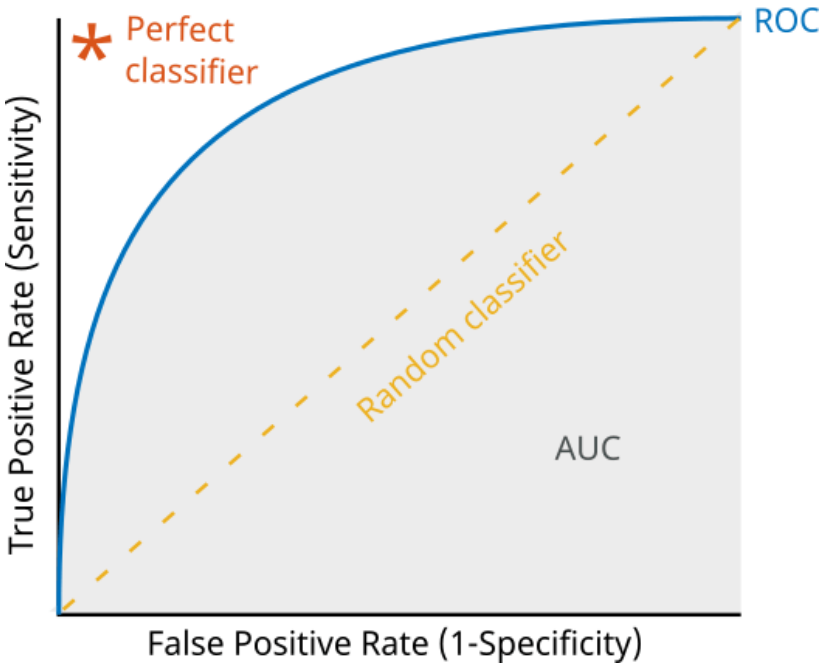

Figure 3.15 AUC and ROC

3.6 Conclusion

In this chapter, we discussed the modeling part of our model, where we discussed the research design, procedures, and dataset preparation and preprocessing. Also, we discussed the model development and building. Furthermore, we discussed CNN and transfer learning algorithms training. Finally, we introduced the performance metrics for model evaluation.

Chapter 4

Experiments and Results

4.1 Experiments

To have a valuable and more accurate comparison of the results, we have split the data into train, valid, and test, with the percentage of 70:20:10 respectively, then we have validated models by using a validation dataset during training. In addition, after training prediction has been done on both valid and test (unseen data) datasets, we could simply analyze our result. We have trained the data through a customized CNN model and some of the transfer pre-trained learning architectures, such as; VGG16, VGG16, InceptionV3, Xception, EfficientNetB7, ResNet50, and DenseNet196, were used as base models. The ImageNet weights were initiative weights and we replaced their top layers with five Dense layers. The Dense layers had 512, 256, 2048, 1024, and 512 channels, and, considering the objective of this work, which is binary classification, a Dense layer with 1 channel was selected. Then, the ensemble approach was implemented, which helped us in covering the weaknesses in the models and improve the final performance. Therefore, we combined several models and merged them to apply the average ensemble technique. The best combination, which is a mix of the best five models, has been selected as the proposed model.

The Computing Environment in this work was done on Lenovo Laptop Intel(R) Core (TM) i7-8550U CPU @ 1.80GHz 1.80 GHz, RAM 16.0 GB, SSD: 240 G, and HDD: 1 TB with Windows 10 Pro. For applying experiments, Google Colab was used with the

subscription to Colab Pro. All implementation codes were run via GPU-High RAM resources and all images and CSV's files were uploaded to Google Drive.

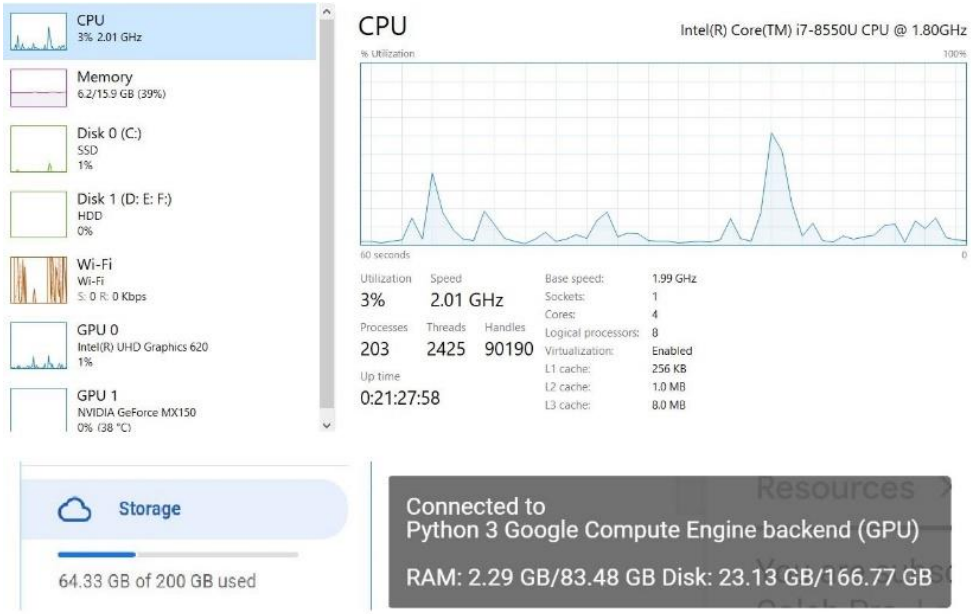

Figure 0.1: Computational Environment

4.2 Results Analysis

For results analysis, we calculated the confusion matrix, which shows the number of correct and wrong classifications, to make the results more suitable to our goal, since we are doing research in the medical field, which is minimizing the wrong classifications. Moreover, with the help of the confusion matrix all of the important metrics; Accuracy, Precision, Recall, and F1 scores were calculated for all architectures. Finally, we plotted the AUC/ROC curve to compare how each model performed for both classes.

4.2.1 CNN Model

The CNN model on validation data presented an overall accuracy of 91.72%, precision (positive predicted values) of 91.78%, recall of 91.72%, and F1-score of

91.72%. And this was the result of the performance of the model on validation data. The learning curves are depicted in the 4.2 diagram; it observed the gradual increase in accuracy during training, as it fluctuated around 88% to 92% on validation data. As shown in Table 4.1, it could achieve 90.96% accuracy, 91.16% precision, 90.96% recall, and 90.95% F1 score on the test dataset.

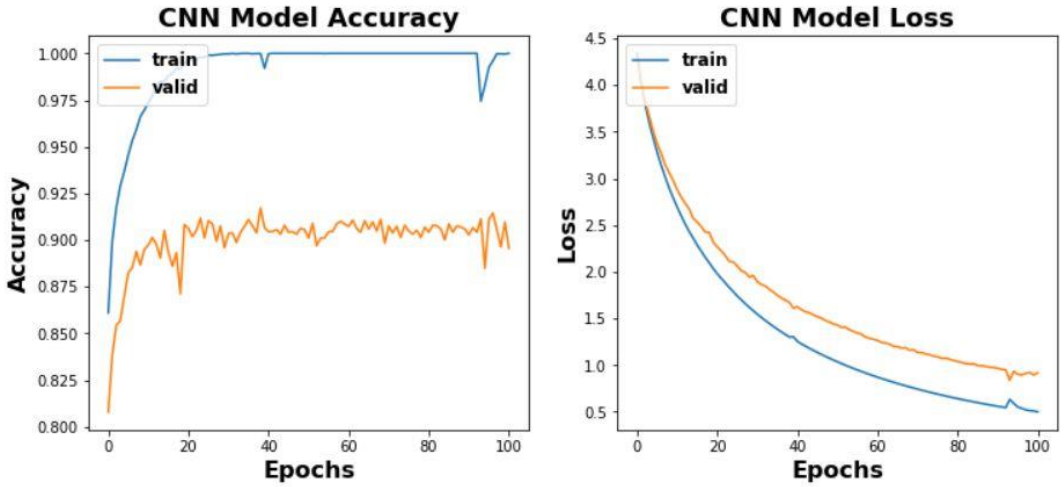

Figure 0.2: Learning Curve - CNN Model

Table 0.1: Classification Report - CNN Model

Classification Report CNN Model

	precision	recall	f1-score	support
Benign	0.88	0.94	0.91	813
Malignant	0.94	0.87	0.91	813
accuracy			0.91	1626
macro avg	0.91	0.91	0.91	1626
weighted avg	0.91	0.91	0.91	1626

*****Performance Metrics*****

Accuracy: 0.9096
Precision: 0.9116
Recall: 0.9096
f1_metric: 0.9095

The Confusion Matrix, as shown in figure 4.3 of the CNN model on the dataset test, tells us that 102 malignant cases were classified as benign, and 45 benign cases were wrongly assigned to the malignant class. The area under the curve was 0.95 for each class as shown in figure 4.4.

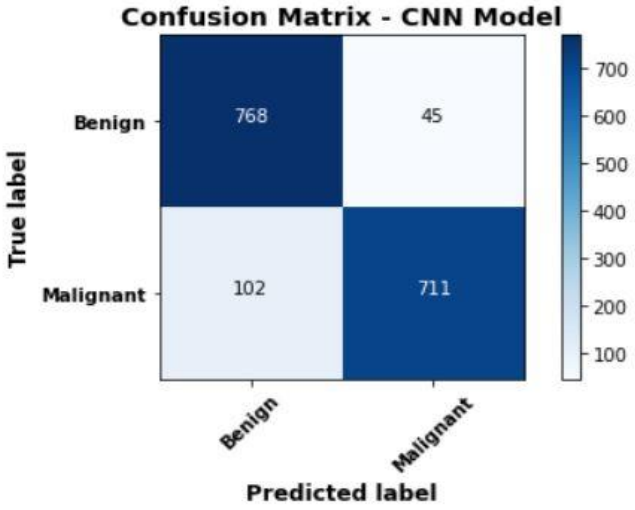

Figure 0.3: Confusion Matrix - CNN Model

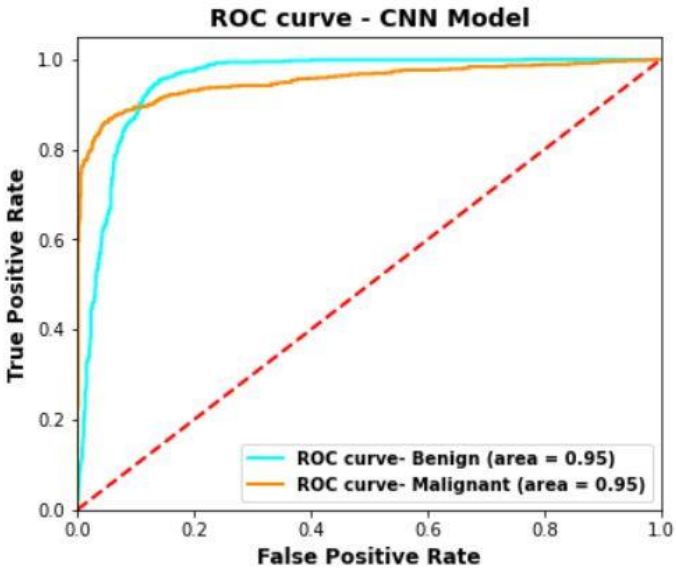

Figure 0.4: AUC/ROC Curve - CNN Model

4.2.2 ResNet50

The ResNet50, which is our first transfer learning model, had an overall accuracy of 94.02%, a precision of 94.05%, a recall of 94.02%, and an F1-score of 94.02% on the validation dataset, and there was no overfitting as shown in the learning curve in figure 4.5. ResNet50 model obtained 91.88% accuracy, 91.98% precision, 91.88% recall, and 91.88% F1-score on the test dataset as shown in table 4.2.

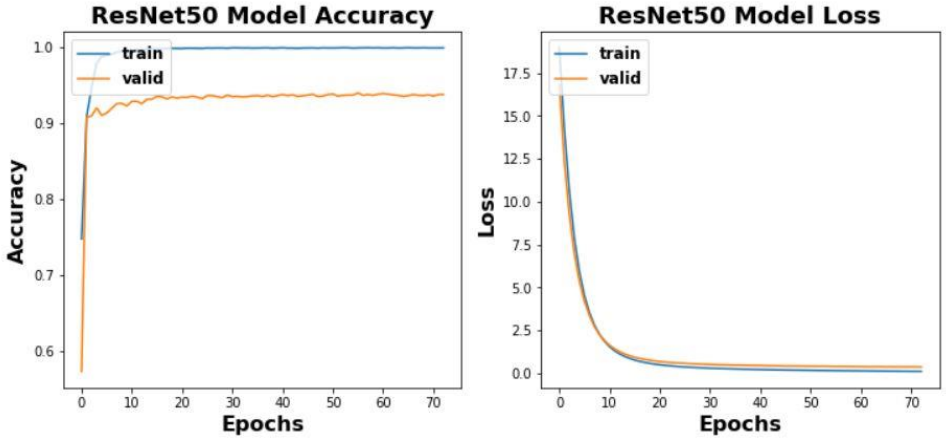

Figure 0.5: Learning Curve - ResNet50 Model

Table 0.2: Classification Report - ResNet50 Model

Classification Report ResNet50 Model

	precision	recall	f1-score	support
Benign	0.90	0.94	0.92	813
Malignant	0.94	0.89	0.92	813
accuracy			0.92	1626
macro avg	0.92	0.92	0.92	1626
weighted avg	0.92	0.92	0.92	1626

*****Performance Metrics*****

Accuracy: 0.9188
Precision: 0.9198
Recall: 0.9188
f1_metric: 0.9188

As we can see in figure 4.6, the confusion Matrix of the test dataset shows that only 86 malignant cases were classified as benign and 46 benign ones were detected as malignant. Hence, the number of misclassifications for malignant images is almost double. The area under the curve was 0.97 for each class as shown in figure 4.7.

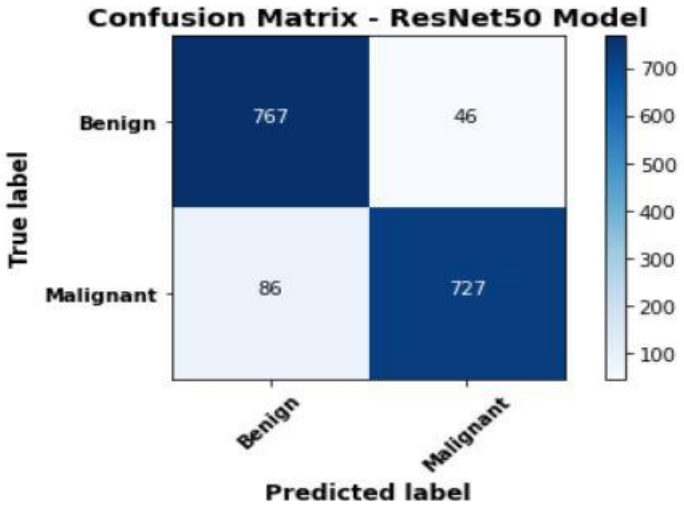

Figure 0.6: Confusion Matrix - ResNet50 Model

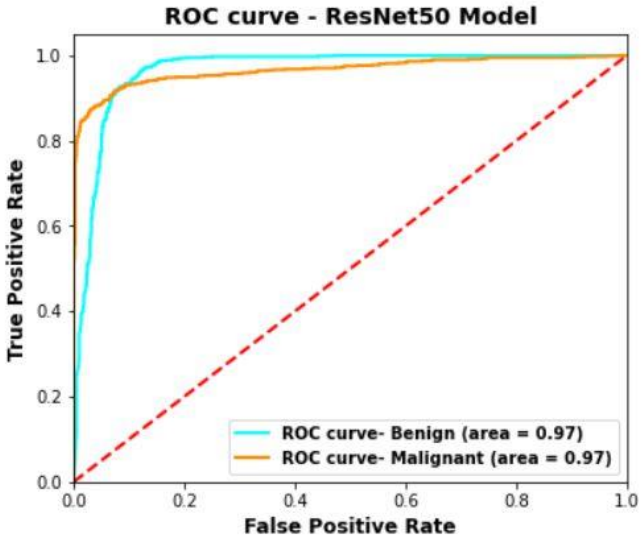

Figure 0.7: AUC/ROC curve - ResNet50 Model

4.2.3 VGG16

The results, which are based on the validation dataset withVGG16 model, had an overall accuracy of 95.12%, a precision of 95.31%, recall of 95.12%, and an F1-score of 95.12%, as the learning curve in figure 4.8 shows. In addition, there was no overfitting and it observed the gradual increase in accuracy during the training, as it fluctuated around 91% to 94% on validation data. On the test dataset, VGG16 model accuracy was 93.67%, precision was 93.93%, recall was 93.67% and F1-score was 93.66, as shown in Table 4.3.

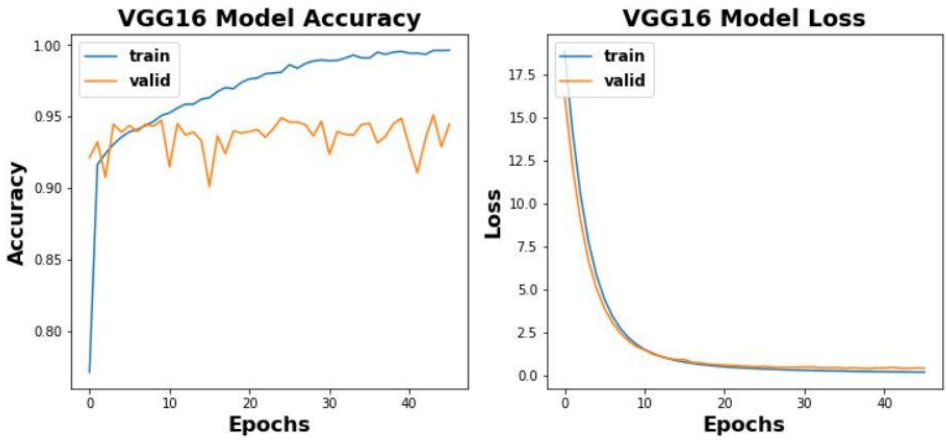

Figure 0.8: Learning Curve -VGG16 Model

Table 0.3: Classification Report -VGG16 Model

```
Classification Report VGG16 Model
              precision    recall  f1-score   support

      Benign       0.91      0.98      0.94       813
   Malignant       0.97      0.90      0.93       813

    accuracy                           0.94      1626
   macro avg       0.94      0.94      0.94      1626
weighted avg       0.94      0.94      0.94      1626


*****Performance Metrics*****

Accuracy:  0.9367
Precision:  0.9393
Recall:  0.9367
f1_metric:  0.9366
```

The Confusion Matrix of the VGG16 model showed that only 83 malignant cases were classified as benign and 20 benign were classified as malignant on the test dataset, as shown in figure 4.9.

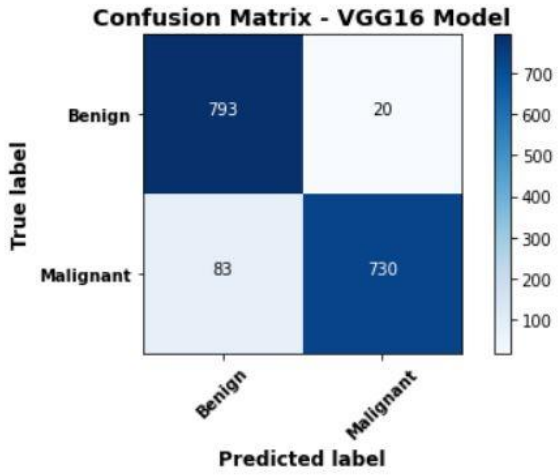

Figure 0.9: Confusion Matrix -VGG16 Model

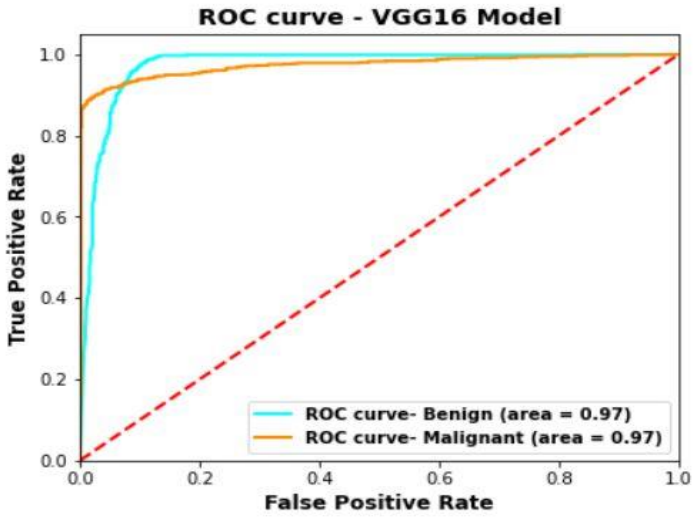

Figure 0.10: AUC/ROC curve -VGG16 Model

Figure 4.10 shows that the AUC of this model in both classes is 0.97, which is similar for both classes, and it also says that our model can recognize both classes as almost the same.

4.2.4 VGG19

The performance of the VGG19 model on the validation dataset had an accuracy of 94.44%, a precision of 94.61 %, a recall of 94.44%, and an F1-score of 94.43%, as the learning curve in figure 4.11 shows there was no overfitting, as it fluctuated around 91% to 95% on validation data. The test dataset had 93.42% accuracy, 93.58% precision, 93.42% recall, and 93.41% F1-score, as shown in Table 4.4.

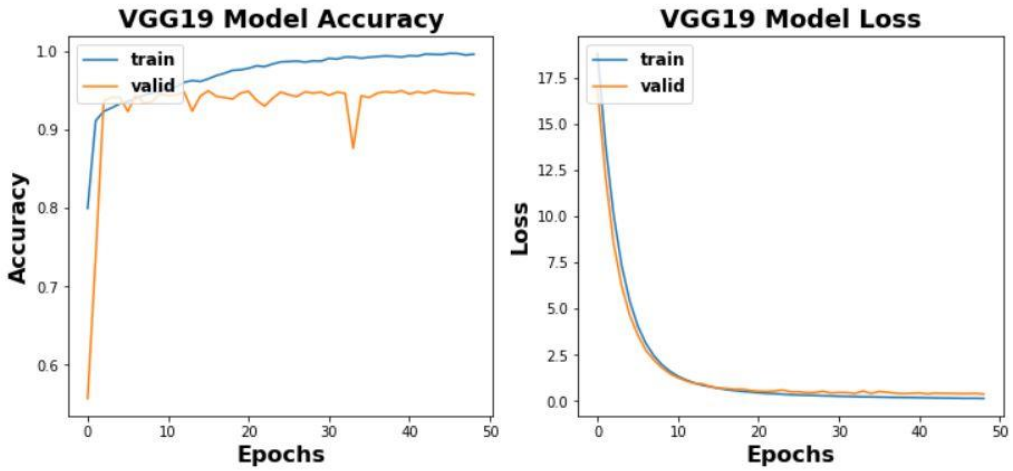

Figure 4.11: Learning Curve-VGG19 Model

Table 0.4: Classification Report -VGG19 Model

```
Classification Report VGG19 Model
              precision    recall  f1-score   support

      Benign       0.91      0.96      0.94       813
   Malignant       0.96      0.90      0.93       813

    accuracy                           0.93      1626
   macro avg       0.94      0.93      0.93      1626
weighted avg       0.94      0.93      0.93      1626


*****Performance Metrics*****

Accuracy:  0.9342
Precision:  0.9358
Recall:  0.9342
f1_metric:  0.9341
----------------------------------------------------------------
```

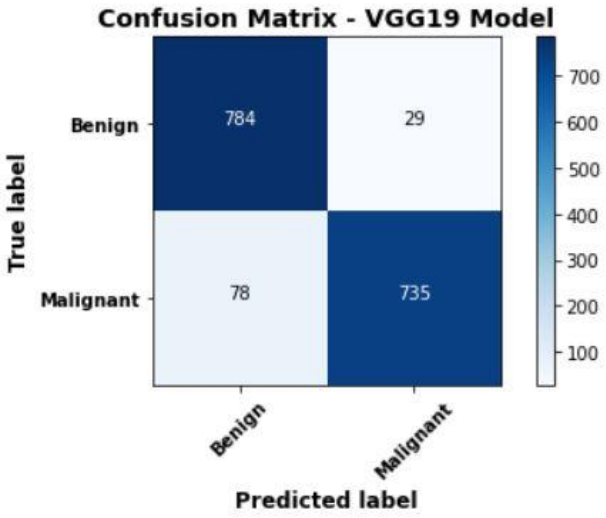

Figure 0.11: Confusion Matrix -VGG19 Model

The confusion matrix, in figure 4.12, shows that there were 78 malignant cases classified as benign, and 29 malignant ones were detected as benign.

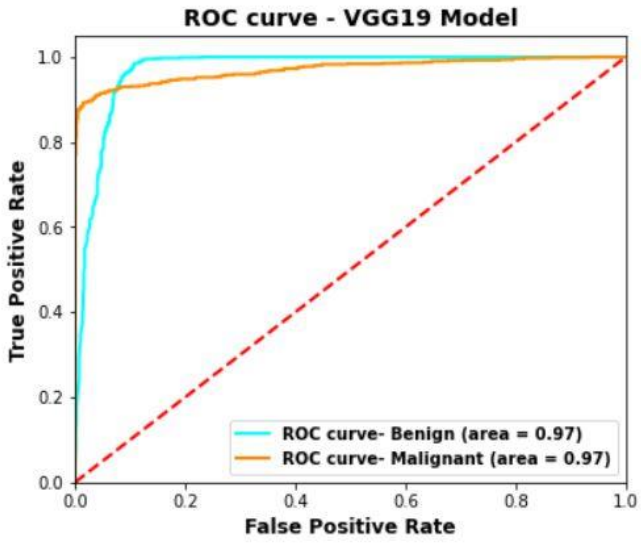

Figure 0.12: AUC/ROC curve -VGG19 Model

The number of malignant cases which had been misclassified is three times more than the benign. Figure 4.13 shows that the area under the curve is 0.97 and both classes follow similar behavior.

4.2.5 EfficientNetB7

The EfficientNetB7 had an accuracy of 94.44%, a precision of 94.54%, a recall of 94.44%, an F1-score of 94.43% on the validation data, and as shown in table 4.5, it had 92.99% accuracy, 93.24% precision, 92.99% re, call and 92.98% F1-score on the test dataset. As demonstrated in figure 4.14, there was no overfitting.

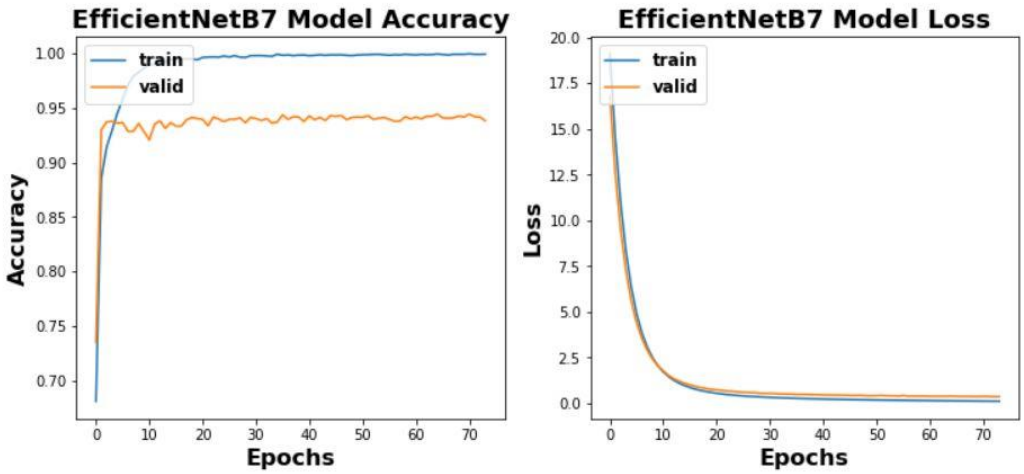

Figure 0.13: Learning Curve - EfficientNetB7 Model

Table 0.5: Classification Report - EfficientNetB7 Model

Classification Report EfficientNetB7 Model

	precision	recall	f1-score	support
Benign	0.90	0.97	0.93	813
Malignant	0.97	0.89	0.93	813
accuracy			0.93	1626
macro avg	0.93	0.93	0.93	1626
weighted avg	0.93	0.93	0.93	1626

*****Performance Metrics*****

Accuracy: 0.9299
Precision: 0.9324
Recall: 0.9299
f1_metric: 0.9298

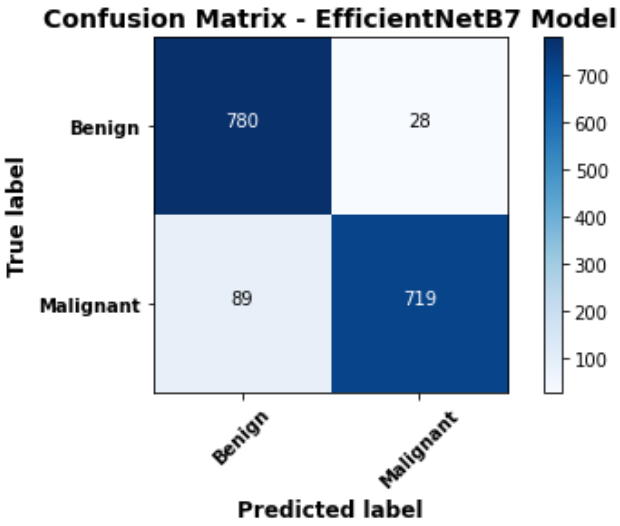

Figure 0.14: Confusion Matrix - EfficientNetB7 Model

Figure 4.15 shows that 89 malignant cases were detected as benign and only 28 benign ones were recognized as malignant.

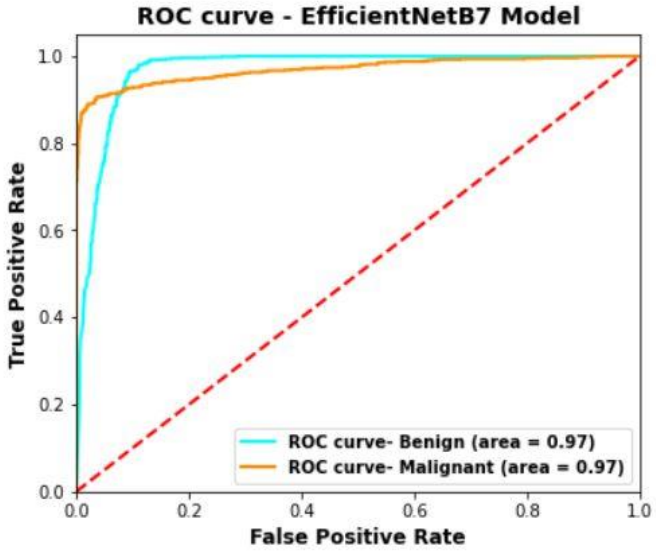

Figure 0.15: AUC/ROC curve - EfficientNetB7 Model

Based on these results, no overfitting was observed for this model. Moreover, the AUC of the model for each class was 0.97 and the trend was also the same, as shown in figure 4.16.

4.2.6 DenseNet169

The DenseNet169 model has an accuracy of 94.54%, a precision of 94.67%, a recall of 94.54%, and an F1-score of 94.54% on validation data. The learning curve, in figure 4.17, shows that there is no overfitting. As shown in Table 4.6, the test dataset's accuracy, precision, recall, and F1-score were 93.30%, 93.48%, 93.30%, and 93.29% respectively.

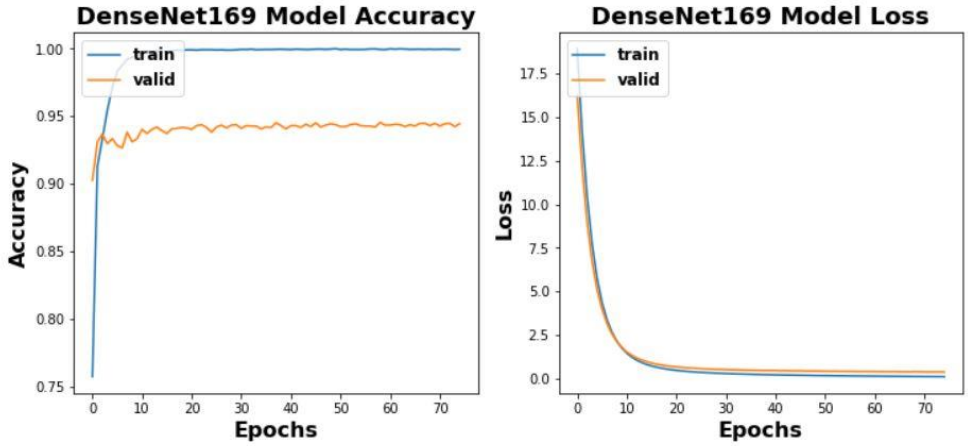

Figure 0.16: Learning Curve - DenseNet169 Model

Table 0.6: Classification Report - DenseNet169 Model

Classification Report DenseNet169 Model

	precision	recall	f1-score	support
Benign	0.91	0.97	0.94	813
Malignant	0.96	0.90	0.93	813
accuracy			0.93	1626
macro avg	0.93	0.93	0.93	1626
weighted avg	0.93	0.93	0.93	1626

*****Performance Metrics*****

Accuracy: 0.9330
Precision: 0.9348
Recall: 0.9330
f1_metric: 0.9329

The Confusion Matrix in figure 4.18 shows that the DenseNet169 model wrongly classified 85 malignant cases as benign and 20 benign ones as malignant.

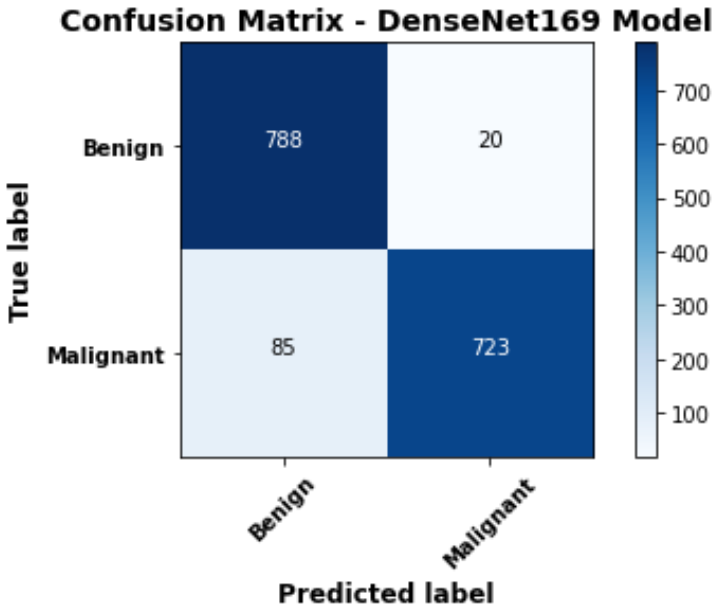

Figure 0.17: Confusion Matrix - DenseNet169 Model

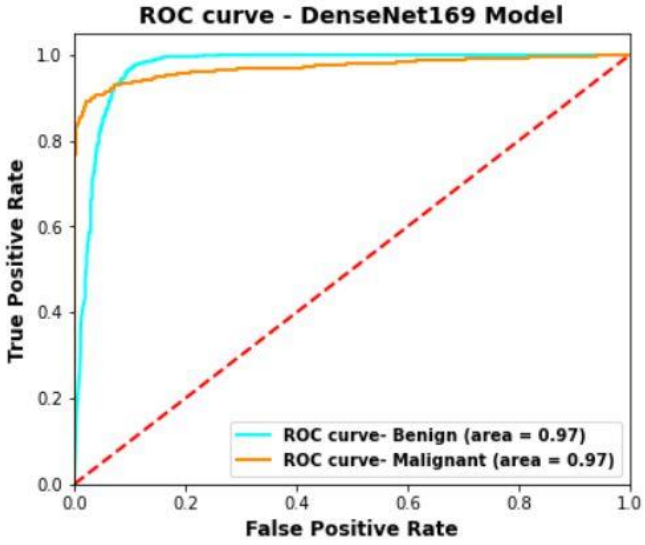

Figure 0.18: AUC/ROC curve - DenseNet169 Model

Figure 4.19 shows that the area under the curve for DenseNet169 has an exactly similar trend in both classes with a result of 0.97, which is similar to most of the models.

4.2.7 Xception

The Xception model produced an accuracy of 93.34%, a precision of 93.46%, a recall of 93.34%, and an F1-score of 93.33% on the validation dataset, while the result was 90.59% accuracy, 91.02% precision, 90.59% recall and 90.57% F1-score on the test dataset, as shown in Table 4.7.

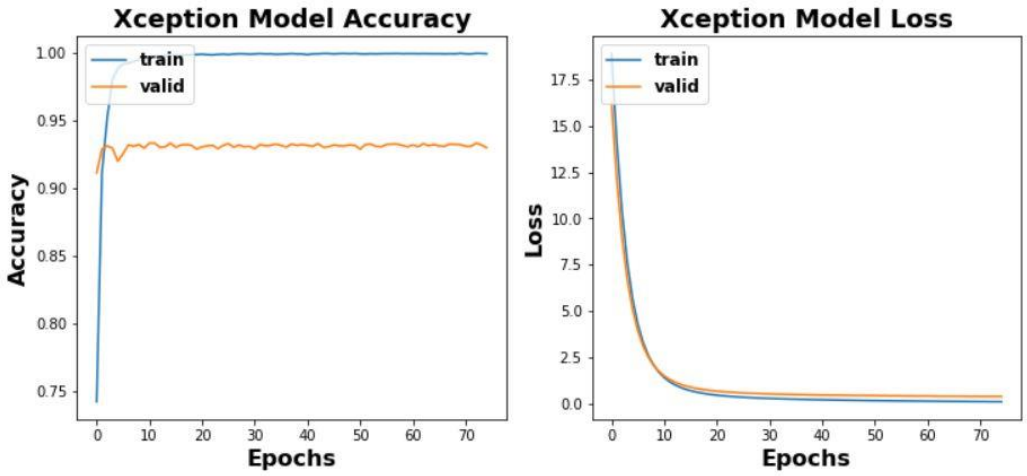

Figure 0.19: AUC/ROC curve - DenseNet169 Model

Table 0.7: Confusion Matrix - Xception Model

Classification Report Xception Model

	precision	recall	f1-score	support
Benign	0.87	0.96	0.91	813
Malignant	0.95	0.85	0.90	813
accuracy			0.91	1626
macro avg	0.91	0.91	0.91	1626
weighted avg	0.91	0.91	0.91	1626

*****Performance Metrics*****

Accuracy: 0.9059
Precision: 0.9102
Recall: 0.9059
f1_metric: 0.9057

In the confusion matrix, in figure 4.21, there are 98 and 45 images wrongly detected as benign and malignant respectively and it observed that the number of misclassifications in the malignant class is around double of the benign ones. The AUC of the model, which is shown in figure 4.22 is 0.95 for both classes with the same trend.

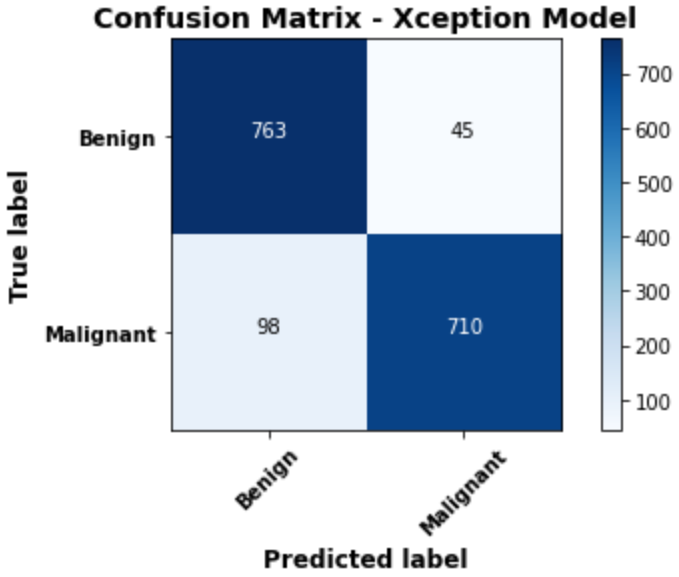

Figure 0.20: Confusion Matrix - Xception Model

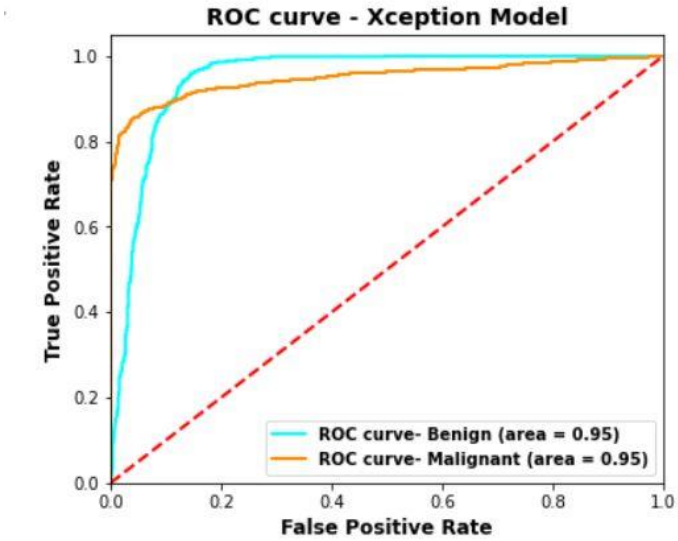

Figure 0.21: AUC/ROC curve - Xception Model

4.2.8 Inception V3

The results show that the Inception v3 model has an accuracy of 92.55%, a precision of 92.74%, a recall of 92.55%, and an F1-score of 92.54% in the validation dataset, and figure 4.23 shows that there is no overfitting. The model gave a 90.65% accuracy, a 91.33% precision, a 90.65% recall, and a 90.61% f1 score on the test dataset, as shown in Table 4.8.

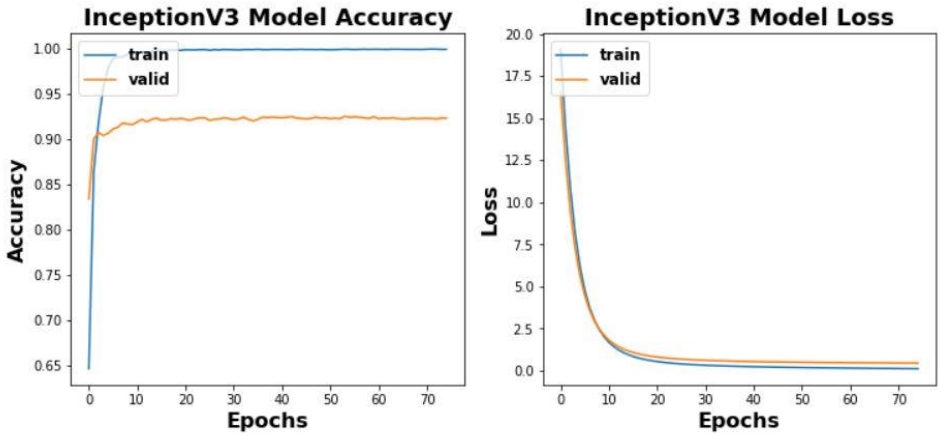

Figure 0.22: Learning Curve - Inception V3 Model

Table 0.8: Learning Curve - Inception V3 Model

Classification Report InceptionV3 Model

	precision	recall	f1-score	support
Benign	0.86	0.97	0.91	813
Malignant	0.97	0.84	0.90	813
accuracy			0.91	1626
macro avg	0.91	0.91	0.91	1626
weighted avg	0.91	0.91	0.91	1626

*****Performance Metrics*****

Accuracy: 0.9065
Precision: 0.9133
Recall: 0.9065
f1_metric: 0.9061

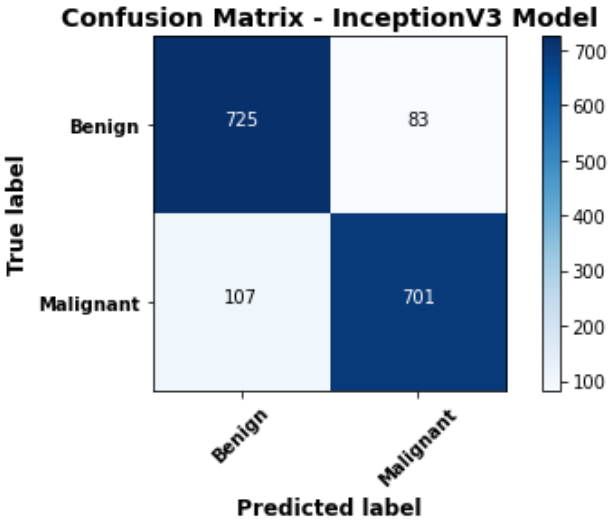

Figure 0.23: Confusion Matrix - Inception V3 Model

From figure 4.24, 107 malignant cases were classified as benign and 83 benign images were classified as malignant. Moreover, the area under the curve of this model is 0.96 for both classes as shown in figure 4.25.

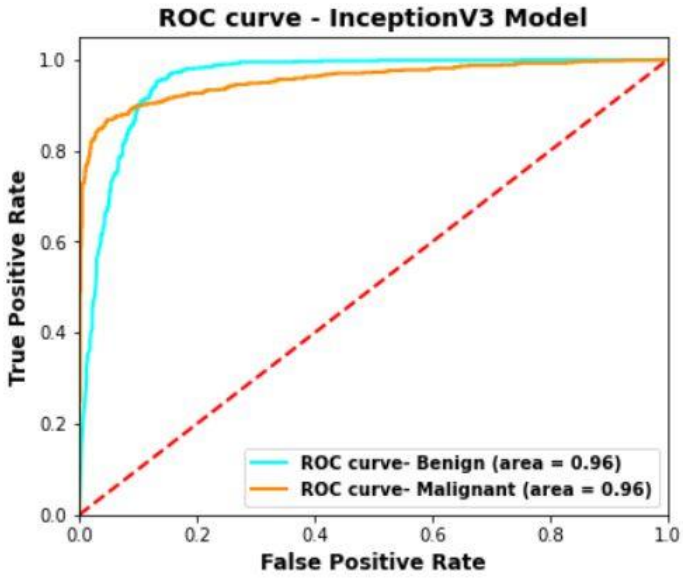

Figure 0.24: AUC/ROC curve - Inception V3 Model

4.2.9 Ensemble Model

In the ensemble approach, an averaging with different combinations of models is performed to achieve the highest result. In the following we will explain this in detail:

A. Ensemble with all models

After obtaining the average of all models' predictions, the performance of the ensemble model on the validation data becomes as the following: 95.02% accuracy, 95.30% precision, 95.03% recall, and 95.01% F1-score. In the same circumstances, the results on the test data are 93.33% accuracy, 93.71% precision, 93.30% recall, and 93.28% F1-score, as shown in Table 4.9. The confusion matrix of this approach in figure 4.26 shows 94 malignant cases and 15 benign cases were classified incorrectly.

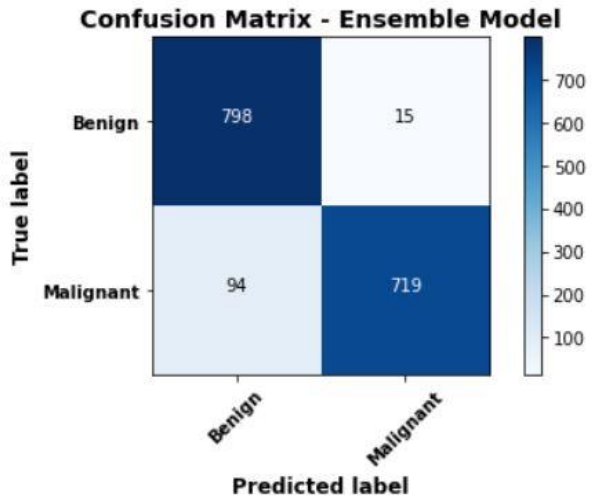

Figure 0.25: Confusion Matrix of Ensemble on all models

Table 0.9: Ensemble All Models Classification Report

Classification Report - Ensemble Model

	precision	recall	f1-score	support
Benign	0.89	0.98	0.94	813
Malignant	0.98	0.88	0.93	813
micro avg	0.93	0.93	0.93	1626
macro avg	0.94	0.93	0.93	1626
weighted avg	0.94	0.93	0.93	1626
samples avg	0.93	0.93	0.93	1626

*****Performance Metrics*****

Accuracy: 0.9330
Precision: 0.9371
Recall: 0.9330
f1_metric: 0.9328

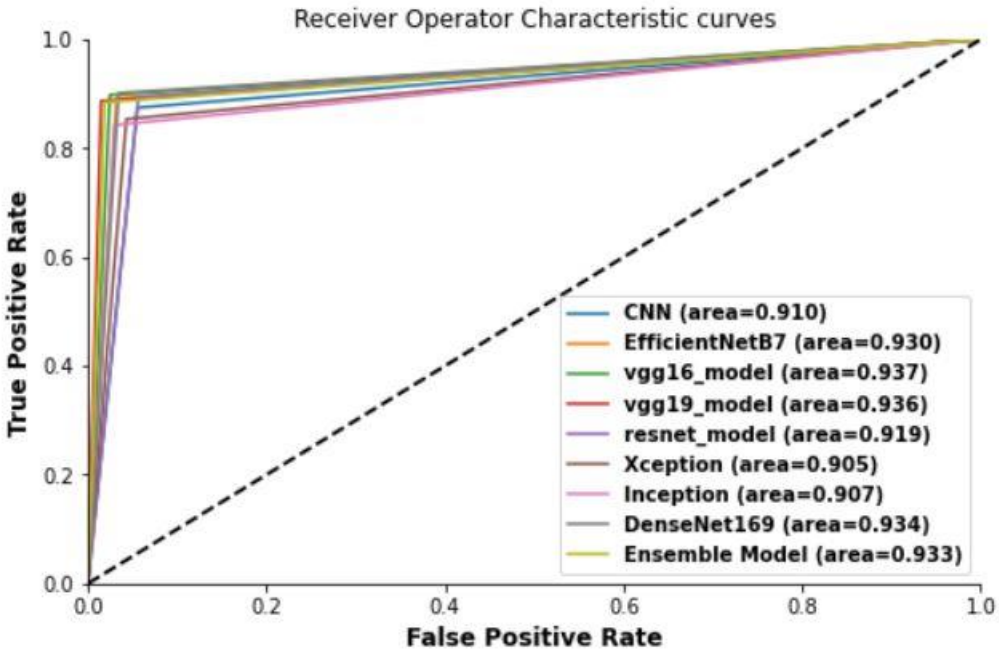

Figure 0.26: Roc Curve for All Models

Figure 4.26 showed the area under the curve for each model and the proposed ensemble model.

B. Ensemble All without InceptionV3 and Xception

After averaging all models' predictions, except Inceptionv3 and Xception, the performance on the validation data becomes 95.16% accuracy, 95.41% precision, 95.16% recall, and 95.15% F1-score. And as shown in table 4.10, the performance got 93.42% accuracy, 93.81% precision, 93.42% recall, and 93.40% F1-score for test data. The confusion matrix of this approach in figure 4.27 shows 95 malignant cases and 15 benign cases were classified incorrectly.

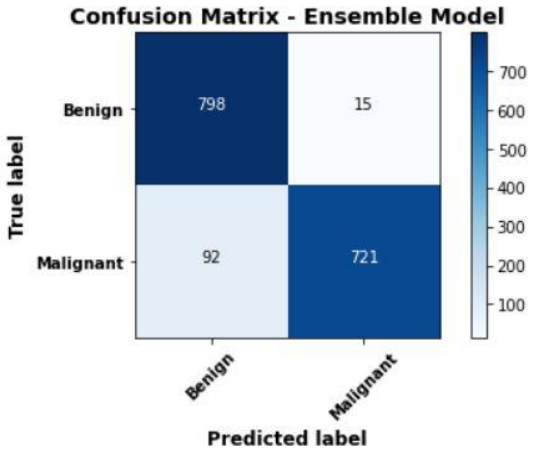

Figure 0.27: Confusion Matrix of Ensemble All without InceptionV3 and Xception

Table 0.10: Classification Report for Ensemble All Except Inception and Xception Models

Classification Report - Ensemble Model

	precision	recall	f1-score	support
Benign	0.90	0.98	0.94	813
Malignant	0.98	0.89	0.93	813
micro avg	0.93	0.93	0.93	1626
macro avg	0.94	0.93	0.93	1626
weighted avg	0.94	0.93	0.93	1626
samples avg	0.93	0.93	0.93	1626

*****Performance Metrics*****

Accuracy: 0.9342
Precision: 0.9381
Recall: 0.9342
f1_metric: 0.9340

There were 92 and 15 were detected wrongly, in this model the result improves a bit after taking those architectures away.

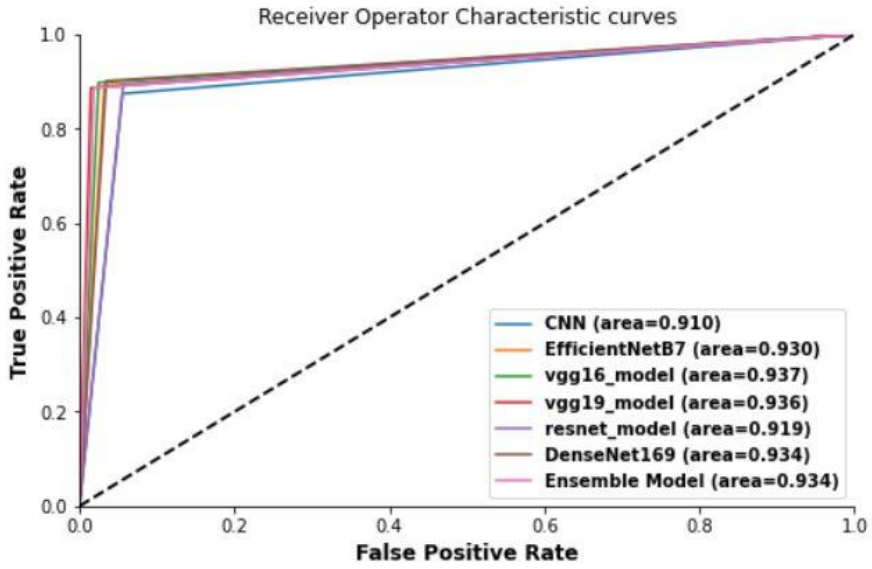

Figure 0.28: Roc Curve for All Models Except Inception and Xception

Figure 4.28 showed the area under the curve for each model and the proposed ensemble model.

C. Ensemble Best 5 Models CNN, EfficientNetB7, VGG16, VGG19 and DenseNet196

In this approach, the selected models; CNN, EfficientNetB7, VGG16, VGG19, and DenseNet196, gave performance metrics, after averaging prediction of selected models, of 94.99% and 93.79% accuracy, 95.28% and 94.18% precision, 94.99% and 93.79% recall, 94.98% and 93.77% F1-score for both validation and test datasets respectively, as shown in Table 4.11. The confusion matrix shows 89 misclassifications for malignant cases and 12 for benign ones, as shown in figure 4.29.

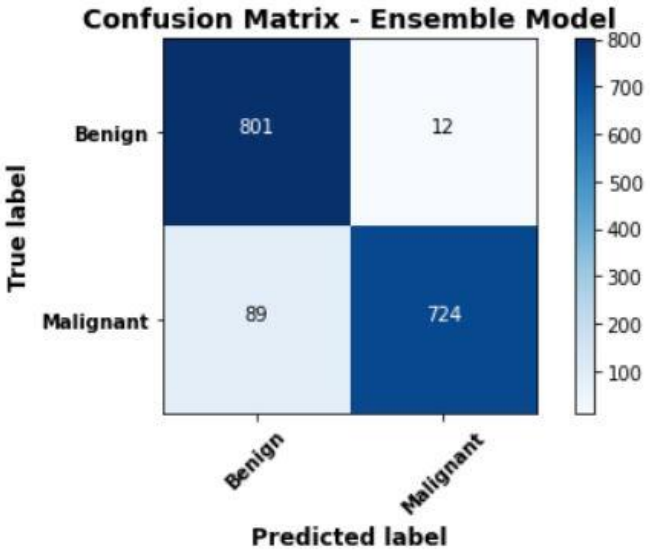

Figure 0.29: Confusion Matrix of Ensemble Best 5 Models

Table 0.11: Classification Report of Ensemble Best 5 Models

Classification Report - Ensemble Model

	precision	recall	f1-score	support
Benign	0.90	0.99	0.94	813
Malignant	0.98	0.89	0.93	813
micro avg	0.94	0.94	0.94	1626
macro avg	0.94	0.94	0.94	1626
weighted avg	0.94	0.94	0.94	1626
samples avg	0.94	0.94	0.94	1626

*****Performance Metrics*****

Accuracy: 0.9379
Precision: 0.9418
Recall: 0.9379
f1_metric: 0.9377

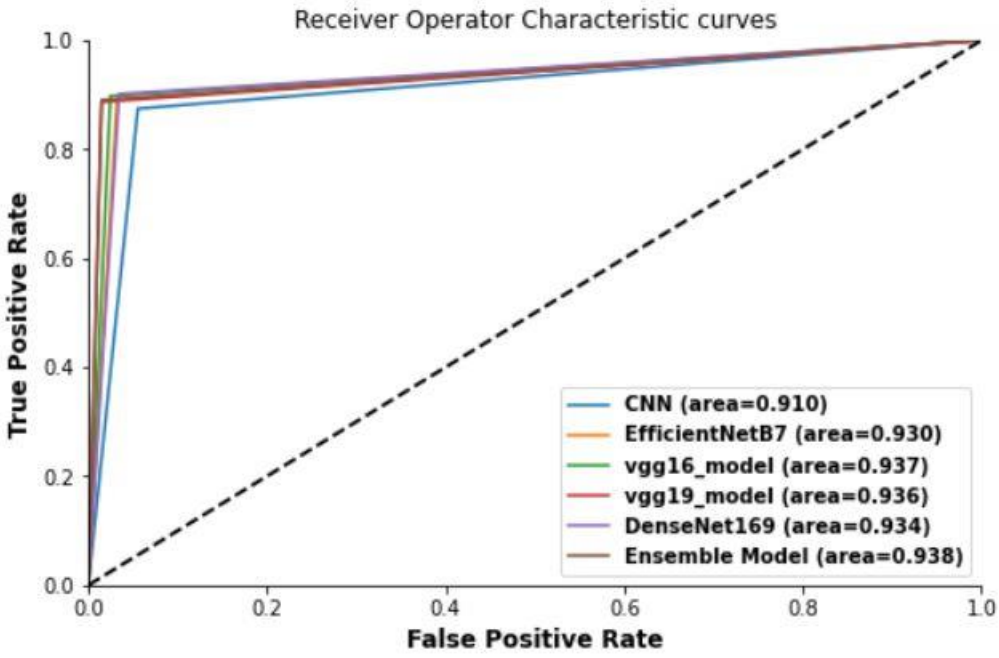

Figure 04.30: Roc Curve for Ensemble Best 5 Models

Figure 4.30 showed the area under the curve for each model and the proposed ensemble model.

D. Ensemble for best three models, EfficientNetB7, VGG19, and DenseNet196

Here we selected the top 3 high-performance models to average their predictions; the selected models are EfficientNetB7, VGG19, and DenseNet196. They achieved 94.85% accuracy, 95.08% precision, 94.85% recall, and 94.85% f1-score on the validation dataset and obtained 93.36%, 93.72%, 93.36%, 93.34% for accuracy, precision, recall, and f1-score respectively on the test dataset as shown in table 4.12. The confusion matrix shows 91 misclassifications for malignant cases and 17 for benign ones, as shown in figure 4.32.

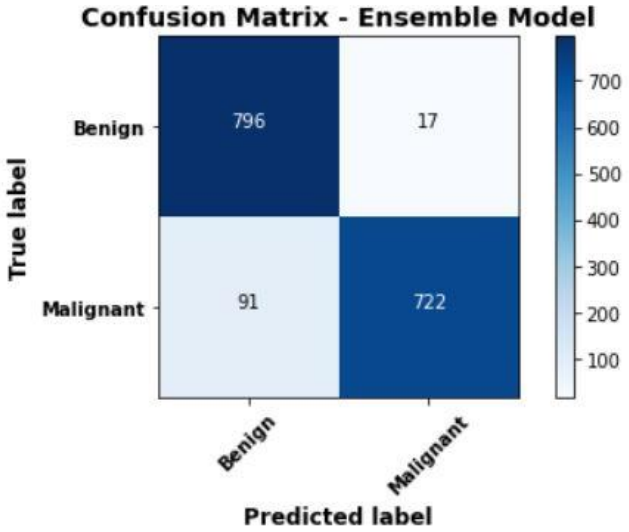

Figure 4.31: Confusion Matrix of Ensemble Best 3 Models

Table 0.12: Classification Report of Ensemble of Best 3 Models

Classification Report - Ensemble Model

	precision	recall	f1-score	support
Benign	0.90	0.98	0.94	813
Malignant	0.98	0.89	0.93	813
micro avg	0.93	0.93	0.93	1626
macro avg	0.94	0.93	0.93	1626
weighted avg	0.94	0.93	0.93	1626
samples avg	0.93	0.93	0.93	1626

*****Performance Metrics*****

Accuracy: 0.9336
Precision: 0.9372
Recall: 0.9336
f1_metric: 0.9334

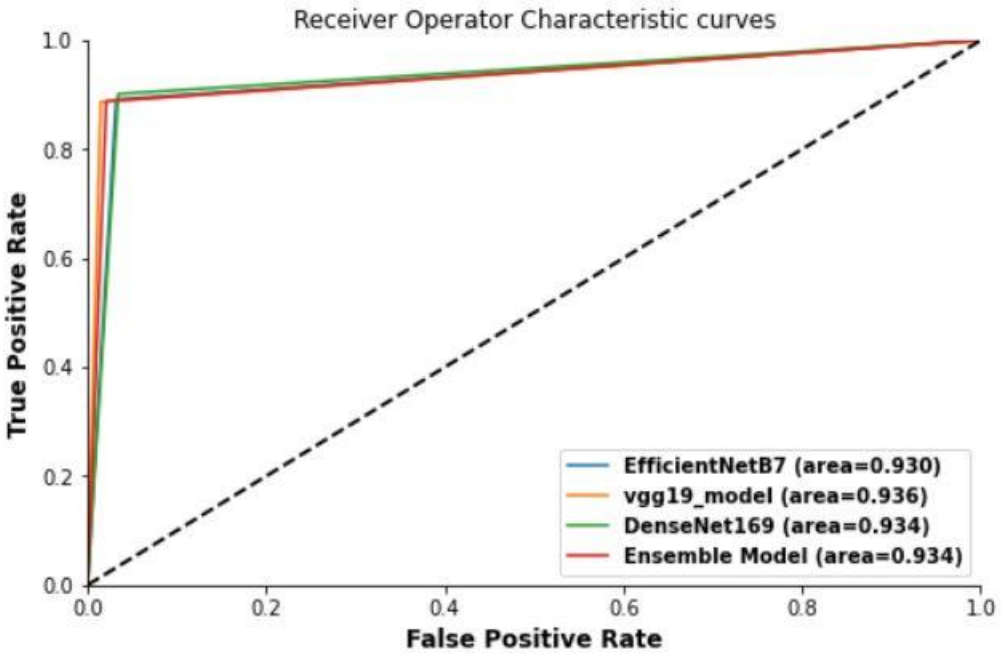

Figure 0.32: Roc Curve for of Ensemble Best 3 Models

Figure 4.32 showed the area under the curve for each model and the proposed ensemble model.

E. Ensemble for best four models, EfficientNetB7, VGG19, VGG16, and DenseNet196

As for our final approach, the four top models were selected for the ensemble approach and the obtaining performance was 95.30% accuracy, 95.55% precision, 95.30% recall, and 95.29% f1-score for validation dataset and 93.73% accuracy, 94.11% precision, 93.73% recall and 93.71% f1-score for the test dataset as shown in table 4.13. Furthermore, the confusion matrix in figure 4.34 shows 89 images were recognized as benign and 13 images were detected as malignant incorrectly.

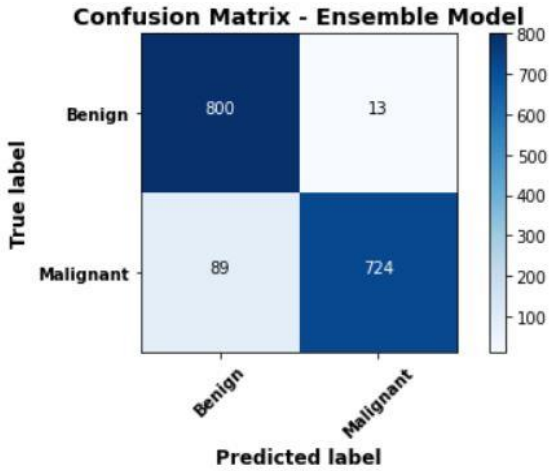

Figure 0.33: Confusion Matrix of Ensemble Best 4 Models

Table 0.13: Classification Report of Ensemble Best 4 Models

Classification Report - Ensemble Model

	precision	recall	f1-score	support
Benign	0.90	0.98	0.94	813
Malignant	0.98	0.89	0.93	813
micro avg	0.94	0.94	0.94	1626
macro avg	0.94	0.94	0.94	1626
weighted avg	0.94	0.94	0.94	1626
samples avg	0.94	0.94	0.94	1626

*****Performance Metrics*****

Accuracy: 0.9373
Precision: 0.9411
Recall: 0.9373
f1_metric: 0.9371

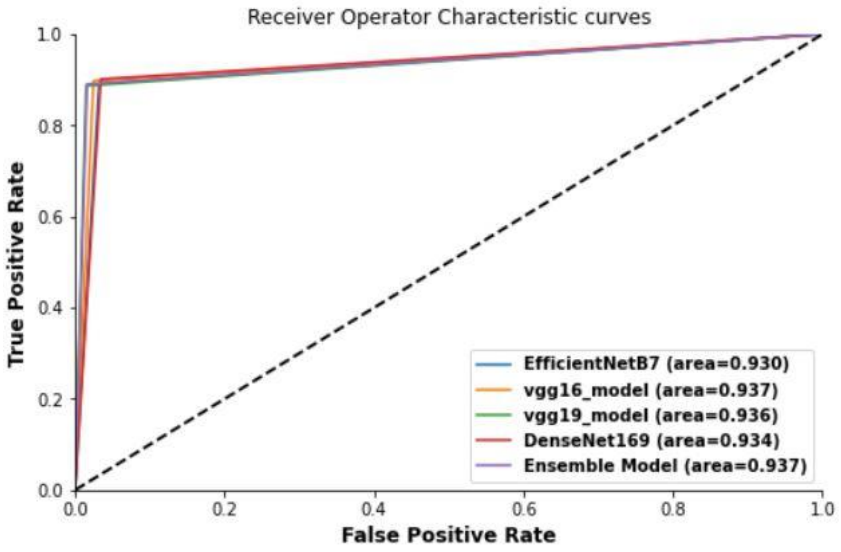

Figure 0.34: Roc Curve for of Ensemble Best 4 Models

Figure 4.34 showed the area under the curve for each model and the proposed ensemble model.

4.3 Models Comparison

The performance summary of all individual models is shown in table 4.1 and figure 4.44. In this work, customized CNN and 7 different pre-trained models, trained on microscopic images of skin cancer, in addition to 5 various ensemble approaches were performed with the goal of performance improvement. Table 4.14 shows a comparison of all models' accuracy results, the ensemble of the best five models obtained the highest accuracy on test data. In addition, all the ensemble models achieved higher accuracy as compared to single models on the test dataset, except for vgg16 and vgg19. In the case of dataset validation, the ensemble model of the best four models with an accuracy of 95.3% was the top model, while in the case of individual models; single vgg16 has the best validation accuracy with 95.12%. While in the test dataset case, the top single model was vgg16 with 93.67 accuracies, 93.93% precision,

93.67 recall, 93.66 F1-score, and AUC 0.937. Moreover, the top ensemble model was the ensemble approach of the best five models with 93.79 accuracies, 94.18 precision, 93.79 recall, 93.77 f1-score, and AUC 0.938

Table 0.14: Models Results in Comparison

Models	Val Accuracy	Accuracy	Precision	Recall	F1_Score	AUC_Score
CNN	91.72	90.96	91.16	90.96	90.95	0.91
EfficientNet	94.44	92.99	93.24	92.99	92.98	0.93
Vgg16	95.12	93.67	93.93	93.67	93.66	0.937
Vgg19	94.44	93.42	93.58	93.42	93.41	0.936
DenseNet169	94.54	93.3	93.48	93.3	93.29	0.934
ResNet50	94.02	91.88	91.98	91.88	91.88	0.919
Xception	93.34	90.59	91.02	90.59	90.57	0.905
Inception	92.55	90.65	91.33	90.65	90.61	0.907
==========	====	====	====	====	====	====
Ensemble Model- ALL Models	95.02	93.33	93.71	93.3	93.28	0.933
Ensemble Model-ALL Except Inception and Xception	95.16	93.42	93.81	93.42	93.4	0.934
Ensemble Model- Best 3(Efficient_vgg19_densenet)	94.85	93.36	93.72	93.36	93.34	0.934
Ensemble Model-Best 4(Efficient_Vgg16_vgg19_densenet)	95.3	93.73	94.11	93.73	93.71	0.937
Ensemble Model- Best 5(CNN_Efficient_Vgg16_vgg19_densenet)	94.99	93.79	94.18	93.79	93.77	0.938

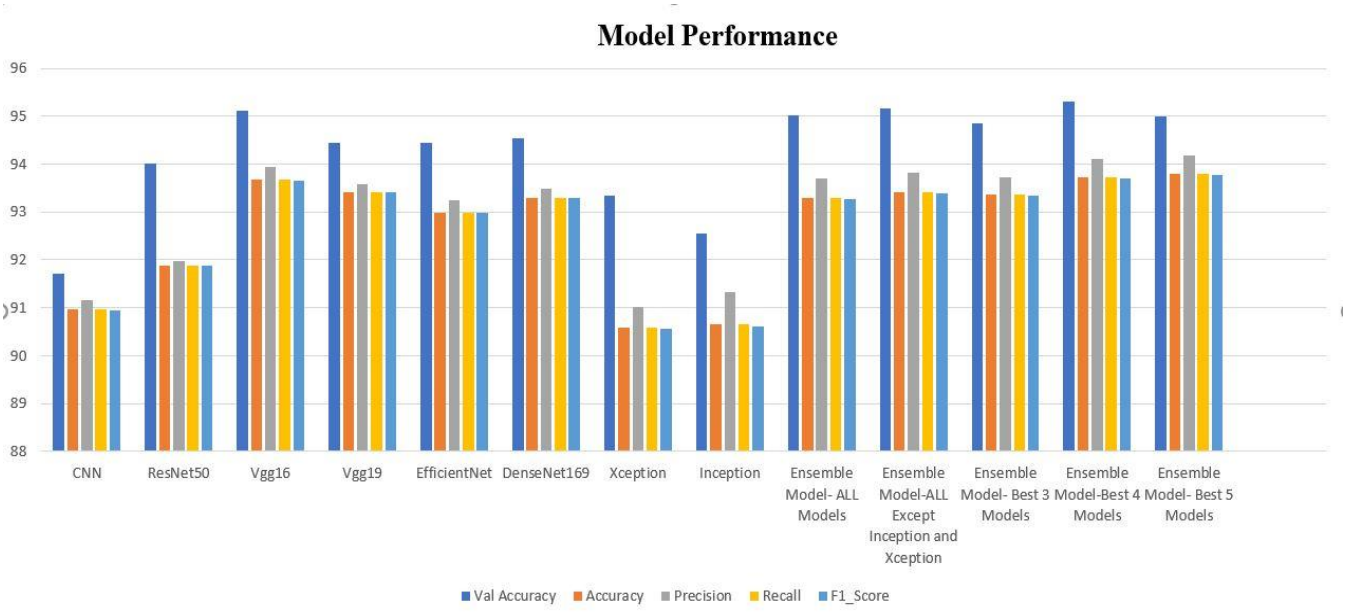

Figure 0.35: Models Results

On the other hand, the InceptionV3 and Xception models stand at the lowest level with 90.65% and 90.59% accuracy on test datasets respectively.

In this thesis, the max voting ensemble technique was used to improve the performance of the final model. This technique is used when there are models with different algorithms or configurations, thus, combining the models' predictions by selecting the most frequent answer among all models, helps to reduce variance and the chance of overfitting and improves final performance. It is clear, the outcome of the ensemble model is completely related to the performance of the models because, more correct predictions for each sample image, create a higher probability of selecting the correct prediction during voting. Therefore, selecting the top five models give the confidence to have a greater accuracy by ensemble approach.

4.4 Comparisons with other works

In this section, the performance of our proposed model is compared with several state-of-the-art approaches that had been put forth to identify skin cancer using microscopic images. The outcomes of several state-of-the-art methods described in the literature review are compared in table 4.2 with the corresponding proposed method.

Table 0.15: Dataset-Related Research Results

Author	Year	Model	Accuracy	AUC_Score
Mohammed Rakeibul Hassan et.al [34]	2021	VGG16	93.18	-
Kartikeya Agarwal [20]	2018	DenseNet	86	-
Michael Cabanillas [37]	2021	CNN	88.87	-
Satin Jain et al [26]	2021	Xception	90.48	-
Jaisakthi S M [40]	2022	EfficentNet	-	0.9465
Abhinav Sagar et.al [39]	2020	ResNet50	93.5	-
Khalil Aljohani [31]	2022	VGG19	76	-
Qishan Ha et.al [24]	2020	Ensemble	-	0.95
==========	====	========	====	====
Proposed Method	2023	Ensemble Model- Best 5(CNN_Efficient_Vgg16_vgg19_densenet)	93.7	0.938

Table 4.2 shows that the proposed ensemble model outperforms all other models of previous works, which were mentioned in the literature, with 94.37% accuracy and

0.978 AUC. Abhinav Sagar et.al, in 2020, applied ResNet50 and could achieve 93.5% accuracy, and Mohammed Rakeibul Hassan et.al, in 2021, used VGG16 and could achieve 93.18% accuracy.

Chapter 5

Conclusion and Future Work

5.1 Introduction

A summary of the whole process and findings of this work will be explained in this chapter. After that, some future works which may be helpful to boost performance will be suggested.

5.2 Conclusion

The main reason for skin cancer is overexposure to sunlight and its probability increases when sunburn and blistering occur regularly. UV rays from the sun affect the DNA of the skin and cause the growth of abnormal cells which divide fast in a disorganized way and create a mass of cancer cells. They have various symptoms but, mostly a change in the growth of the skin occurs or old moles appear. The number of people who suffer from skin cancer is not the same in every country and completely relates to various factors, such as; the level of sunlight, the genetic history of people, food style, pollution of the environment, etc. Although various methods of treatment have been invented and developed, still, many people die because of skin cancer every year. We found whilst doing this work, that the main reason for the high number of deaths due to skin cancer is late detection. This occurs because most of the time the process of detection is done by experts and the number of specialists in such domains is low, especially in a country with fewer skin cancer cases. Another reason is that

people, especially in counties with rare skin cancer cases, are not aware of how they can protect themselves from getting the disease and what its symptoms are. In addition, the lack of enough experience increases human mistakes, and disagreements that happen between specialists during the process of diagnosis impact accurate detection. All of the aforementioned often cause postponement of the healing process for patients, as a result, the percentage of successful treatments is not that high. It was observed that the process of detection begins with verifying clinical samples and classifying them into cancerous or non-cancerous. The concept of classification means assigning a label to each group of data based on their similarity factors, and this is what we do every day while deciding between various solutions. This selection can be done between two different subjects or more. Fortunately, machine learning algorithms and deep learning facilitate this process and make it possible to have extremely correct decisions in many circumstances. Algorithms try to fetch every detail from the given data and based on the collected information, and their features, they find the most suitable label for that specific data. For the algorithms to be able to do that, they need enough samples of each class during the training phase. In this work, our goal was a binary classification and the dataset did not have enough data in the malignant class, hence, additional preprocessing steps were done while modeling to handle this problem before feeding our algorithms.

As mentioned before, our data is image type, which contains a huge amount of conceptual data based on previous research, so, traditional ML algorithms cannot perform well in this type of problem because computational complexity is high. However, today, as a result of vast improvements in deep learning algorithms, which is

a subset of machine learning, powerful image classification models have been developed and have been conducted in the real world. For instance, in cases like animal breed detection, driverless cars, and many others. Also, it can be used for medical problems, such as; breast cancer, alimentary tract cancer detection, or prediction of kidney diseases. Thus, the process of skin cancer detection can be accelerated and become more affordable if a precise application is developed to classify images, whether benign or malignant, and that is the golden target of this work.

For achieving this goal, we utilized CNN architecture and some state-of-the-art transfer learning CNN architectures that were trained on immerse images from 1000s of classes, which include; ResNet50, VGG16, VGG19, EfficientNetB7, DenseNet169, Xception, and Inception v3. The best individual model was VGG16 which outperformed as compared to all other architectures, and the best ensemble model was an ensemble of five models (CNN, VGG16, VGG19, EfficientNetB7, DenseNet169), Its accuracy is good enough to persuade doctors to start treatment at the earliest time. Although wrong detection imposes a lot of stress and costs for the patient, as well as for doctors, the number of wrong detections was way less in this model in comparison to others, among all which makes the model trustable.

Thanks to researchers who have spent time developing extraordinary methods and techniques, we can now use ensemble learning with minimum effort, and then we combine the results of our models by summing them. Although the accuracy of this model is a bit less than our proposed model, the false detections have been reduced in both classes. Thus, this model is not only able to detect skin cancer with acceptable certainty but also, decreases the number of wrong diagnoses, which reduces a lot of

effort and difficulties for people, we got slightly higher results with the proposed ensemble approach, and the best ensemble was for ensemble model of five models (CNN, VGG16, VGG19, EfficientNetB7, and DenseNet196) together.

We did not implement complex preprocessing, we only implemented resizing to prepare requirements of architecture and normalizing to reduce the computational cost. In the augmentation matter, two steps were implemented; the first step helped us solve the imbalance problem, and the second one was conducted to feed architectures with more diverse data, by changing the direction of images. Moreover, we did not employ complex architectures and tried to keep them as simple as possible, and this helped us to easily manage the cost.

5.3 Challenges and Limitations

The first challenge we faced in this work was the imbalanced data. As it was mentioned before, the number of malignant images was way less when compared to the benign class. This would impact the performance of any model; therefore, it was important to deal with this issue. Here, augmentation techniques, such as; rotation, zooming, random distortion, random contrasting, random change in brightness, and random cropping were applied to malignant images to increase the number of samples in this class. The second challenge was ensemble learning; as the performance of CNN and pre-trained architectures were not close to each other, the normal ensemble approach did not have a higher performance than most of the single models. To solve this issue, five different combinations of models were selected, and for each group, the average ensemble technique was applied. The best outcome of all the groups compared

together is the combination of CNN, VGG16, VGG19, EfficientNetB7, and DenseNet196 models as they transcend all the other groups.

The other challenge was generalizing the models, which are related to the inputs. In the medical field, the direction of microscopy images can be diverse; therefore, preparing suitable data for models to help them see data of different directions and collect more precise information during training is essential. For this matter, we applied online flipping horizontally and vertically to help models to generalize well.

5.4 Future Work

For future improvement, other techniques of preprocessing, such as color transformation can be used, because the edges of photos do not have much information in some cases and color transformation can create some contrast between normal and infected skin. Also, Generative Adversarial Networks (GAN) Model can be used to generate realistic images to balance the dataset.

Considerations of the high performance of transfer learning architectures in contrast to designed CNN, and other architectures such as ResNet101, InceptionResNetV2, MobileNet, and AlexNet can be tested, too. There are different methods of ensemble learning, and the simple one is used here to control the cost. So, other techniques, such as; weighted average or machine learning algorithms can be tested in future works. Finally, there were not enough images for different types of skin cancer in the dataset, thus, we did not perform multiclass classification. Multiclass classification is the case of the availability of sufficient data that will be helpful to select the best treatment path. So, future improvements on this side need to be taken into consideration.

References

[1] Yamaguchi, Yuji, Janusz Z. Beer, and Vincent J. Hearing. "Melanin mediated apoptosis of epidermal cells damaged by ultraviolet radiation: factors influencing the incidence of skin cancer." Archives of Dermatological Research 300.1 (2008): 43-50.

[2] Gordon, Randy. "Skin cancer: an overview of epidemiology and risk factors." Seminars in oncology nursing. Vol. 29. No. 3. WB Saunders, 2013.

[3] Haggenmüller, S., Maron, R.C., Hekler, A., Utikal, J.S., Barata, C., Barnhill, R.L., Beltraminelli, H., Berking, C., Betz-Stablein, B., Blum, A. and Braun, S.A., "Skin cancer classification via convolutional neural networks: a systematic review of studies involving human experts.," European Journal of Cancer, 156, pp.202-216., 2021.

[4] Silpa, S. & V, Chidvila, "A REVIEW ON SKIN CANCER," International Research Journal of Pharmacy. 4. 10.7897/2230-8407.04814., 2013.

[5] Brinker, T.J., Hekler, A., Utikal, J.S., Grabe, N., Schadendorf, D., Klode, J., Berking, C., Steeb, T., Enk, A.H. and Von Kalle, C., "Skin cancer classification using convolutional neural networks: a systematic review.," Journal of medical Internet research, 20(10), p.e11936., 2018.

[6] Chaturvedi, S.S., Tembhurne, J.V. and Diwan, T., 2020. A multi-class skin Cancer classification using deep convolutional neural networks. Multimedia Tools and Applications, 79(39), pp.28477-28498.

[7] Jinnai, S., Yamazaki, N., Hirano, Y., Sugawara, Y., Ohe, Y. and Hamamoto, R., 2020. The development of a skin cancer classification system for pigmented skin lesions using deep learning. Biomolecules, 10(8), p.1123.

[8] Das, Kinnor, et al. "Machine learning and its application in skin Cancer." International Journal of Environmental Research and Public Health 18.24 (2021): 13409.

[9] Höhn, J., Hekler, A., Krieghoff-Henning, E., Kather, J.N., Utikal, J.S., Meier, F., Gellrich, F.F., Hauschild, A., French, L., Schlager, J.G. and Ghoreschi, K., 2021. Integrating patient data into

skin cancer classification using convolutional neural networks: systematic review. Journal of Medical Internet Research, 23(7), p. e20708.

[10] Joanna Korjakowska, Study on the Smartphone Application for Early Detection of Malignant Melanoma, 2018.

[11] Dorj, U.O., Lee, K.K., Choi, J.Y. and Lee, M., 2018. The skin cancer classification using deep convolutional neural network. Multimedia Tools and Applications, 77(8), pp.9909-9924.

[12] Estee and Allan Current and emerging technologies in melanoma diagnosis: the state of the art,2009.

[13] Siegel , Hannah , Ahmedin , A Cancer Journal for Clinicians,2022.

[14] Apalla Z, Lallas A, Sotiriou E, Lazaridou E, Ioannides D. Epidemiological trends in skin cancer. Dermatol Pract Concept 2017;7(2):1.

[15] Hussain, S., Mubeen, I., Ullah, N., Shah, S.S.U.D., Khan, B.A., Zahoor, M., Ullah, R., Khan, F.A. and Sultan, M.A., 2022. Modern Diagnostic Imaging Technique Applications and Risk Factors in the Medical Field: A Review. BioMed Research International, 2022.

[16] Haenssle, H.A., Fink, C., Schneiderbauer, R., Toberer, F., Buhl, T., Blum, A., Kalloo, A., Hassen, A.B.H., Thomas, L., Enk, A. and 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), pp.1836-1842.

[17] Curiel-Lewandrowski, C. et al. (2019). Artificial Intelligence Approach in Melanoma. In: Fisher, D., Bastian, B. (eds) Melanoma. Springer, New York, NY. https://doi.org/10.1007/978-1-4614-7322-0_43-1.

[18] Kaur, R.; Gholam Hosseini, H.; Sinha, R.; Lindén, M. Melanoma Classification Using a Novel Deep Convolutional Neural Network with Dermoscopic Images. (2022), Sensors 2022, 22, 1134.

[19] Kalouche, S., Ng, A. and Duchi, J., 2016. Vision-based classification of skin cancer using deep learning. 2015, conducted on Stanfords Machine Learning course (CS 229) taught.

[20] Kumar, V. and Choudhury, T., 2019. Real-time recognition of malignant skin lesions using ensemble modeling.

[21] Agarwal, K. and Singh, T., 2022. Classification of Skin Cancer Images using Convolutional Neural Networks. arXiv preprint arXiv:2202.00678.

[22] Cruz, M.V., Namburu, A., Chakkaravarthy, S., Pittendreigh, M. and Satapathy, S.C., 2020. Skin cancer classification using convolutional capsule network (CapsNet), 2020.

[23] Bill Cassidy, Connah Kendrick, Andrzej Brodzicki, Joanna Jaworek-Korjakowska, Moi Hoon Yap, Analysis of the ISIC image datasets: Usage, benchmarks and recommendations, (2022), Science Direct, Medical Image Analysis, Volume 75,102305, ISSN 1361-8415,

[24] Y. Zhang and C. Wang, "SIIM-ISIC Melanoma Classification with DenseNet," (2021), IEEE 2nd International Conference on Big Data, Artificial Intelligence and Internet of Things Engineering (ICBAIE), 2021, pp. 14-17.

[25] Qishen Ha, Bo Liu, Fuxu Liu, Identifying Melanoma Images using EfficientNet Ensemble: Winning Solution to the SIIM-ISIC Melanoma Classification Challenge, 2020, arXiv:2010.05351.

[26] Elansary, I., Ismail, A., Awad, W. (2022). Efficient Classification Model for Melanoma Based on Convolutional Neural Networks. In: Hassanien, A.E., Bhatnagar, R., Snášel, V., Yasin Shams, M. (eds) Medical Informatics and Bioimaging Using Artificial Intelligence, Studies in Computational Intelligence, vol 1005. Springer.

[27] Jain, S., Singhania, U., Tripathy, B., Nasr, E.A., Aboudaif, M.K. and Kamrani, A.K., 2021. Deep Learning-Based Transfer Learning for Classification of Skin Cancer. Sensors, 21(23), p.8142.

[28] S. Das and D. Das, "Skin Lesion Segmentation and Classification: A Deep Learning and Markovian Approach," (2021), IEEE Mysore Sub Section International Conference (MysuruCon), 2021, pp. 546-551.

[29] SIIM-ISIC melanoma classification Dataset, Anna Zawacki, Brian Helba, George Shih, Jochen Weber, Julia Elliott, Marc Combalia, Nicholas Kurtansky, NoelCodella, Phil Culliton, Veronica Rotemberg, Kaggle, 2020.

[30] Dildar, M., Akram, S., Irfan, M., Khan, H.U., Ramzan, M., Mahmood, A.R., Alsaiari, S.A., Saeed, A.H.M., Alraddadi, M.O. and Mahnashi, M.H., 2021. Skin cancer detection: a review using deep learning techniques. International journal of environmental research and public health, 18(10), p.5479.

[31] S. Sasikala, S. Arun Kumar, S.N. Shivappriya, Priyadharshini T, "Towards Improving Skin Cancer Detection Using Transfer Learning," Biosc.Biotech.Res. Comm. Special Issue Vol 13 No 11 (2020) Pp-55-60, 2020.

[32] Aljohani, K.; Turki, T, "Automatic Classification of Melanoma Skin Cancer with Deep Convolutional Neural Networks," AI 2022, 3, 512–525, 2022.

[33] Kaur, R.; GholamHosseini, H.; Sinha, R.; Lindén, M. Melanoma Classification Using a Novel Deep Convolutional Neural Network with Dermoscopic Images. (2022), Sensors 2022, 22, 1134.

[34] Kumar A and Vatsa A, (2022) Untangling Classification Methods for Melanoma Skin Cancer. Front. Big Data 5:848614.

[35] Mohammed Rakeibul Hasan, Mohammed Ishraaf Fatemi, Mohammad Monirujjaman Khan, Manjit Kaur, Atef Zaguia, Comparative Analysis of Skin Cancer (Benign vs. Malignant) Detection Using Convolutional Neural Networks, (2021), Hindawi, Article ID 5895156.

[36] Karki, Sagar and Kulkarni, Pradnya and Stranieri, Andrew, Melanoma Classification Using EfficientNets and Ensemble of Models with Different Input Resolution, 2021, Association for Computing Machinery.

[37] Milton, M.A.A., 2019. Automated skin lesion classification using an ensemble of deep neural networks in ISIC 2018: Skin lesion analysis towards melanoma detection challenge. arXiv preprint arXiv:1901.10802.

[38] Michael Cabanillas-Carbonell, Randy Verdecia-Peña, "Convolutional Neural Network Architecture for Skin Cancer Diagnosis," European Journal of Molecular & Clinical Medicine Volume 08, Issue 03, 2021.

[39] Bhuvaneshwari Shetty, Roshan Fernandes, Anisha P Rodrigues et al. , "Skin Lesion Classification of Dermoscopic Images using Machine Learning and Convolutional Neural Network," Research Square , 2022.

[40] Abhinav Sagar, Dheeba J Vellore, "Convolutional Neural Networks for Classifying Melanoma Images," biorxiv.org, 2021.

[41] S M, J., P, M., Aravindan, C. et al., "Classification of skin cancer from dermoscopic images using deep neural network architectures," Multimedia Tools Appl, 2022.

[42] Md. Maruf Hossain, Emrana Kabir Hashi, "Skin Cancer Classification Using Ensembles of Multi-Scale Efficientnets Model," Big Data, BMI, 2021.

[43] Grignaffini, F.; Barbuto, F.; Piazzo, L.; Troiano, M.; Simeoni, P.; Mangini, F.; Pellacani, G.; Cantisani, C.; Frezza, F., "Machine Learning Approaches for Skin Cancer Classification from Dermoscopic Images: A Systematic Review," Algorithms 2022, 15, 438, 2022.

[44] Bassel, A.; Abdulkareem, A.B.; Alyasseri, Z.A.A.; Sani, N.S.; Mohammed, H.J, "Automatic Malignant and Benign Skin Cancer Classification Using a Hybrid Deep Learning Approach," Diagnostics 2022, 12, 2472., 2022.

[45] A. Romero Lopez, X. Giro-i-Nieto, J. Burdick and O. Marques, "Skin lesion classification from dermoscopic images using deep learning techniques," International Conference on Biomedical Engineering (BioMed), 2017, pp. 49-54,, 2017.

[46] Weiss, K., Khoshgoftaar, T.M. & Wang, D. A survey of transfer learning. J Big Data 3, 9 (2016). https://doi.org/10.1186/s40537-016-0043-6.

[47] Neema, M., Nair, A.S., Joy, A., Menon, A.P. and Haris, A., 2020. Skin lesion/cancer detection using deep learning. International Journal of Applied Engineering Research, 15(1).

[48] Yinhao Wu et al, Skin Cancer Classification With Deep Learning: A Systematic Review, 2022.

[49] Kaiming He et al., Deep Residual Learning for Image Recognition, arXiv:1512.03385v1 [cs.CV] 10 Dec 2015.

[50] S. Tammina, "Transfer learning using vgg-16 with deep convolutional neural network for classifying images.," International Journal of Scientific and Research Publications (IJSRP), 9(10), pp.143-150., 2019.

[51] Chan HP, Hadjiiski LM, Samala RK. Computer-aided diagnosis in the era of deep learning. Med Phys. 2020 Jun;47(5):e218-e227. doi: 10.1002/mp.13764. PMID: 32418340; PMCID: PMC7293164.

[52] Sudha, V. and Ganeshbabu, T.R.,, "A convolutional neural network classifier VGG-19 architecture for lesion detection and grading in diabetic retinopathy based on deep learning," CMC-Computers Materials & Continua, 66(1), pp.827-842., 2021.

[53] Alzubaidi, L., Zhang, J., Humaidi, A.J. et al. Review of deep learning: concepts, CNN architectures, challenges, applications, future directions. J Big Data 8, 53 (2021).

[54] Passricha, V., Aggarwal, R.K. A comparative analysis of pooling strategies for convolutional neural network based Hindi ASR. J Ambient Intell Human Comput 11, 675–691 (2020). https://doi.org/10.1007/s12652-019-01325-y.

[55] Khalil, Muhammad & Tehsin, Samabia & Humayun, Mamoona & Zaman, Noor & AlZain, Mohammed. , "Multi-Scale Network for Thoracic Organs Segmentation," Computers, Materials, and Continua. 70. 3251-3265. 10.32604/cmc.2022.020561, 2022.

[56 Huang, G., Liu, Z., Van Der Maaten, L., & Weinberger, K. Q. (2017). Densely connected convolutional networks. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 4700-4708).

الملخص

تصنيف سرطان الجلد باستخدام تقنيات التعلم العميق

يعد سرطان الجلد أحد أكثر أنواع السرطانات انتشارًا في العالم، ويعتبر سرطان الجلد، على نطاق واسع، أكثر أنواع السرطانات شيوعًا وربما من أكثر أنواع السرطان المسببة للموت. على غرار جميع أنواع السرطان الأخرى، فإن الاكتشاف المبكر للمرض يزيد من احتمالية نجاح العلاج؛ فلحسن الحظ، ومن خلال تقدم الذكاء الاصطناعي في معالجة الصور وتوافر مجموعات البيانات التاريخية لسرطان الجلد، أصبح من الممكن باستخدام خوارزميات التعلم العميق تطوير نموذج تصنيفي يسهل اكتشاف المرض والحصول على نتيجة أكثر دقة لسرطان الجلد. فإن النموذج سيساعد فرق المختبرات في التحليل وسيؤثر بشكل إيجابي في طول المدة الزمنية لاكتشاف المرض بالإضافة إلى انخفاض التكلفة مقارنًة بالطريقة التقليدية. حيث أنه باستخدام هذا النظام الذي سيساهم في الاكتشاف المبكر للسرطان سيوفر الجهد والوقت والعديد من الأرواح البشرية.

في هذه الدراسة، تم تنفيذ نموذج **CNN** ومختلف النماذج القائمة على التعلم، مثل؛ **ResNet50** و**VGG16** و**VGG19** و**EfficientNetB7** و**DenseNet169** و**Xception** و**InceptionV3**. بالإضافة إلى ذلك، تم تطبيق نهج التجميع المتوسط على مجموعات مختلفة من النماذج. حيث أن مثل هذه التطبيقات تساعد على زيادة كفاءة وامكانية الكشف عن السرطان. وقد تم استخدام مجموعة بيانات تحدي تصنيف **SIIM- Melanoma ISIC** لتدريب النموذج المقترح واختباره. قمنا بتنفيذ الازدياد على مرحلتين؛ أولاً، من خلال تطبيق تقنيات متعددة على فئة الصور الخبيثة لزيادة عدد الصور في هذه الفئة وحل مشكلة عدم التوازن. ثانيًا، من خلال إجراء التقليب العشوائي لليسار ولليمين وأساليب التقليب لأعلى ولأسفل في كلا الفئتين من البيانات لإنشاء مجموعة متنوعة من الصور من الصور الموجودة لدينا.

تم تقييم أداء الطريقة المقترحة باستخدام مقاييس كمية معروفة، مثل؛ الدقة والضبط والاستدعاء ونتيجة **f1**. تتفوق نتيجة مجموعات النماذج الخمسة بكونها من أفضل النماذج، وهي **CNN**و**EfficientNetB7** و**VGG16** و**VGG19** و**DenseNet169** على جميع الطرز الأخرى، بدقة 93.79٪ وضبط 94.18٪ واسترجاع 93.79٪ ونتيجة f1 94.77٪ **في** مجموعة بيانات الاختبار غير المرئية. تشير النتائج إلى أن النموذج المقترح هو الأكثر

كفاءة وموثوقية لأتمتة تصنيف سرطان الجلد مقارنة بالأعمال السابقة، ويقصّر عملية الكشف عن المرض وهو ذو قيمة عالية ستساهم في إنقاذ حياة المرضى.

الكلمات الرئيسية: **CNN**، التعلم العميق، الورم الميلاني، سرطان الجلد، نقل التعلم.

.