ISTANBUL TECHNICAL UNIVERSITY
ELECTRICAL-ELECTRONICS FACULTY

CLASSIFICATION OF SKIN LESIONS USING ARTIFICIAL INTELLIGENCE

SENIOR DESIGN PROJECT

Dilge DAĞ
Rabia BATIR

ELECTRONICS AND COMMUNICATION
ENGINEERING DEPARTMENT

AUGUST 2024

ISTANBUL TECHNICAL UNIVERSITY
ELECTRICAL-ELECTRONICS FACULTY

CLASSIFICATION OF SKIN LESIONS USING ARTIFICIAL INTELLIGENCE

SENIOR DESIGN PROJECT

Dilge DAĞ
040190097

Rabia BATIR
040190076

ELECTRONICS AND COMMUNICATION ENGINEERING
DEPARTMENT

Project Advisor: Prof. Dr. Işın ERER

AUGUST, 2024

İSTANBUL TEKNİK ÜNİVERSİTESİ
ELEKTRİK-ELEKTRONİK FAKÜLTESİ

YAPAY ZEKA KULLANILARAK CİLT LEZYONLARININ SINIFLANDIRILMASI

LİSANS BİTİRME TASARIM PROJESİ

Dilge DAĞ
040190097

Rabia BATIR
040190076

Proje Danışmanı: Prof. Dr. Işın ERER

ELEKTRONİK VE HABERLEŞME MÜHENDİSLİĞİ BÖLÜMÜ

AUGUST, 2024

We are submitting the Senior Design Project Report entitled as "CLASSIFICATION OF SKIN LESIONS USING ARTIFICIAL INTELLIGENCE". The Senior Design Project Report has been prepared as to fulfill the relevant regulations of the Electronics and Communication Engineering Department of Istanbul Technical University. We hereby confirm that we have realized all stages of the Senior Design Project work by ourselves and we have abided by the ethical rules with respect to academic and professional integrity .

Dilge DAĞ ....
040190097

Rabia BATIR ....
040190076

FOREWORD

We are deeply grateful for the opportunity to contribute to this project and extend our heartfelt acknowledgments to all those who have supported and inspired us throughout this journey, especially Prof. Dr. Işın ERER.We also want to sincerely thank each and every one of our lecturers at Istanbul Technical University who helped us with our graduate studies.

August 2024

Dilge DAĞ
Rabia BATIR

TABLE OF CONTENTS

ABBREVIATIONS

ABCD   **:** Asymmetry, Border, Color, Diameter

AK   **:** Actinic Keratosis

ANN   **:** Artificial Neural Network

AUC   **:** Area Under the Curve

BCC   **:** Basal Cell Carcinoma

CNN   **:** Convolutional Neural Network

FKM   **:** Fuzzy K-Means

DNA   **:** Deoxyribonucleic Acid

FN   **:**False Negative

FP   **:**False Positive

FPN   **:**Feature Pyramid Networks

GLCM   **:** Gray-Level Co-occurrence Matrix

IDM   **:** Inverse Difference Moment

ISIC   **:** International Skin Imaging Collaboration

KNN   **:** K-Nearest Neighbors

LR   **:** Lojistic Regression

mAP   **:** Mean Average Precision

NV   **:** Nevus

PR   **:** Precission- Recall

PAN   **:** Path Aggregation Networks

RGB   **:** Red-Green-Blue

SCC   **:** Squamous Cell Carcinoma

SVM   **:**Support Vector Machines

TN   **:**True Negative

TP   **:**True Positive

VL   **:**Vascular Lesions

YOLO   **:**You Only Look Once

SYMBOLS

p **:** Distance parameter (represents different types of distance)

LIST OF TABLES

LIST OF FIGURES

CLASSIFICATION OF SKIN LESIONS USING ARTIFICIAL INTELLIGENCE

SUMMARY

Skin lesions are aberrant patches of skin that are not like the surrounding skin. Certain lesions are only cosmetic in nature, but others might be signs of cancer. Since skin cancer is so common, it is essential to correctly diagnose malignant tumors in order to receive appropriate treatment. There are two types of skin cancer: non-melanoma and melanoma, the latter being very hazardous. The low contrast in images between lesioned and non-lesioned areas makes diagnosing melanoma difficult, necessitating experience and causing physicians' assessments to differ. Low contrast, skin texture, and climate and regional differences are other challenges for automated skin analysis. However, developing an effective automated analytic method could help dermatologists identify skin malignancies more quickly, accelerate treatment, and improve patient outcomes.

This project aims to use a deep learning-based system as well as traditional machine learning methods in feature extraction and classification processes to accurately classify skin lesions among eight classes. Using the renowned ISIC archive in the field of skin lesion analysis, the project aims to leverage advances in machine learning to improve diagnostic capabilities. Various data augmentation techniques and undersampling are used to eliminate the prominent imbalance between the classes in data sets. Data augmentation techniques used include rotation, width and height shifts, zoom and pan. By artificially expanding the dataset through these transformations, it is aimed to reduce overfitting and increase the generalization capacity of the model. This approach has shown performance improvements, especially in underrepresented classes, and the results are reported. The project takes a look at traditional machine learning classifiers such as K-Nearest Neighbors (KNN) and Support Vector Machines (SVM) alongside advanced deep learning models. Feature extraction uses the ABCD method (Asymmetry, Border, Color, Diameter), a standard clinical approach in dermatology for lesion evaluation. Texture analysis is performed using gray-level co-occurrence matrix (GLCM), which captures statistical texture properties such as contrast, correlation, and entropy. For deep learning 3 pre-trained models: Xception, VGG16 and

ResNet50 are built upon and fine-tuned. YOLOv5 and YOLO5n are also used to tackle the task at hand by regarding it as an object detection task. By integrating these methodologies, the project seeks to develop a robust and accurate system for classifying skin lesions, contributing to advances in dermatological diagnostics and ultimately improving patient outcomes. This approach aims to make healthcare more accessible by providing a reliable diagnostic tool for both experienced dermatologists and less experienced healthcare professionals, as well as for patients in remote areas.

YAPAY ZEKA KULLANILARAK CİLT LEZYONLARININ SINIFLANDIRILMASI

ÖZET

Deri lezyonları, çevredeki deriye benzemeyen anormal deri lekeleridir. Bazı lezyonlar doğası gereği yalnızca kozmetiktir, ancak diğerleri kanser belirtileri olabilir. Cilt kanseri çok yaygın olduğundan, uygun tedaviyi alabilmek için kötü huylu tümörlerin doğru şekilde teşhis edilmesi önemlidir. İki tür cilt kanseri vardır: melanom dışı ve melanom; ikincisi çok tehlikelidir. Lezyonlu ve lezyonsuz alanlar arasındaki görüntü kontrastının düşük olması, melanom tanısını zorlaştırmakta ve hekimlerin değerlendirmelerinin farklı olmasına neden olmaktadır. Bu durum, tanının doğru bir şekilde yapılabilmesi için deneyim gerektirmektedir.Düşük kontrast, cilt dokusu, iklim ve bölgesel farklılıklar, otomatik cilt analizinin diğer zorluklarıdır. Ancak etkili bir otomatik analitik yöntemin geliştirilmesi, dermatologların cilt malignitelerini daha hızlı tanımlamasına, tedaviyi hızlandırmasına ve hasta sonuçlarını iyileştirmesine yardımcı olabilir. Bu proje, cilt lezyonlarını sekiz sınıf arasında doğru bir şekilde sınıflandırmak amacıyla derin öğrenmeye dayalı bir sistemin yanı sıra özellik çıkarma ve sınıflandırma süreçlerinde geleneksel makine öğrenme yöntemlerini kullanmayı hedeflemektedir. Cilt lezyonu analizi alanında sıkça kullanılan ISIC kütüphanesini kullanan proje, teşhis yeteneklerini geliştirmek için makine öğrenimindeki ilerlemelerden yararlanmayı amaçlamaktadır.Veri setindeki sınıflar arasındaki belirgin dengesizliği ortadan kaldırmak için çeşitli veri çoğaltma ve alt örnekleme teknikleri kullanılmaktadır. Bunlara döndürme, genişlik ve yükseklik kaydırmaları, yakınlaştırma ve kaydırma dahildir. Bu dönüşümler yoluyla veri setinin yapay olarak genişletilmesi ile aşırı uyumun azaltılması ve modelin genelleme kapasitesinin artması hedeflenmiştir. Bu yaklaşım, özellikle az temsil edilen sınıflarda performans iyileşmeleri göstermiştir ve elde edilen sonuçlar raporlanmıştır. Bu, özellikle az temsil edilen numunelerde güvenilir performans için çok önemlidir. Özellik çıkarma, lezyon değerlendirmesinde dermatolojide standart bir klinik yaklaşım olan ABCD yöntemini (Asimetri, Sınır, Renk, Çap) kullanır. Doku analizi, kontrast, korelasyon ve entropi gibi istatistiksel doku özelliklerini yakalayan gri düzeyli birlikte oluşum matrisi (GLCM) kullanılarak gerçekleştirilir. Derin öğrenme için üç önceden eğitilmiş model: Xception, VGG16 ve ResNet50 üzerine inşa edilip, ince ayar yapılmıştır. Projeye obje tanıma yaklaşımıyla sınıflandırma yapmayı denemek adına YOLOv5 ve YOLO5n modelleri kullanılacaktır. Proje, bu metodolojileri kullanarak cilt lezyonlarını sınıflandırmak, dermatolojik teşhislerdeki ilerlemelere katkıda bulunmak ve sonuçta hasta sonuçlarını iyileştirmek için sağlam ve doğru bir sistem geliştirmeyi amaçlamaktadır.Bu yaklaşım, hem deneyimli dermatologlar hem de daha az

deneyimli sağlık profesyonelleri için, ayrıca uzak bölgelerdeki hastalar için güvenilir bir teşhis aracı sunarak sağlık hizmetlerini daha erişilebilir hale getirmeyi hedefler.

1. INTRODUCTION

1.1 The Main Objectives and Purpose of Project

Cancer is a disease that occurs as a result of uncontrolled proliferation and growth of cells in a tissue of the body under the influence of various genetic and environmental factors. Almost every tissue in the body can carry abnormalities in the DNA sequence that can cause cancer. Skin cancer is also a common type of cancer. When there is a suspicion that a skin lesion can be cancerous it is important to swiftly distinguish whether the tumor is malignant or benign as an early diagnosis can prevent fatalities. Invasive melanoma, a malign type of skin cancer, causes the vast majority of deaths caused by skin cancer [1].

Dermoscopy is the non-invasive examination of skin using a device called a dermoscope that is equipped with a powerful lighting source and magnifying glasses. It is used to evaluate skin lesions, mainly pigmented ones. With the aid of a dermoscope it can become easier to diagnose and classify skin cancer. Accurate predictions increase by %49 when dermoscopic techniques are used; however, this increase only proved to be true for experienced professionals. When used by the less experienced the technique gave similar diagnostic accuracy compared to unaided examinations [2]. This can especially cause difficulties for accurate diagnosis when expertise is hard to find. Another method for diagnosis is biopsy, the surgical removal and pathological examination of the lesion. Even though it provides a definitive diagnosis it is quite time consuming thus not practical for such a common type of cancer [3]. These procedures also place a financial burden on the economy as just melanoma's cost to the U.S. healthcare system in 2011 had been more than $8 billion.

With a highly accurate and fast diagnosis technique that doesn't depend on the availability of an expert unnecessary fatalities can be prevented with early diagnosis and the financial load can be decreased. In a 2019 study, deep learning-based algorithms have proven to outperform 136 of the 157 dermatologists of various levels of experience [4]. Thus, it can be predicted that artificial intelligence trained on dermoscopic images can be the solution to early diagnosis of skin cancer and the classification of malign and benign lesions.

1

The primary objective of this project is to develop a robust and efficient machine learning model capable of accurately distinguishing between 8 classes of skin leasions. Traditional diagnostic processes often rely on time-consuming and resource-intensive pathological examinations. By leveraging advanced machine learning methods, the aim is to streamline and expedite the diagnostic process, ultimately providing a faster and potentially more accessible means of identifying malignancies in moles. Ready dataset will be used in the project, but at the end real-life data will be tested and the feasibility of the project will be observed.

The significance of this project lies not only in its potential to offer a quicker and more efficient diagnostic tool but also in reducing the reliance on invasive procedures. The model's performance will be critically assessed, and the most suitable machine learning method will be identified based on its ability to provide reliable and consistent results across different datasets.

The project intends to develop medical technology by effectively adopting this machine learning-based diagnostic approach, providing a quick and non-invasive way to differentiate between types of moles and give insight about the malignant and benign status of the mole. The ultimate objective is to increase skin cancer early detection, which will result in better patient outcomes and a more efficient healthcare system. At the beginning of the project, it was aimed to achieve similar results as previous projects. Approximately 80% accuracy was achieved in the applied methods.

1.2 Literature Review

In recent years, significant advancements have been made in the field of skin disease classification, driven by the increasing availability of large dermatoscopic image datasets like HAM10000. Researchers have extensively explored the application of machine learning and deep learning models to improve diagnostic accuracy and aid in early detection of skin cancer. Among these models, Convolutional Neural Networks (CNNs) have demonstrated superior performance compared to traditional ML algorithms, owing to their ability to capture intricate patterns in dermatoscopic images. One such study by Shetty et al. [5] analyzed the classification accuracy of various Machine Learning algorithms and Convolutional Neural Network models. Their findings indicated that CNN models outperformed traditional machine learning algorithms, achieving the highest accuracy of 95.18%. This superior performance of CNN models highlights their ability to capture and learn complex patterns within

2

dermatoscopic images. A study by Adebiyi et al. [6] used the ALBEF (Align before Fuse) model for multimodal deep learning and compared it to image-only models like Inception-v3, DenseNet121, and ResNet50. The ALBEF model outperformed them with an accuracy of 94.11% and an AUROC score of 0.9426 on the HAM10000 dataset, highlighting the potential of multimodal deep learning in enhancing skin cancer detection accuracy. Deif et al. [7] proposed four different CNN architectures for the classification of skin lesions: VGG16, VGG19, MobileNet, and InceptionV3. They used the HAM10000 dataset for training, verification, and testing, reporting that the accuracy rates of the VGG16, VGG19, MobileNet, and InceptionV3 architectures were 87.42%, 85.02%, 88.22%, and 89.81%, respectively. Taşar [8] evaluated pre-trained CNN models (VGG16, ResNet50, DenseNet121, MobileNet, Xception) using the HAM10000 dataset. The modified DenseNet121 achieved the highest accuracy of 94.29%, followed by VGG16 (93.28%), ResNet50 (87.10%), MobileNet (83.10%), and Xception (80.05%). Esteva et al. [9] achieved dermatologist-level classification performance using a single CNN model trained on a large dataset of dermoscopic images. Building on such successes, researchers have investigated ensemble methods to leverage the complementary strengths of different CNN models. Shahin et al. [10] proposed an automated framework employing an ensemble approach by combining ResNet-50 and Inception V3 architectures to classify seven different skin lesion types. This approach achieved a validation accuracy of up to 0.899. Kassem et al. [11] reported performance indicators such as accuracy, sensitivity, specificity, and precision, with respective values of 94.92%, 79.8%, 97%, and 80.36%. Their suggested solution performed better when there were fewer photos in each class overall, addressing the issue of class imbalance. Moldovan et al. [12] leveraged a method based on transfer learning, achieving an accuracy of 85% on the HAM10000 dataset. Thwin and Park [13] used an ensemble of VGG16, Inception-V3, and ResNet-50 for skin lesion classification on the ISIC 2018 dataset. Their weighted averaging method achieved 91% accuracy on the original dataset and 97% on the oversampled dataset. When evaluated on the HAM10000 dataset, they achieved 90% accuracy on the original dataset and 96% on the oversampled dataset, demonstrating the effectiveness of ensemble models in improving classification performance and addressing dataset imbalance. In Tschandl et al.'s [14] study, 511 participants from 63 countries took part in lesion prediction trials, including 283 board-certified dermatologists, 118 dermatology residents, and 83 practicing physicians. Out of 90 studies on the accuracy of automated diagnostic systems for melanoma, 57 provided sufficient data for quantitative analysis, with nine directly comparing machine learning algorithms to human experts. On average, algorithms achieved 2.01 more accurate diagnoses than human

readers. A large dermatoscopic dataset of pictures of benign lesions and melanoma was used by Haenssle et al. [15] to train their InceptionV4 deep learning model. They evaluated the model's performance against that of 58 dermatologists using a test set of 100 cases, comprising 25 cases of melanoma and 75 cases of benign lesions. Dermatologists attained an average sensitivity of 86.6% and specificity of 71.3%, in contrast to the deep learning model's 95% sensitivity and 63.8% specificity. Four modified pre-trained models—EfficientNetB0, VGG19, ResNet50V2, and MobileNet—were tested in a Ranjan et al. [16] study to see how well they classified skin lesions. Comparable metric scores were displayed by ResNet50V2 and MobileNet, which outperformed EfficientNetB0 and VGG19 in the results. In the Hossain et al. [17] study, different pre-trained deep neural networks were combined using the Max Voting Ensemble Technique to classify skin cancer using the HAM10000 and ISIC 2018 datasets. With an accuracy of 93.18% and an AUC score of 0.9320, this approach outperformed individual models in classification performance. In Elyasi and Moghadam's study [18], skin cancer images were classified using two different datasets using the Kepler Mapper algorithm and the Xception neural network. The study demonstrated the power of TDA algorithms (Mapper and persistent homology) and neural networks using the classification performed by the Xception neural network and the analysis of network layers. While traditional deep learning models only capture Euclidean neighborhood relationships, this study extracts both Euclidean and topological features using TDA methods. Alam et al [19] achieved an average accuracy of 0.9103 on the classification task trained on HAM10000 using Efficient-Net B4 instead of U-Net encoder in the segmentation network. The paper by Al-Tuwaijari and Mohammed [20] reviews machine learning methods for skin condition classification, noting that while artificial neural networks, support vector machines, and Naive Bayes show variable performance depending on dataset size and feature selection, convolutional neural networks achieve superior accuracy but are time-consuming. They also highlight that low resolution and noise often hinder image-based classification, while texture feature-based methods generally offer better diagnostic accuracy. Also, The HAM10000 dataset was used in the study by Elshahawy et al. [21] to develop a new melanoma detection model that included YOLOv5 with ResNet50. Values of 99.0%, 98.6%, 98.8%, 99.5%, 98.3%, and 98.7% were achieved for the performance criteria of sensitivity, recall, DSC, accuracy, and MAP (0.0-0.5). Lastly, in the study conducted by Xu [22], they investigated the effect of dataset size on the image classification performance of eight networks AlexNet, ResNet18, ResNet34, etc. We trained these classifiers using different ratios of samples from the HAM10000 dataset. The models yielded the highest average accuracy of 85.09% when the

dataset size was 5%. They were trained on datasets ranging from 1% to 100%. Notably, ResNet101 and ResNet152, trained on 100% and 5% of the datasets, respectively, achieved the maximum accuracy of 90.07% and 90.02%.

1.3   Our Contributions

The project on skin lesion classification using artificial intelligence tackled the significant issue of data imbalance by employing a range of data augmentation techniques. These techniques included rotation, shifting, zooming, and panning, all of which were strategically applied to the dataset. By augmenting the data in these ways, the project aimed to improve the model's ability to generalize from training data to unseen data, effectively reducing the risk of overfitting. This approach enhanced the model's robustness and reliability in real-world applications. In terms of feature extraction, the project integrated both the ABCD rule and advanced texture analysis methods. The ABCD rule, which assesses asymmetry, border irregularity, color variation, and diameter, provided a structured approach to evaluating the primary characteristics of skin lesions. Complementing this, texture analysis techniques such as the Gray-Level Co-occurrence Matrix (GLCM) were employed to capture and quantify texture features within the lesions. GLCM helps in understanding the spatial relationship between pixels, which is crucial for distinguishing between different types of lesions based on texture patterns.

For the classification tasks, several pre-trained deep learning models were utilized and fine-tuned through transfer learning. These models included Xception, VGG16, and ResNet50, each of which has demonstrated exceptional performance in various image classification tasks. Transfer learning allowed the project to leverage the rich feature representations learned by these models on large-scale datasets, adapting them to the specific nuances of skin lesion images. This approach significantly boosted the accuracy and effectiveness of the classification system. Additionally, the project incorporated the YOLO model for real-time detection and classification of skin lesions. YOLO's capability to perform object detection in real-time made it particularly suited for clinical environments, where timely and accurate identification of lesions is crucial. The implementation of YOLO enhanced the system's practical applicability, providing healthcare professionals with a powerful tool for the early detection and assessment of skin lesions.

2. MATERIALS AND METHODS

2.1 Dataset

The International Skin Imaging Collaboration (ISIC) dataset collection is a comprehensive repository of high-quality dermoscopic images curated for the purpose of advancing research in the field of dermatology. ISIC archive **is** known for their diverse and extensive collection of images, which cover a wide range of skin conditions, including both benign and malignant lesions. These datasets are annotated with detailed metadata, such as lesion type, diagnosis, and additional clinical information, provided by expert dermatologists.

The dataset utilized in this study comprises dermoscopic images sourced from the ISIC Archive. It consists of 11,720 images, categorized into eight distinct classes of skin lesions. The classes included are Nevus, Melanoma, Pigmented Benign Keratosis, Dermatofibroma, Squamous Cell Carcinoma, Basal Cell Carcinoma, Vascular Lesion, and Actinic Keratosis. These images reflect a diverse range of skin conditions, encompassing both benign and malignant cases. This dataset is characterized by an imbalance in class representation, with the majority class, Nevus, comprising 66% of the total images. Table 2.1 indicates the number of images in the dataset as well as the class distribution.

Table 2.1 :   Image and class distribution in dataset.

	Original
Nevus	7737
Melanoma	1305
Pigmented Benign Keratosis	1338
Dermatofibroma	160
Squamous Cell Carcinoma	229
Basal Cell Carcioma	622
Vascular Lesion	180
Actinic Keratosis	149

The HAM10000 dataset, short for "Human Against Machine with 10000 training images," is a well-known collection of dermoscopic images of pigmented lesions [14]. It includes 10,015

images with majority class nevus making up %67 of the dataset, categorized into seven classes: Melanocytic Nevi, Melanoma, Benign Keratosis-like Lesions, Basal Cell Carcinoma, Actinic Keratosis, Vascular Lesions, and Dermatofibroma. The HAM10000 dataset is publicly avaliable through the ISIC library.

Our dataset, derived entirely from the ISIC library, includes similar skin condition categories, with the addition of another class, the overlap in categories and the high-quality, expert-curated nature of both datasets provide a solid foundation for drawing meaningful comparisons. This allows us to evaluate the performance of machine learning and deep learning models across different but related datasets. The comparison of the attained results are valid and insightful therefore our reults will be compared with those achieved by training with HAM10000 dataset.

2.1.1   Actinic keratosis (AK)

Actinic keratosis is a precancerous skin condition caused by prolonged exposure to ultraviolet (UV) radiation. It appears as rough, scaly patches on sun-exposed skin. If untreated it has the potencial to progress to squamous cell carcinoma [23].

2.1.2   Basal cell carcinoma (BCC)

Basal cell carcinoma (BCC) is a type of skin cancer that is generally considered "benign" because it grows slowly and has a low tendency to metastasize (spread to other parts of the body). However, if left untreated, BCC can damage surrounding tissue and cause severe tissue loss. Therefore, it is important to detect and treat it early [23].

2.1.3   Dermatofibrama (DF)

Dermatofibroma is abenign, fibrous skin tumor that is firm and often feels like a small lump under the skin. It can be brownish in color. It is generally harmless and does not require treatment unless it causes discomfort or cosmetic concerns [23].

2.1.4 Melanoma (ML)

Melanoma is a serious form of skin cancer that develops from melanocytes. It often appears as a new or changing mole or spot on the skin. It can be life-threatening if not detected and treated early. It has a higher potential to spread to other parts of the body [23].

2.1.5 Nevus(NV)

Commonly known as a mole, a nevus is a benign growth of melanocytes ,pigment-producing cells, in the skin. Nevus can be flat or raised, and their color can range from pink to brown or black. They can be present from birth or develop over time. Nevuses are usually harmless, but changes in a nevus can indicate melanoma, which is a type of skin cancer [23].

2.1.6 Pigmented benign keratosis (BK)

Benign skin lesions that are typically brown or black, raised, wart-like growths on the skin caused by an overgrowth of skin cells. As indicated by the name they are not cancerous. It can be mistaken for other skin conditions but usually requires no treatment other than for cosmetic reasons or if they become irritated [23].

2.1.7 Squamous cell carcinoma (SCC)

Squamous cells make up the outer layer of skin.Squamous cell carcinoma is a type of skin cancer that originates in the squamous cells. It often looks like a scaly, red patch or a sore that doesn't heal. Without early treatment it can be aggressive and without treatment can spread to other parts of the body[23].

2.1.8 Vascular lesion (VL)

Vascular lesions can be described as abnormal growths or malformations of blood vessels, which can appear as red or purple marks on the skin. Some are due to genetic conditions, while others may develop due to aging or sun exposure.Even though they are often benign some can cause complications or cosmetic concerns, and might require treatment depending

on their location and severity [23].Figure 2.1 illustrates samplesfrom all the classes taken from the dataset. It is evident after taking a look at the data that there is no easy distinction between the classes.
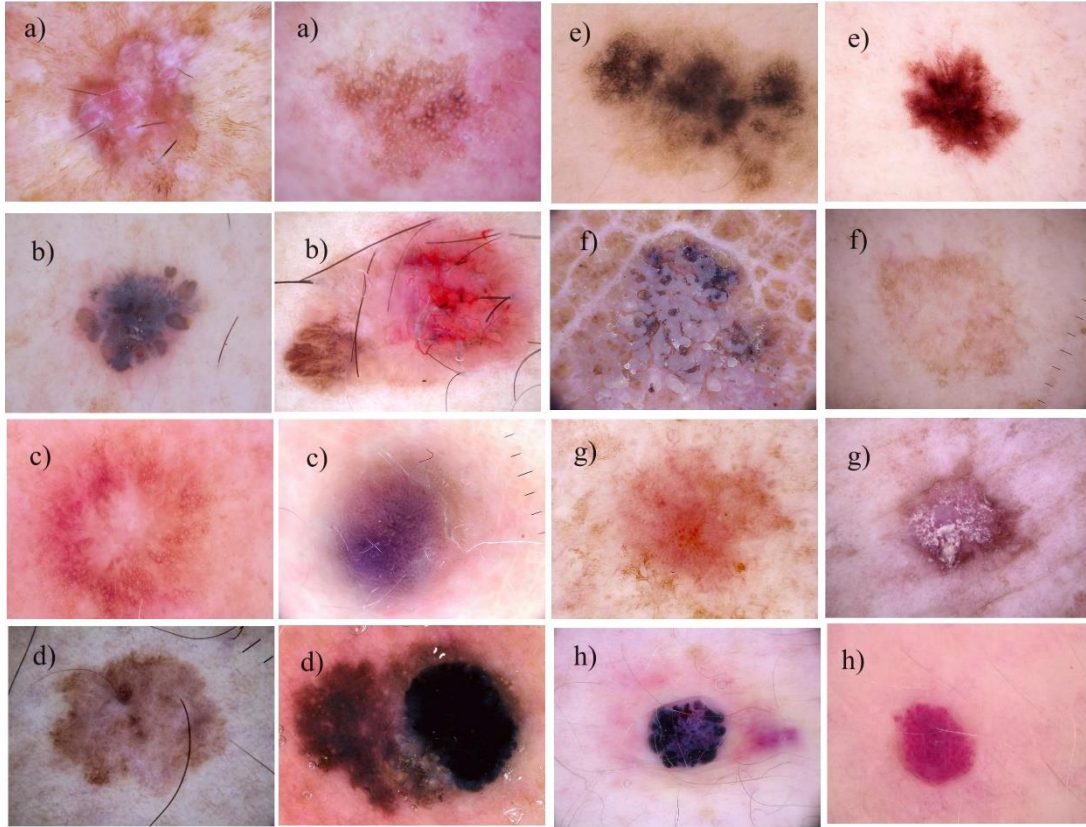
Figure 2.1 :   Sample images from the dataset a)AK, b) BCC, c)DF, d)ML, e)NV, f)BK, g)SCC, h)VL.

2.2 Data Preprocessing

As a first step in preprocessing the data, normalization is done. Normalization ensures that all input data is on a consistent scale, facilitating faster and more effective learning by the model. This step also improves convergence during training, leading to a more stable and efficient learning process. By providing normalized input data, the model can detect patterns and features more accurately, ultimately enhancing its performance. Without normalization, the model may perform poorly due to inconsistencies in pixel value scales across different images. To address this, pixel values were normalized to a range of [0, 1] by dividing each pixel value by 255.

The dataset has inconsistencies when it comes to the image resolutions, this can have a substantial impact on the model output. Consistent image sizes help stabilize the training process, preventing issues like gradient explosion, which can occur if the input data has widely varying characteristics. We resized all images to a uniform resolution of 224×224 pixels to ensure that the input dimensions are consistent across all samples. Additionally this leads to better feature extraction, improves the stability and efficiency of the training process.

2.3 Data Augmentation

Data augmentation techniques are used to increase training data without actually collecting new data. It involves applying various transformations to existing data points, such as rotations, flips, zooms, shifts, and changes in brightness or contrast. This process helps create multiple versions of the same data, enhancing the model's ability to generalize and improving its robustness to variations in input data. All the data augmentation techniques applied are given in Table 2.2. By artificially enlarging the dataset, data augmentation helps mitigate overfitting and improves the performance of machine learning models.Since data augmentation can improve the model's performance, it is particularly advantageous when working with little datasets.

Table 2.2 :   Data augmentation strategies.

Type	Value	Description
Rotation Range	10	Rotational degree between 0 and value
Width Shift	0.1	Shift the image on X-Scale
Height Shift	0.1	Shift the image on Y-Scale
Zoom Range	0.1	Scale upto which the image is zoomed
Horizontal Flip	True	Horizontal image flipping
Vertical Flip	True	Flipping the image vertically

Rotation Range: Images are rotated arbitrarily within a predetermined degree range.
Width Shift : Randomly moves pictures left or right along their horizontal axis by a predetermined   percentage   of   their   width
Height Shift: Images are shifted vertically (up or down) by a predetermined percentage of

their entire height, just like with width shift. **Zoom Range:** Applies a random range of zoom to photos**.** **Horizontal Flip:** This technique randomly rotates pictures from left to right. **Vertical Flip:** This technique randomly rotates pictures from top to bottom.

As a way to tackle class imbalance Alsaidi et al. [24]'s approach will be implemented and compared to the results of previous results with only flipping. According to the research about imbalanced dermoscopic image datasets, when trained on a dataset balanced through augmentation, model hyperparameters and accuracies compared to models trained on imbalanced dataset showed a clear and significant increase in accuracies. On their approach the augmented data almost perfectly balances the overall data for model training, meaning all the classes roughly had the same number of samples. In order to use this method for class imbalancewe will be applying all the augmentation methods on the Table 2.2, however in order to limit the size of the new dataset we will be randomly undersampling the majority class, nevus. With this method we are left with 4500 images belonging to nevus class, the remaining classes will be randomly augmented until the number of images in the class match that of nevus. We will be referring to this new balanced dataset asthe augmented dataset, the non-augmented dataset will be referred to as the original dataset.Class distribution and number of images after undersampling and augmention is shown in Table 2.3. The class imbalance is non-existant in the augmented dataset. For the YOLO dataset 1000 images from each class is labeled and used. Use of original data is prioritized and only when the original number of images for the class failed to reach 1000 augmented images are added to complete the subset to 1000 images. The dataset that will be used for YOLO required manually labellingeach image with bounding boxes.Since both the original dataset and abalanced dataset will be used to train YOLO we limited the number of images per class due the limited availability of resources and time requird for labeling such a large datatset. All the datasets' class distributions are shown in Table 2.3.

Table 2.3 :   Image and class distribution in datasets.

HAM10000	Original	Augmented and balanced	YOLO
Nevus	7737	4500	1000
Melanoma	1305	4500	1000
Pigmented Benign Keratosis	1338	4500	1000

HAM10000	Original	Augmented and balanced	YOLO
Dermatofibroma	160	4500	1000
Squamous Cell Carcinoma	229	4500	1000
Basal Cell Carcioma	622	4500	1000
Vascular Lesion	180	4500	1000
Actinic Keratosis	149	4500	1000

Both the original and augmented datasets are divided into training, test and validation subsets. The class distributions of subsets match those of the main datasets. The split is 80% training, 20% test for machine learning and 60% training, 20% test, 20% validation for deep learning and YOLO. The total number of images in each subset is shown in the Table 2.4.

Table 2.4 :   Class distribution of each dataset. Machine learning (ML), deep learning (DL), YOLO

Method	Dataset	Training	Test	Validation
ML	Original	9376	2344	-
ML	Augmented	28800	7200	-
DL	Original	7032	2344	2344
DL	Augmented	21600	7200	7200
YOLO	Original	7032	2344	2344
YOLO	Augmented	21600	7200	7200

2.4 Assessment of Success

The results are assesed based of certain performance metrics that are calculted to reflect how well the model does the classification and gives insight on how the model can be improved.

2.4.1   Traditional machine learning and deep learning

Accuracy: defines the recognition rate of the system. It is the ratio of correctly predicted instances to the total instances. Accuracy is a straightforward metric but can be misleading in cases of imbalanced datasets. For example, if 95% of your data belongs to one class, a model that always predicts this class will have 95% accuracy but will be useless in identifying the minority class.

Sensitivity: also known as recall or true positive rate, measures the proportion of actual positives that are correctly identified by the model. It indicates how well the model can identify positive cases. High recall is crucial in situations where it's important not to miss

any positive cases, such as in medical diagnostics. Most real positive cases are found when a model has a strong recall. It might, nevertheless, also result in more false positives.

Precision: is the ratio of correctly predicted positive observations to the total predicted positives. It reflects how many of the positive predictions were actually correct. Precision is important when the cost of false positives is high. High precision means that when the model predicts a positive case, it is likely to be correct.

The F1 score: is the harmonic mean of precision and recall, providing a balance between the two. It gives a single metric that balances the trade-offs between precision and recall. The F1 score is particularly useful when you need to balance precision and recall, or when you have an imbalanced dataset. It provides a more comprehensive view of the model's performance compared to accuracy, which might not differentiate well in such scenarios.

A true negative (TN) is achieved when the model correctly predicts the absence of the positive class. Conversely, a false positive (FP) occurs when the model incorrectly predicts the presence of the positive class. Additionally, a false negative (FN) occurs when the model inaccurately predicts the absence of the positive class. When the model correctly predicts the existence of the positive class, this is known as a true positive (TP). Stated otherwise, the model accurately detects occurrences belonging to the positive class. The formulas for the metrics mentioned above can se seen in Table 2.5. Throughout the project, the different machine learning and deep learning methods mentioned above will be performed on the dataset. Finally, augmented data will be tested in the project and performance metrics will be reported.

Table 2.5 :   Performance metrics' formulas.

Success Criterias	Formula
Accuracy	(TP+TN) /(TP+TN+FP+FN)
Sensitivity	TP/(TP+FN)
Presicion	TP/(TP+FP)
F1-score	TP/ (TP + (FP+FN) /2)

2.4.2 YOLO

Intersection oveR Union (IoU):is a metric used to measure the accuracy of an object detector on a particular dataset. It calculates the overlap between the predicted bounding box and the ground truth bounding box.A higher IoU indicates a better prediction. IoU is crucial because it helps to determine if a predicted bounding box is close enough to the ground truth bounding box.Low IoU might indicate poor localization, even if the objects are detected.

Precision-Recall Curve (PR): This curve plots precision (y-axis) against recall (x-axis) at different confidence thresholds. It helps to visualize the trade-off between precision and recall. A higher area under the curve indicates a better model performance.

Mean Average Precision (mAP): is a common evaluation metric used in object detection that summarizes the precision-recall curve for all classes.mAP is then computed as the mean of the APs for all classes. mAP provides an overall measure of model accuracy, taking into account both precision and recall across different classes. A higher mAP indicates better performance.

Confidence Score: Each detection in YOLO comes with a confidence score that indicates the likelihood that the detected object actually exists and the accuracy of the bounding box. The model's confidence in its predictions. High-confidence detections are more likely to be correct. However, a high confidence score does not guarantee a correct detection if the IoU is low.

Table 2.6 :   Performance metrics' formulas.

Success Criterias	Formula
IoU	(Area of Overlap) / (Area of Union)
Precision	TP/ (TP + FP)
Recall	TP/(TP+FN)

2.5 Machine Learning

Algorithms called machine learning classifiers are made to divide incoming data into various groups or classifications. Classifiers in supervised machine learning use labeled datasets for training in order to discover patterns and connections between labels and features. Non-linear classifiers, like random forest, catch complicated patterns, whereas linear classifiers, like logistic regression, presume linear correlations. A non-parametric, lazy learning technique called K-Nearest Neighbors (KNN) classifies data points according to the majority class of their k-nearest neighbors in feature space. Support vector machine(SVM) can be used as linear or non-linear depending on the type of kernel used. SVM works by finding the hyperplane that best separates the classes in the feature space.

2.5.1   Classifiers

Classifiers are algorithms designed to categorize input data into different classes or categories. In supervised machine learning, they rely on labelled datasets during training to learn patterns and relationships between features and labels. During training, classifiers adjust their internal parameters to minimize the difference between predicted and true labels, in other words to reduce error. The process involves using features, properties of the input data, and assigning labels or classes. Linear classifiers, such as logistic regression, assume linear relationships, while non-linear classifiers, such as decision trees, capture complex patterns. K-Nearest Neighbors (KNN), on the other hand, classifies data points based on the majority class among their k-nearest neighbours in feature space. Because it is a non-parametric, lazy learning algorithm, it does not make any assumptions about the distribution of underlying data while it is being trained. After training, classifiers are tested on a separate dataset to assess performance using metrics like accuracy and precision. Once satisfactory, they can be deployed for predictions on new data. Challenges include overfitting and underfitting, mitigated by proper feature selection, model tuning, and diverse training data. AI classifiers find applications in image recognition, speech processing, fraud detection, and medical diagnosis, automating decision-making based on learned patterns from previous data [25].

2.5.1.1 K-nearest neighbours

K-Nearest Neighbors (KNN) stands out as a widely used and simple supervised learning technique that can be applied to both classification and regression challenges. Working as a non-parametric lazy learning method, KNN stands out by not explicitly learning the parameters during training; instead, it relies on the entire training dataset for decision making. KNN, often classified as an example-based learning algorithm, derives its predictions by evaluating the similarity between examples. In fact, when a new sample arrives, the KNN algorithm first starts by calculating the distance of the sample to all samples in the training set to determine which class the sample belongs to. Then, the k closest examples in the training set are selected for the new example and the new example is assigned to the class with its k closest neighbors. This unique approach makes KNN a versatile and intuitive algorithm for a variety of machine learning tasks. The effectiveness of the k-Nearest Neighbors (kNN) method is largely affected by the choice of the k parameter. It is very important to choose an appropriate value for k as it significantly affects the classification results. One approach to address this sensitivity is to experiment with various values of k by running the algorithm multiple times. By evaluating the performance of the algorithm for different k values, the k that gives the best classification results can be selected [25]. This iterative process helps reduce the bias introduced by the selection of k in the kNN method.

Different distance functions such as Minkowski Distance, Manhattan Distance, Euclidean Distance and Cosine Distance can be preferred in determining the distance between neighbors. A generalization of several distance metrics, such as the Manhattan distance and the Euclidean distance, is the Minkowski distance. It includes a parameter p that defines the kind of distance metric and is defined for real-valued vector spaces. In a normed vector space, the formula for the Minkowski distance between two points, x and y, is as follows:

$$D(X,Y) = \sum_{i=1}^{n} |x_, - y_i|^p {}^{\frac{1}{p}}$$

In the Minkowski distance formula, the kind of distance metric is determined by the parameter p. This is equivalent to the Manhattan distance when p = 1 and the Euclidean distance when p = 2. It symbolizes a more inclusive version of distance for other values of p.

16

The distance metric needs to meet the following requirements:

a. Non-negativity: The distance between any two points is always non-negative.

$$d(x,y) \geq 0$$

b. Identity: If and only if two points are identical, then there is no distance between them.

$$d(x,y) = 0 \; if \, and \, only \, if \; x = y$$

c. Symmetry: The distance from point x to point y is the same as the distance from point y to point x.

$$d(x,y) = d(y,x)$$

d. Triangle Inequality: The sum of the distances between two points x and y, and between y and z, is always greater than or equal to the distance between x and z.

$$d(x,y) + d(y,z) \geq d(x,z)$$

2.5.1.2 Support vector machine
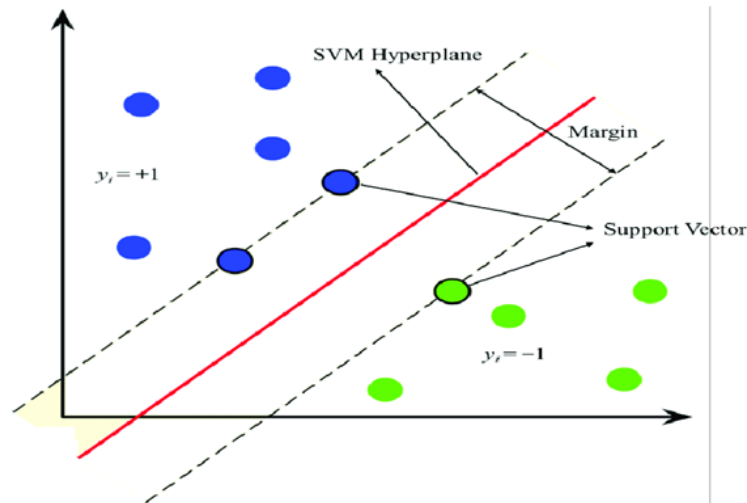
Figure 2.2 : Support vector machine model [26].

SVM is a machine learning approach typically used for classification, aiming to find the optimal hyperplane that separates different classes of samples. Figure 2.2 illustrates an SVM classifier that separates two classes using a line. In this project, SVM will be applied for

multi-class classification.Given n training data sets (x1,y1) to (xn,yn), where x represents the features and y the class labels, SVM seeks to find hyperplanes that separate the different classes. For multi-class classification, this involves constructing multiple binary classifiers, usually using "one-vs-rest" or "one-vs-one" approaches. Any hyperplane is defined as the collection of points x that satisfy the following conditions:

$$w * x - b = 0$$

Here w acts as the hyperplane's normal vector. Hard margin and soft margin are the two sorts of margins that are possible. Hard margin is utilized if the training data can be entirely and linearly separated without any errors (perfect seperation) while soft margins allow some errors. The goal is to maximize the margin between classes.If there is a mistake, the margin either shrinks or the exact margin is not successful

$$y = argmax_j\left(w_j{}^T * x + b_j\right)$$

Each class j is assigned based on which hyperplane gives the highest value. In this project, the SVM model will classify images into multiple classes, assigning them based on this multi-class decision rule[26].

2.5.1.3 Logistic regression
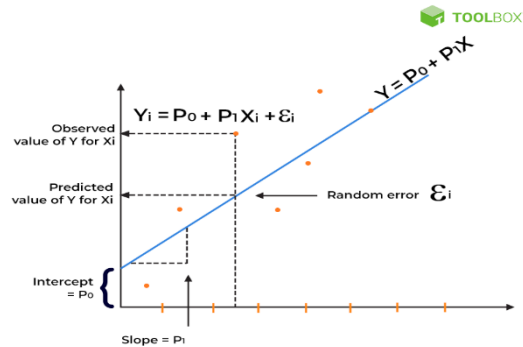
Figure 2.3 :   Linear regression model[25].

In Figure 2.3, a linear regression model is presented, and the straight line depicted here is the regression line described as $Y = P_0 + P_1 * x$. Linear regression models will be used in this project because they are fast in training and prediction processes. Depending on the data set

size, the model can often be trained in a short time and quickly used to predict new observations.

The equation of a linear regression line is given by

$$Y = a + b * X$$

where X represents the explanatory variable, and Y is the dependent variable. In this equation, b is the slope of the line, and a is the intercept, representing the value of Y when X is equal to 0. Linear regression attempts to find a line that best fits the points in the data set. This line is usually determined by the method of least squares, that is, it is true that the total squared errors are minimum.

2.5.1.4 Random forest

Random forest is a popular ensemble learning method that excels in both classification and regression tasks. It is recognized as a non-parametric, bagging-based algorithm that builds multiple decision trees during the training phase and merges their outputs to improve the overall prediction accuracy. Unlike parametric models, Decision Forests do not make assumptions about the underlying data distribution, allowing them to effectively capture complex patterns and relationships within the dataset[27].
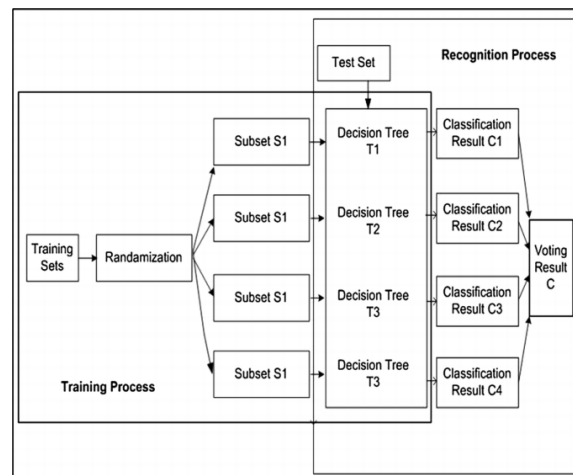
Figure 2.4 :   Decision tree model [27].

It builds multiple decision trees during training and combines their outputs to enhance accuracy. Each tree is trained on a random subset of the data and features, which introduces diversity and reduces overfitting. For classification, the final prediction is determined by majority vote among the trees; for regression, the average prediction is used. A decision tree models can be seen in Figure 2.4.

While powerful, Decision Forests can overfit if trees are too deep and may lack interpretability due to the ensemble nature. Despite these challenges, they are a robust and versatile choice for various machine learning tasks.

2.5.2   Segmentation

By using characteristics like color, intensity, or texture, a picture can be segmented into discrete segments or areas. A major component of the image, like an item or the background, is often represented by one segment. By focusing on evaluating and interpreting these significant portions of the image, machine learning algorithms are able to improve their overall comprehension and analysis of the picture.

2.5.2.1 OTSU binarization

Otsu binarization is a thresholding technique used in image processing to convert a grayscale image into a binary image [28]. It automatically selects the optimal threshold value that minimizes the intra-class variance of the black and white pixels. This method is effective for images with a bimodal histogram, where pixel intensities can be clearly separated into two distinct groups. The process involves computing the histogram of the grayscale image, finding the threshold value that minimizes the weighted sum of variances of the two classes (foreground and background), and then applying this threshold to convert the grayscale image into a binary image. Pixels above the threshold are set to one value (e.g., white), and pixels below the threshold are set to another value (e.g., black). Otsu's method is widely used for its simplicity and effectiveness, especially in images with clear foreground and background separation.

2.5.2.2 Morphological operations

Morphological operations are a set of image processing techniques used to analyze and manipulate the structure of objects within an image. They are primarily applied to binary images but can also be used on grayscale images. These operations rely on the shape and structure of objects and typically involve the use of a small shape or template known as a structuring element. Most common operations include: erosion, dilation, opening, closing and morphological gradient.

2.5.3   ABCD feature extraction

The ABCD Rule is a dermatological tool used to evaluate moles and skin lesions for signs of melanoma. It stands for A-asymmetry, B-border, C-color, and D-diameter. As mentioned in the work of Johr [29],following pattern analysis, the second algorithm to be devised was the ABCD rule of dermatoscopy, which was an initial attempt to streamline the procedure.Another tool used is the seven-point checklist, which has irregular shape, irregular color, and large diameter in common with the ABCD rule. Moreover, the seven-point checklist provides a simplification of standard pattern analysis. These clinical features are essential for identifying high-risk lesions and differentiating benign from malignant conditions. By integrating these characteristics with advanced computational techniques, the predictive accuracy of melanoma detection models can be significantly improved. As a combination of the two tools and the metadata available, our chosen features for the machine learning model will be extracted from asymmetry, border, color, and texture analysis. In conclusion, by combining traditional clinical techniques like the ABCD rule and the seven-point checklist with innovative approaches such as image processing and machine learning, the algorithm aims to provide a robust and reliable tool for melanoma screening and diagnosis. The iterative testing and refinement of different approaches will guide the development towards the most satisfactory results, ultimately contributing to more effective melanoma detection and patient care.
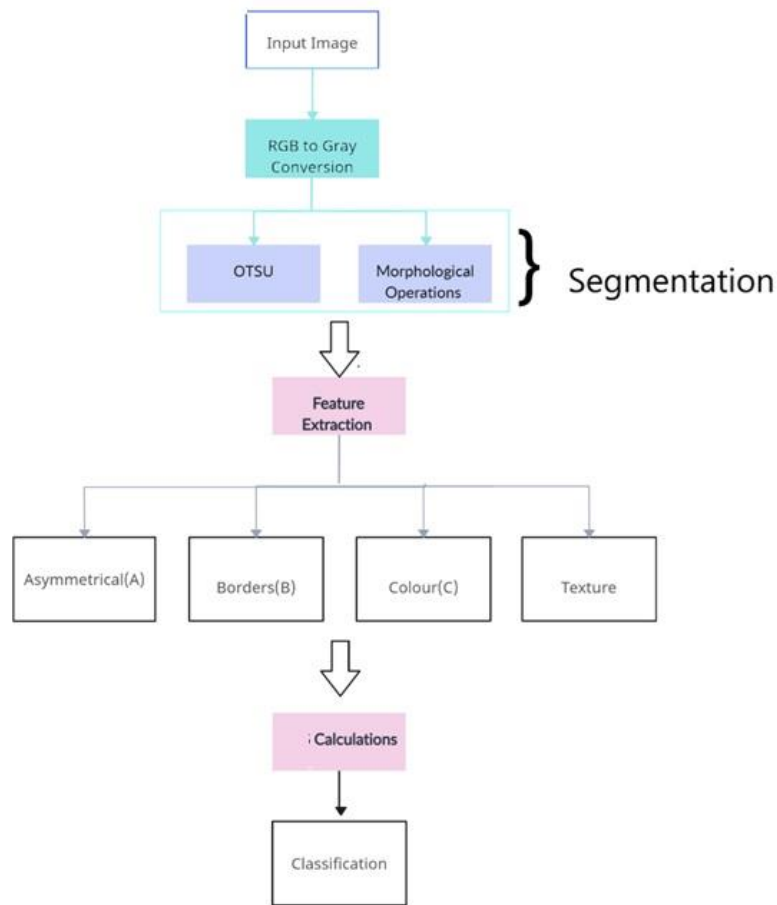
Figure 2.5 :   Machine learning strategy flowchart.

2.5.3.1 Asymmetry

Look for irregularities in shape. Asymmetrical lesions may raise concerns. The asymmetry score calculation involves comparing the original image to its horizontally flipped version. This process quantifies the asymmetry of each image by comparing its pixel distribution with that of its horizontally flipped version, providing a numerical measure of asymmetry for further analysis and classification

2.5.3.2 Border

Examine the edges of the lesion. Uneven, notched, or blurred borders can be indicative of potential issues. The method employs contour analysis to compute the irregularity of lesion

borders. It does so by first detecting contours in binary images obtained through segmentation. The largest contour, representing the lesion boundary, is identified and used to calculate its perimeter and area. These metrics are then used to compute circularity, a measure of how closely the contour resembles a perfect circle. The inverse of circularity yields the border irregularity score, providing a quantitative assessment of border irregularities.

2.5.3.3 Color

Evaluate the color distribution within the lesion. Uneven color or multiple colors may be a sign of concern. Three 8-bit layers make up an RGB image: red, green, and blue. To ascertain the RGB ratios of interest for a specific segmented image, the individual RGB colors can be retrieved and evaluated. The various colors are determined by the RGB mix ratio. The mean and standard deviation between the channels of two different color spaces, as well as other parameters like color asymmetry, histogram distance, and centroidal distance, were computed to quantify the colors in a lesion. As a measure of color asymmetry, color skewness and color kurtosis scores are calculated. The histogram distances are calculated using euclidean distance whereas centroid distances are calculated using absolute distance. The color spaces considered were RGB and HSV.

2.5.3.4 Texture

The spatial arrangement of intensity and color in an image defines its texture, and various methods can be employed to characterize it. Some approaches rely on pixel statistics, while others leverage gradient histograms for texture analysis.The utilization of gradient histograms has proven highly effective in addressing challenging problems, showcasing excellent results across various applications. In this study, we will employ histogram-based techniques to capture and describe the distinctive features of image texture. Haralick texture features will be calculated from the grayscale version of the image [30].This involves computing the gray-level co-occurrence matrix (GLCM) and extracting statistical measures [30]. We picked the Mahotas library for texture analysis which computes 13 Haralick texture features by default. The features calculated are as the following:

- Angular Second Moment (ASM): Also known as energy, it measures the uniformity of the image texture. Higher values indicate more homogeneity.

- Contrast: Measures the local variations in pixel intensities. High contrast values indicate large intensity differences between neighboring pixels.

- Correlation: Measures the linear dependency between pixel intensities at different image locations. It indicates the level of correlation between pixel pairs.

- Variance: Measures the variance of GLCM elements.

- Inverse Difference Moment (IDM): Measures the local homogeneity of the image texture.

- Sum Average: Represents the mean value of GLCM elements.

- Sum Variance: Measures the variance of the sum values of GLCM elements.

- Sum Entropy: Represents the entropy of the sum values of GLCM elements.

- Entropy: Measures the randomness or disorder of pixel intensity distribution in the image.

- Difference Variance: Measures the variance of the differences between neighboring pixel intensities.

- Difference Entropy: Represents the entropy of the differences between neighboring pixel intensities.

- Information Measure of Correlation 1 (IMC1): Measures the correlation between pixel pairs.

- Information Measure of Correlation 2 (IMC2): Additional measure of correlation between pixel pairs.

2.5.4 Standardization

Using Standardization is a data preprocessing technique used to transform the features of a dataset so that they have a mean of 0 and a standard deviation of 1. This process involves subtracting the mean of each feature from the data point and then dividing by the standard deviation of the feature. It helps to bring all features to a similar scale, preventing features with larger scales from dominating the learning process and improving the convergence speed of optimization algorithms. Using our aforementioned methods result in 53 extracted features,

36 of which are features about color and 14 of which are texture features. The features had varying minimum and maximum value ranges thus standardization was necessary.
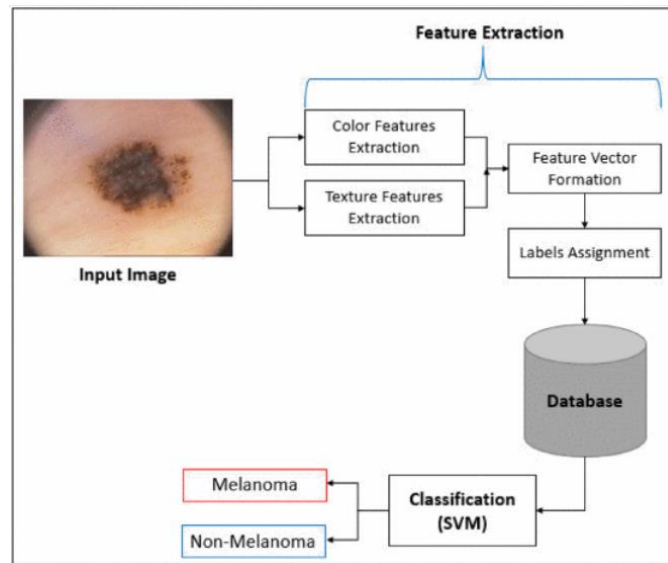
Figure 2.6 :   Example showing different features being used [31]

In short, feature extraction and classification refer to the process where a machine learning model identifies significant features in the dataset and uses them to classify input data. The stages involved in feature extraction and subsequent classification are illustrated in the above Figure 2.6.

2.5.5   Results

In the analysis of machine learning models for lesion classification, four models—K-Nearest Neighbors (KNN), Support Vector Machine (SVM), Logistic Regression, and Random Forest—were evaluated across various performance metrics.

2.5.5.1 Original dataset results
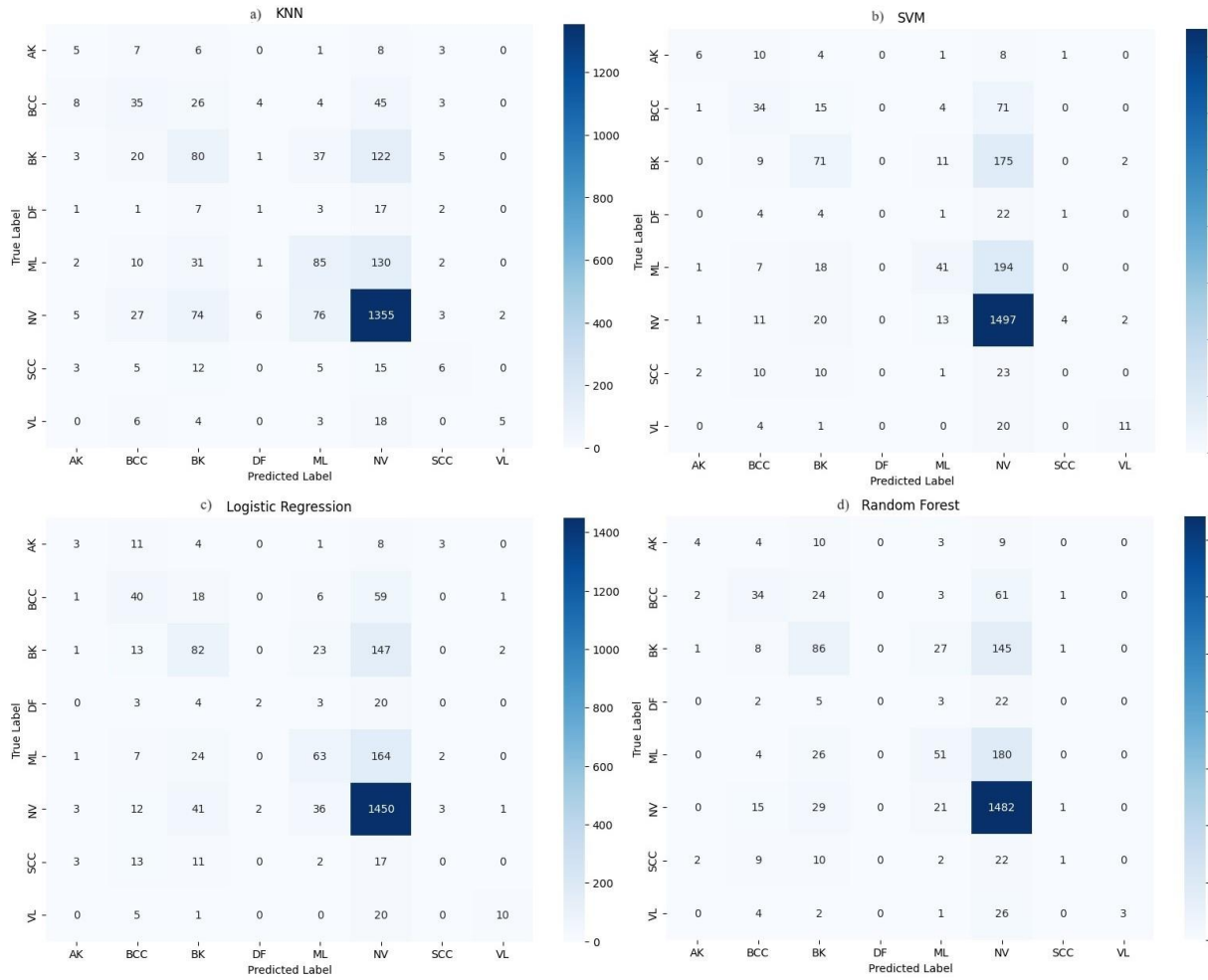
Figure 2.7 :   Confusion matrices of test results for original data with a)KNN, b) SVM, c) Logistic Regression, d) Random Forest.

Figure 2.7 shows four confusion matrices comparing the performance of different machine learning models on a multi-class classification task. Each matrix represents the performance of a model on eight classes (AK, BCC, BK, DF, ML, NV, SCC, VL) that are likely to be associated with lesion classification given the context.

Table 2.7 demonstrates the performance metrics of KNN, SVM and Logistic Regression.

Table 2.7 :   Performance metrics of KNN, SVM, Logistic Regression and Random Forest.

	Accuracy	Precision	Recall	F1-Score
KNN	0.67	0.38	**0.28**	0.30
SVM	**0.71**	0.43	0.27	0.31
Logistic Regression	0.70	0.44	**0.28**	**0.32**
Random Forest	**0.71**	**0.47**	0.25	0.28

The classification models, including K-Nearest Neighbors (KNN), Support Vector Machine (SVM), Logistic Regression and Random Forest were evaluated based on their performance metrics. The accuracy is quite similar across SVM, Logistic Regression, and Random Forest, all being around 0.70-0.71, with KNN trailing slightly behind at 0.67. Random Forest and Logistic Regression have better precision, indicating they make fewer false positives. However, they, along with the other models, have low recall, meaning they miss many true positives. This is likely due to class imbalanced nature of the dataset. Logistic Regression has the highest F1-score, indicating the best balance between precision and recall.

The average accuracy across all models is 0.70 which at first glance seems reasonable. However upon inspecting the confusion matrices all the models show a concentration of predictions on the majority class, nevus , which is a common issue in imbalanced datasets. This indicates that the models are biased towards predicting the majority class, likely due to the disparity in the number of training samples for each class.

The observed trends indicate that the models are not effectively learning to differentiate between classes, especially the minority ones. This leads to underdetection of the minority classes and a false sense of accuracy due to a seemingly high accuracy only resulting from majority class predictions. To tackle these low performance metrics and improve the confusion matrices the data augmentation and balancing strategies are applied.

2.5.5.2 Augmented dataset results
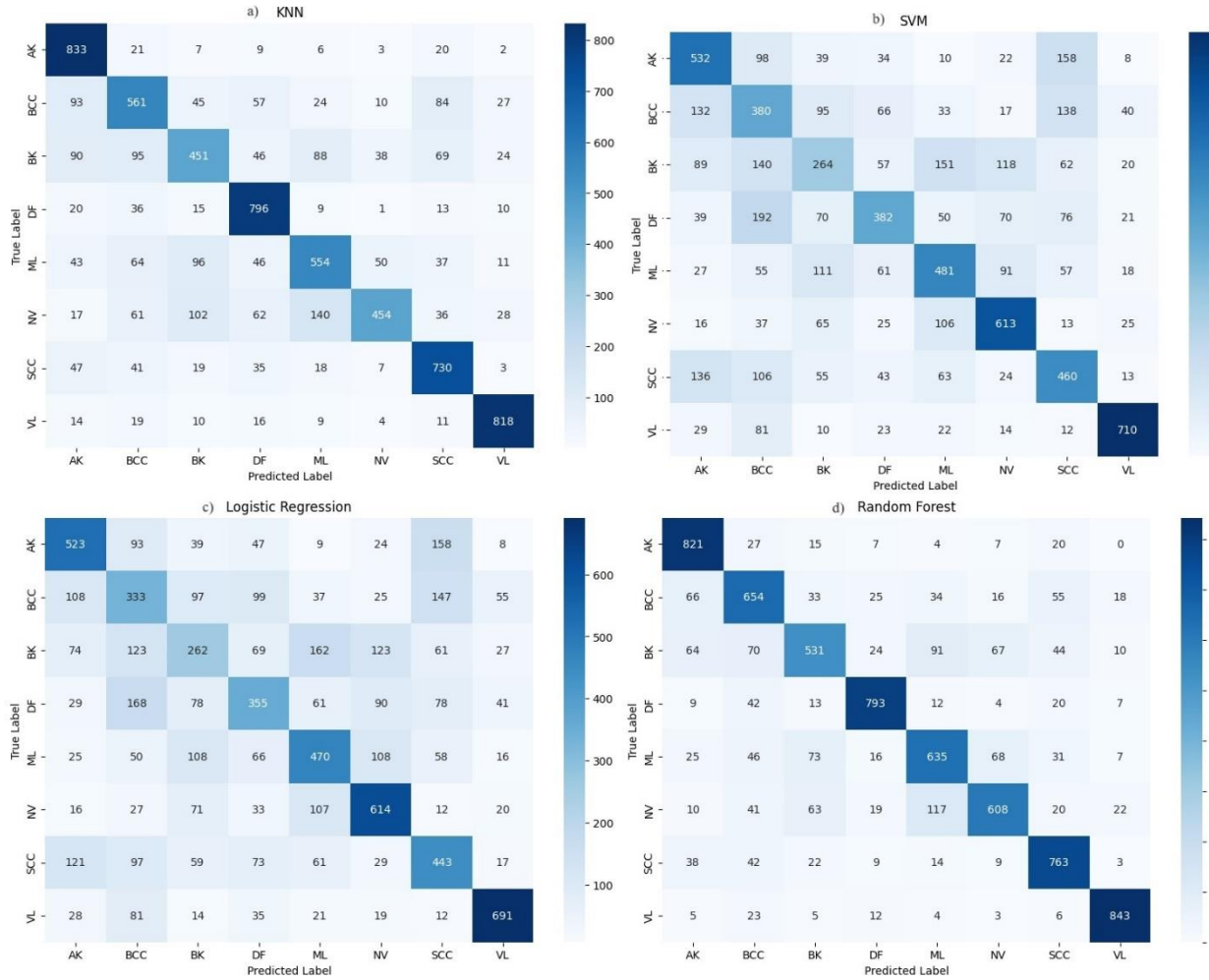
Figure 2.8 : Confusion matrices of augmented data a) KNN, b) SVM, c) Logistic Regression, d) Random Forest.

Table 2.8 : Performance metrics of KNN, SVM and Logistic Regression

	Accuracy	Precision	Recall	F1-Score
KNN	0.72	0.72	0.72	0.71
SVM	0.53	0.53	0.53	0.53
Logistic Regression	0.51	0.51	0.51	0.51
Random Forest	**0.78**	**0.78**	**0.78**	**0.78**

For the augmented dataset Random Forest emerges as the top performer, achieving the highest accuracy at 78% and balanced metrics, indicating that it has effectively leveraged the

balanced dataset. This suggests that Random Forest is well-suited for the current task with the improved data. The improvement in accuracy and balanced metrics reflects a positive outcome of the data balancing efforts. KNN seems to benefit from the enhanced data distribution, showing more reliable performance. The drop in performance for SVM and Logistic Regression suggests that these models may not have fully adapted to the changes in the dataset. They still struggle with achieving high accuracy and balanced metrics despite the class balancing efforts.

The highlighted diagonal in the confusion matrices indicates that the models are performing well in classifying most images correctly. This is a positive sign, as it shows that the models are effectively identifying images of the same class with a relatively high degree of accuracy. Random forest comes up as the model with the most prominent diagonal showcasing its outstanding performance. In summary, the highlighted diagonal in the confusion matrices indicates that the data balancing efforts have positively impacted model performance, leading to more accurate and reliable predictions across classes.

In essence, while the class balancing techniques have improved KNN and Random Forest's performance, SVM and Logistic Regression might need further tuning or alternative approaches to fully benefit from the dataset changes.

2.6 Deep Learning

Deep learning revolutionizes the field of artificial intelligence by enabling computational models to learn intricate representations of data across multiple layers of abstraction. These models have made major advances in a number of disciplines, including object identification, speech recognition, and visual object recognition. They are made up of interconnected processing layers. Neural network topologies, which are modeled after the composition and operation of the human brain, are at the core of deep learning. These networks consist of layers of interconnected nodes, or neurons, with hidden layers between the input and output layers. The term "deep" refers to the numerous hidden layers in these networks, which can range from a few to hundreds or even thousands. Figure 2.9 shows the comparison of machine learning and deep learning.
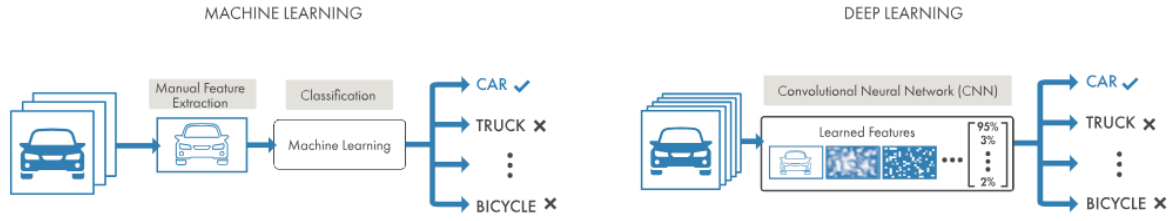
Figure 2.9 : Machine learning and deep learning comparison[32].

Deep learning models learn representations directly from raw data, eliminating the need for manual feature extraction. They achieve this by using large datasets with labeled examples to train the model parameters. Through the process of backpropagation, the model adjusts its internal parameters to optimize the mapping from inputs to outputs, gradually learning to capture complex patterns and relationships in the data. For example, in image processing tasks, lower layers may detect edges and textures, while higher layers combine these features to recognize objects and their relationships.

The power of deep learning lies in its ability to learn hierarchical representations from data, with each layer capturing increasingly complex features. These learned features are not handcrafted by human engineers but instead emerge from the learning process, making deep learning models highly adaptable and capable of handling diverse datasets and tasks.

2.6.1 Transfer learning

Transfer learning, a prevalent approach in image-based predictive modeling, involves repurposing pre-trained models for new tasks. Models like ImageNet, AlexNet, and Inception, known for object classification and general feature learning, serve as common starting points.

In dermatology datasets with limited skin lesion images, models are initialized with weights from networks pre-trained on extensive tasks like ImageNet. Choosing a pre-trained deep learning model trained on the ImageNet dataset is the first step for transfer learning. The technique assumes that pre-trained models possess relevant features. A standard practice is using selected layers as fixed feature extractors, freezing convolutional blocks, and training only the fully connected layers.
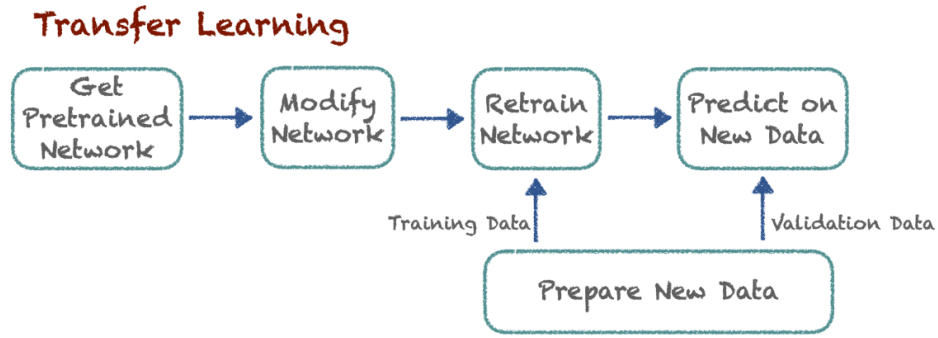
Figure 2.10 : Graphic description of transfer learning[33].

Fine-tuning extends this by selectively training higher-level convolutional layers. The proposed method freezes lower-level layers for generic features, training only the top layers for specificity. Initialization involves using ImageNet weights for the first four convolutional layers and loading saved weights for the final block. This approach optimally balances generic and specific feature utilization for dermatologic image classification. For all the deep learning models other than YOLO we will be using weights from ImageNet as initialization, since YOLO is primarily an object detection model initial weights from COCO (Common Objects in Context) are available thus they will be used.

2.6.2 Convolutional neural networks

One type of deep learning model that works especially well for processing 2D input, such as photos, is the convolutional neural network (CNN). CNNs utilize 2D convolutional layers to integrate learned features with the input data. When a CNN is trained on a dataset of images, it can automatically extract relevant features from the images. This ability for automatic feature extraction makes CNNs highly effective for image classification tasks. Additionally, CNNs can be adapted to classify text and other non-traditional data forms, such as time series.
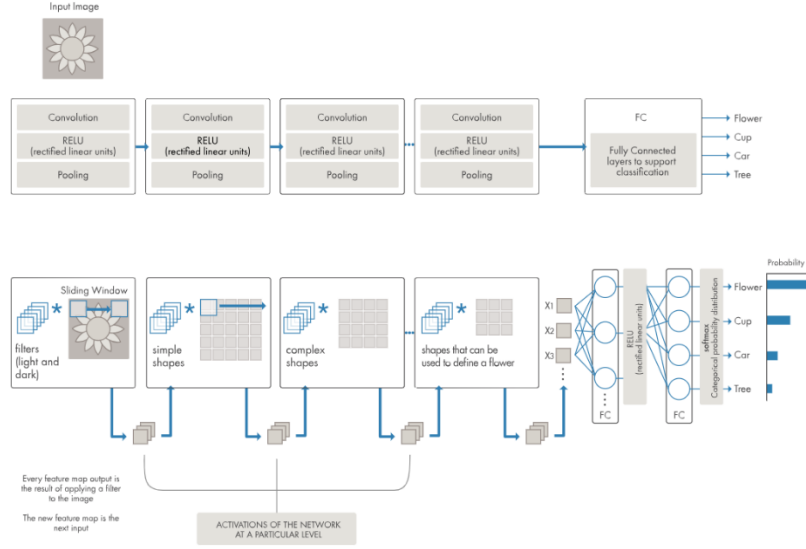
Figure 2.11 : CNN classification model structure [32].

2.6.2.1 Pre-trained models

We chose to use Xception, ResNet50 and VGG16 with weights from ImageNet as our pretrained models which we will build upon.
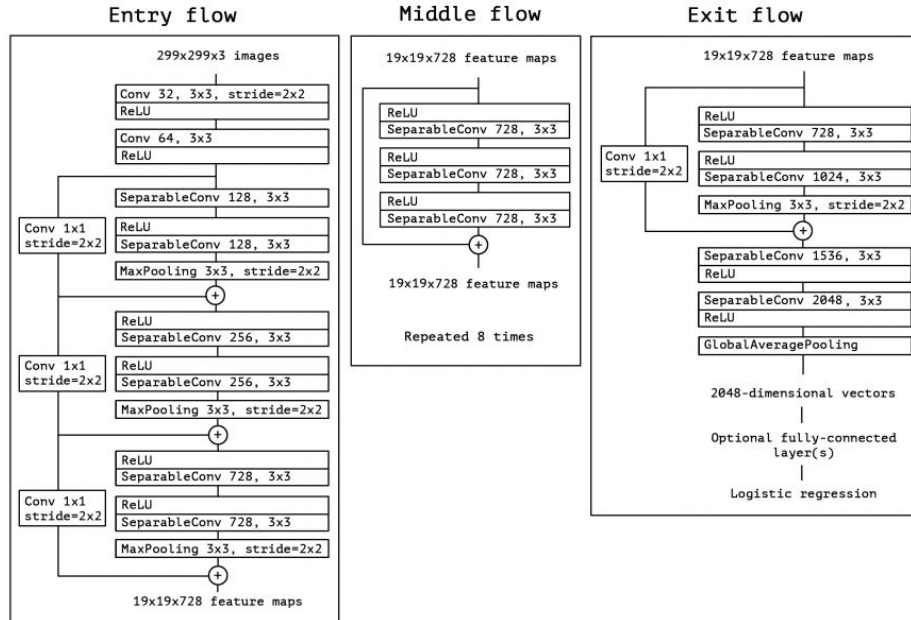
Figure 2.12 : Xception architecture [34]

The Xception model with its 71 layers, the model is designed to strike a balance between depth and computational efficiency. Xception primarily employs depth wise separable convolutions, separating the convolutional operation into depthwise and pointwise convolutions. This innovative approach drastically reduces the computational complexity and parameter count of the model, while maintaining high performance on various image recognition tasks.

The architecture of Xception is structured into three main sections: entry flow, middle flow, and exit flow, comprising a total of 14 modules. Each module, with the exception of the first and last ones, incorporates residual connections to facilitate gradient flow during training. These connections address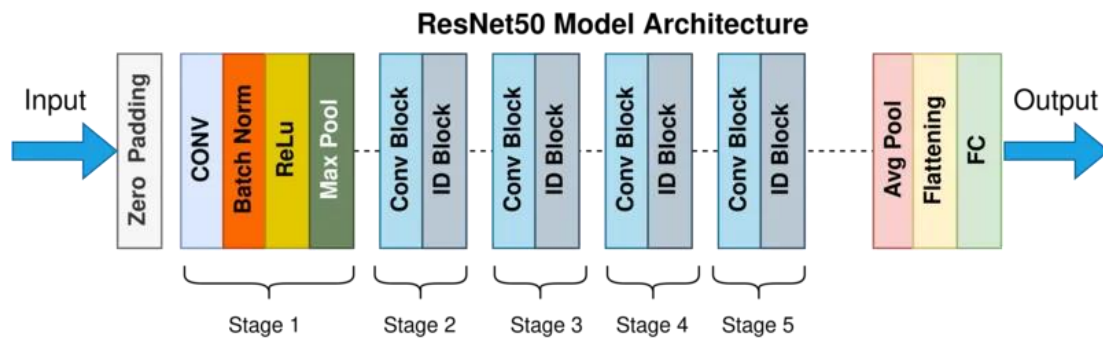 the vanishing gradient problem commonly encountered in deep networks, thereby enhancing the model's training stability and convergence.

Figure 2.13 : ResNet50 model architecture [35]

With its 50 layers, ResNet50 represents a balance between depth and computational efficiency. ResNet50's architecture is distinguished by its 16 residual modules, each comprising three convolutional layers, totaling to 49 convolutional layers. The model also incorporates a fully connected layer with Rectified Linear Unit (ReLU) activation, alongside maximum and average pooling layers. A defining feature of ResNet50 is its innovative residual modules, which introduce skip-connections to address the problem of training truly deep architectures. By allowing each layer to directly propagate its input to succeeding layers, these skip-connections efficiently address the problem of disappearing gradients that are frequently seen in deep networks. This unique jump-type structure significantly alleviates the

risk of gradient disappearance with network depth, thereby reducing the likelihood of overfitting while extracting deeper features.
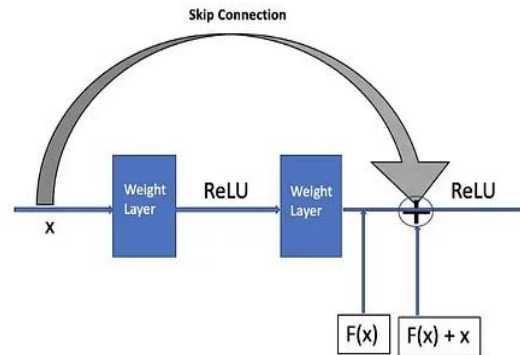
Figure 2.14 : ResNet50's skip-connections structure [36]

VGG-16, shown in Figure 2.15, is a deep learning model consisting of 16 layers, including 13 convolutional layers and 3 fully connected layers.
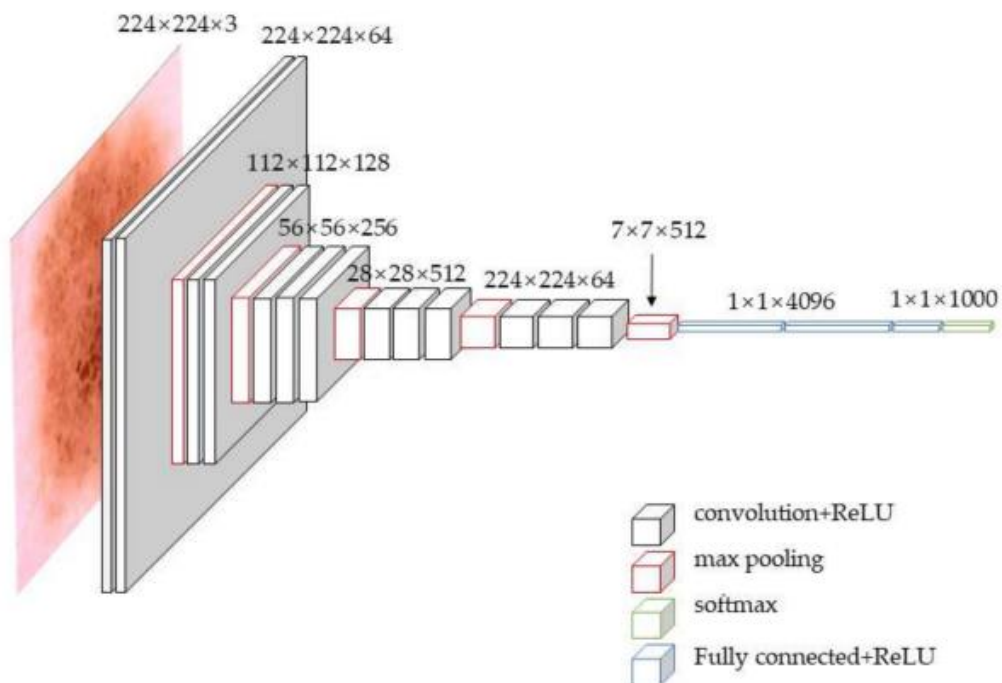
Figure 2.15 : VGG16 architecture [37]

It's well-known for its straightforward design and effectiveness, achieving strong results in tasks like image classification and object recognition. The architecture consists of a series of convolutional layers followed by max-pooling layers, with increasing depth, allowing the model to learn complex hierarchical representations of visual features. Although simpler than some newer architectures, VGG-16 continues to be a popular choice for various deep learning applications due to its versatility and strong performance.

2.6.2.2 Training process

The training process involves using a pre-trained Xception, VGG16 AND ResNet50 models to leverage learned image features. Custom top layers are added. GlobalAveragePooling2D is applied to reduce the output dimensions from the convolutional base to a manageable size while retaining important spatial information.Adense layer with 1024 units and ReLU activation is added for additional learning capacity final layers produces class probabilities.

Initially the layers of the pre-trained model are frozen in order to prevent them from being updated during the first training phase. This focuses learning on the new top layers and avoids disrupting the pre-trained features. The model is compiled with the Adam optimizer and a learning rate of 0.001. The loss function is categorical crossentropy which is suitable for multi-class classification tasks.The model is trained for 10 epochs. After the initial training, fine-tuning is performed by unfreezing some of the base layers and re-training the model to improve classification performance. This enables the model to adapt the pre-trained features more specifically to the skin lesion classification task.The model is recompiled with a lower learning rate of 0.0001 for 15 epochs to refine the weights of the unfrozen layers without drastically altering the previously learned features.The final model, inclusive of the learned weights and architecture, is then saved for deployment or further evaluation.
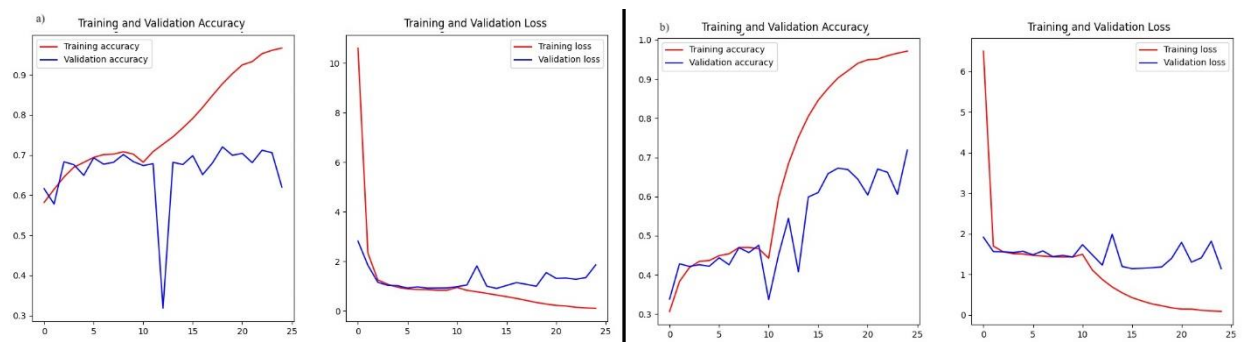
Figure 2.16 : Training and validation accuracy and loss progression of Xception based model trained on a) original data, b) augmented data.

For the Xception based model trained on original data;the training accuracy increases steadily and shows a clear upward trend, the validation accuracy fluctuates and does not show a consistent improvement. The gap between training and validation accuracy widens as training progresses, suggesting potential overfitting. The training loss decreases rapidly and continues to trend downward, however the validation loss fluctuates and does not decrease consistently, which is another indicator of overfitting. When augmentation is done validation accuracy shows improvement and has less fluctuation than the original data, though it still varies. The gap between training and validation accuracy is smaller than in the original data, suggesting better generalization.The validation loss also shows less fluctuation. This suggests that the model benefits from data augmentation, leading to better generalization and reduced overfitting.
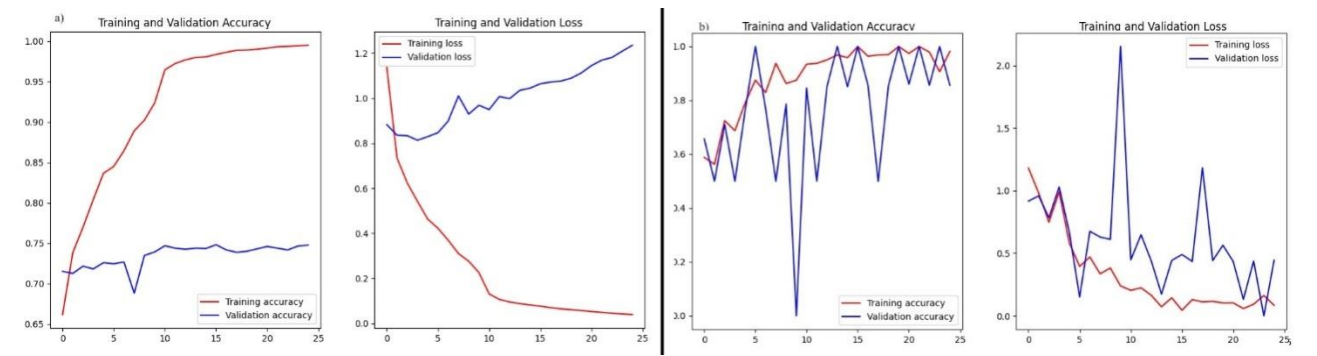
Figure 2.17 : Training and validation accuracy and loss progression of VGG16 based model trained on a) original data, b) augmented data.

Regarding the VGG16 model with original data, the training accuracy for the original data increases quickly, reaching almost 100%, but the validation accuracy plateaus early and remains significantly lower, suggesting strong overfitting.The training loss decreases rapidly, while validation loss decreases initially but then starts to increase, further confirming overfitting. When trained on augmented data; training accuracy improves steadily and reaches a high value, but validation accuracy fluctuates and doesn't improve consistently. Training loss decreases steadily, while validation loss fluctuates, but it doesn't increase as sharply as in the original data, indicating better but still limited generalization.
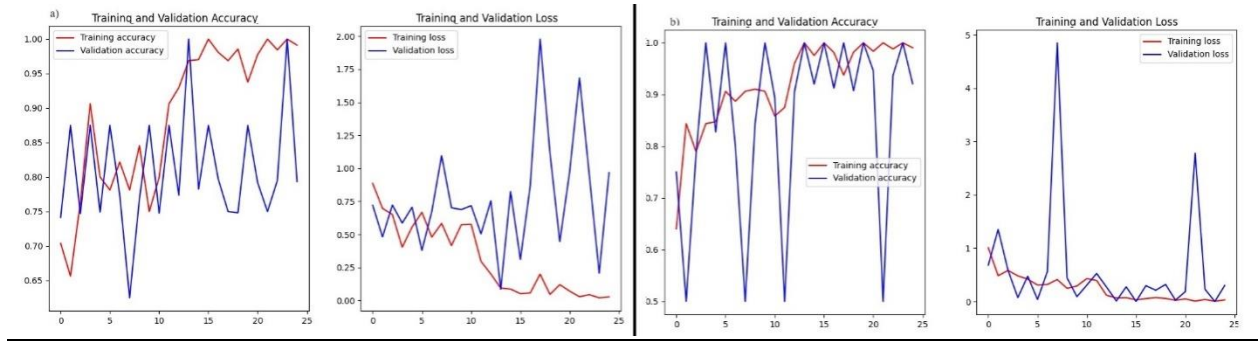
Figure 2.18 : Training and validation accuracy and loss progression of ResNet50 based model trained on a) original data and b) augmented data.

Lastly for ResNet50 model for original data, the training accuracy improves steadily, but validation accuracy fluctuates significantly, indicating that the model might be struggling to generalize well to unseen data.Training loss decreases consistently, while validation loss shows large fluctuations, suggesting overfitting where the model performs well on training data but not on validation data. As for the model trained on augmented data,the training accuracy is high and consistent, but the validation accuracy still shows considerable fluctuations. However, it appears to be more stable towards the end. Training loss decreases smoothly, and validation loss fluctuates but seems to follow a decreasing trend, indicating better generalization compared to the original data.

In summary it can be said that all three models have benefitted from the augmentation as it has improved the overfitting problem. There is still room for improvement as all the models show significant fluctuations during training and while improved overfitting problem is still existant.

2.6.2.3 Deep learning results

Original dataset results

ResNet50 has the highest accuracy at 0.81, followed byVGG16 at 0.76 then Xception at 0.61. This suggests that ResNet50 is generally performing better at correctly classifying the skin lesions compared to the other two models.

ResNet50 has the highest precision at 0.70, followed by VGG16 at 0.61, and Xceptionfalling considerably behind at 0.37. High precision implies that the models are good at minimizing false positives.

ResNet50 has the highest recall at 0.59, indicating it captures more true positive cases compared to VGG16 at 0.43and Xception at 0.37. The relatively lower recall values for VGG16 and Xception shows that there is room for improvement.

The F1-score for ResNet50 is the highest at 0.64, reflecting a good balance between precision and recall. This indicates that ResNet50 handles the trade-off between false positives and false negatives better than the other models.VGG16 has a lower F1-score of 0.48, indicating that while it performs well in terms of precision, its lower recall brings down its overall balance.Xception has the lowest F1-score at 0.33, reflecting its struggles with both precision and recall.

ResNet50 emerges as the best-performing model across all metrics, indicating its robustness in handling the classification task, despite the challenges posed by potential class imbalances.VGG16 performs moderately well but suffers from a lower recall, which suggests it misses more true positive cases. The relatively low recall and F1-scores across all models suggest that the models are struggling with the class imbalance, as they tend to favor the majority class.

The confusion matrices of the models backs up the commentary on the performance metrics, which show a tendency to correctly classify the majority class while underperforming on minority classes.The accuracy is objectively high however this is mainly due to correct detection the majority class, the classification abilities when it comes to minorty classes can benefit from some improvements.
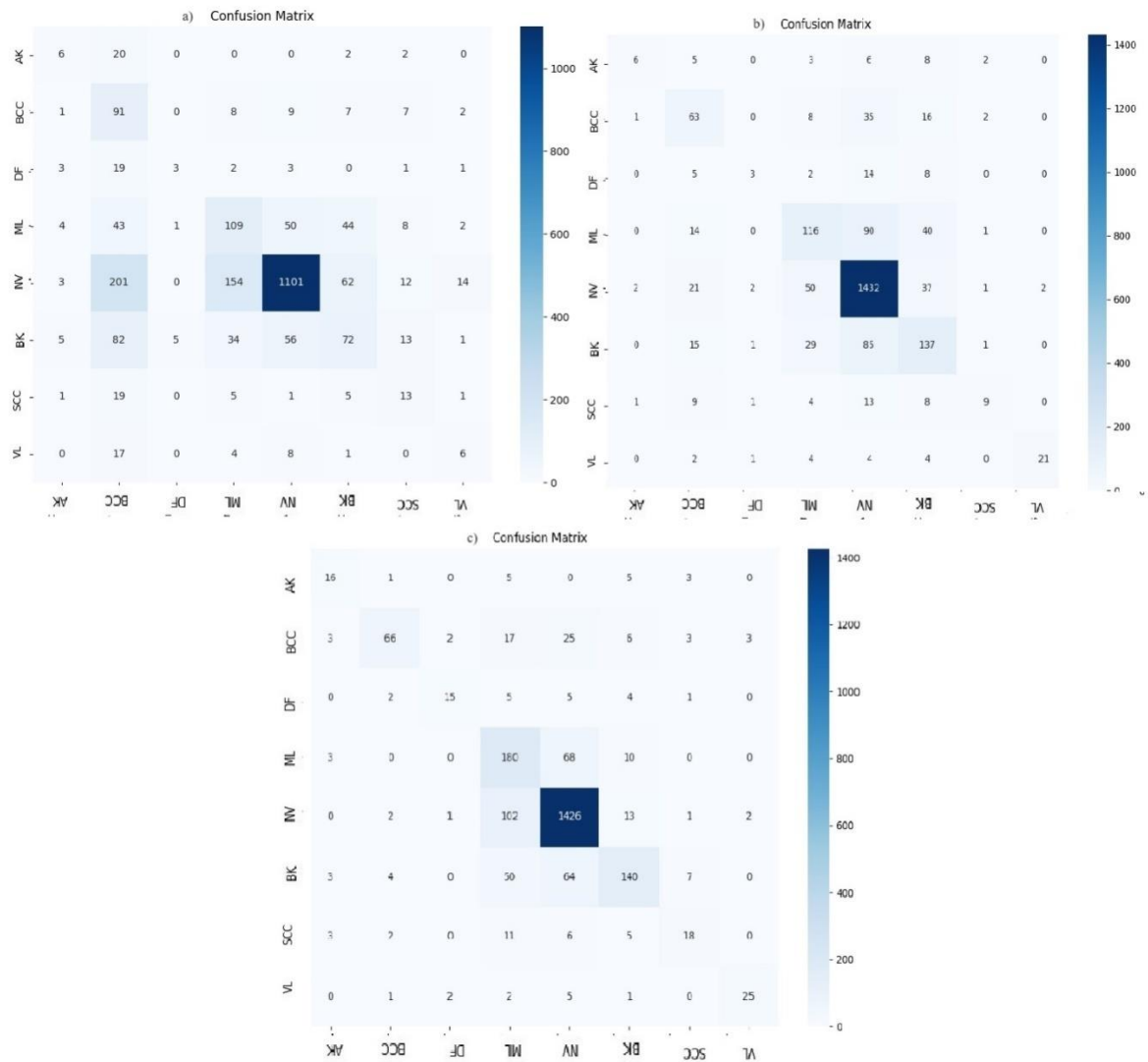
Figure 2.19 : Confusion matrices for original data with a) Xception, b) VGG16 and c) ResNet50 based models.

Table 2.9 :   Performance metric results for Xception, VGG16 and ResNet50 based models'original data results.

	Accuracy	Precision	Recall	F1-Score
Xception	0.61	0.37	0.37	0.33
VGG16	0.76	0.61	0.43	0.48
RestNet50	**0.81**	**0.70**	**0.59**	**0.64**

Augmented dataset results

Post-augmentation, ResNet50 demonstrates superior performance across all metrics, confirming it as the most effective model for classifying skin lesions. Its precision and recall are particularly impressive, indicating it is both accurate and reliable.VGG16 shows marked improvement across all metrics, performing well with balanced precision and recall. It is now a strong performer, though still slightly behind ResNet50.While Xception has improved with augmentation, its performance remains lower than the other models. It has made strides but still requires further refinement, particularly in balancing precision and recall.

Augmentation has clearly enhanced the performance of all models, particularly ResNet50 and VGG16. The results suggest that these models now handle the dataset more effectively, with ResNet50 emerging as the most robust and accurate model. Xception, though improved, still trails behind and could benefit from further tuning or additional augmentation techniques.

The diagonal elements of the confusion matrix represent the number of correct predictions for each class. The highlighted diagonal shows that the majority of the predictions fall along this diagonal, indicating that the models are accurately predicting the correct class for most instances.

The balanced dataset appears to have mitigated a big portion the issues associated with class imbalance, such as bias toward majority classes. This balance is reflected in the consistent higher precision, recall, and F1-scores, indicating that the models are performing better across all classes, not just the dominant ones.

The class balancing strategy has effectively enhanced the models' ability to generalize across all classes, leading to higher accuracy and balanced precision and recall. This underscores the importance of addressing class imbalance in machine learning tasks, especially in classification problems involving medical data, where accurate predictions are crucial.
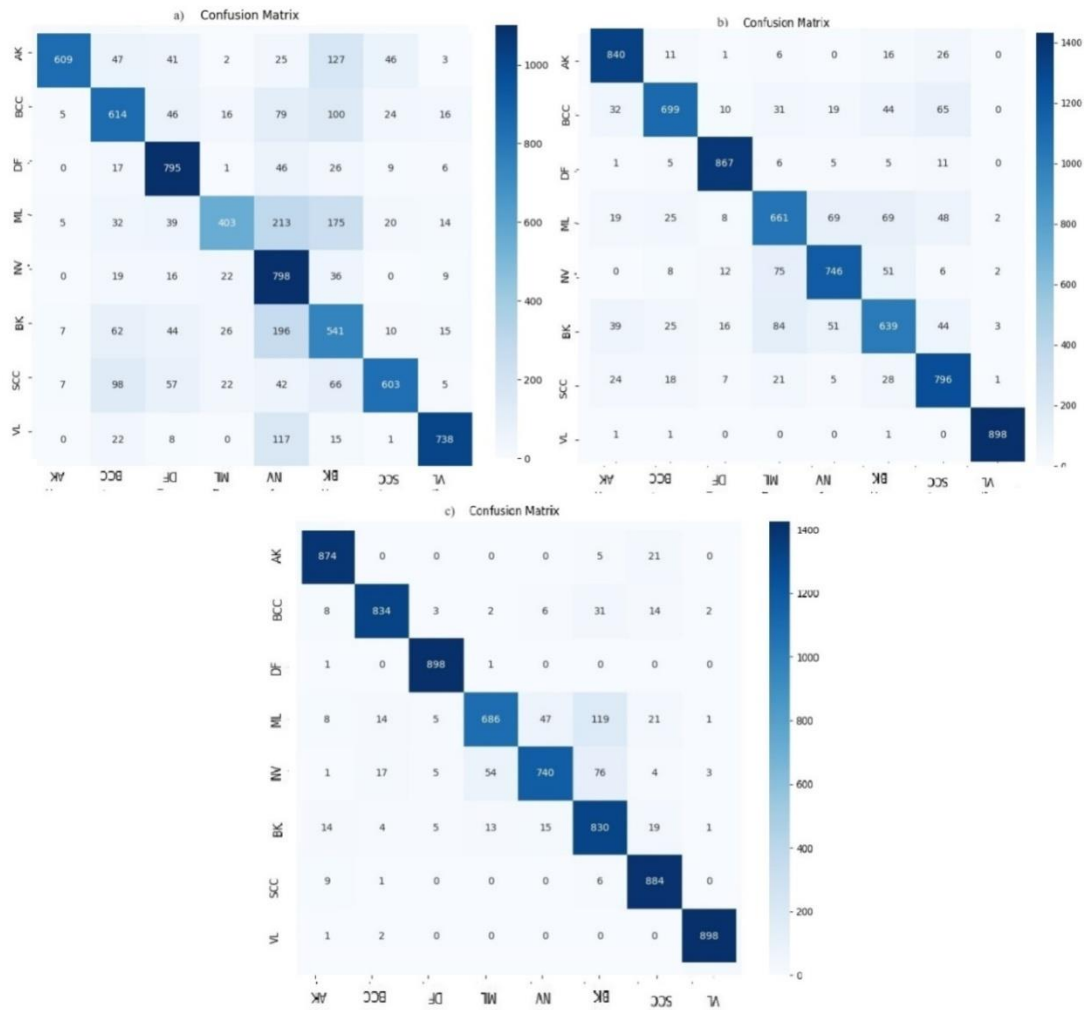
Figure 2.20 : Confusion matrices for collected data test results a) Xception, b) VGG16, c) ResNet50 based models.

The ResNet50 model, as shown in Figure 2.20, outperforms the others by achieving the highest overall classification accuracy and making fewer errors, making it the most reliable model in this context.

Table 2.10 : Performance metric results for Xception, VGG16 and ResNet50 based models for augmented data test results.

	Accuracy	Precision	Recall	F1-Score
Xception	0.71	0.71	0.71	0.70
VGG16	0.85	0.85	0.85	0.85
RestNet50	**0.92**	**0.93**	**0.92**	**0.92**

ResNet50 model comes as the top performer for both dataset and shows and improvement on all metrics when trained with the balanced dataset as can be seen from the table below.

Table 2.11 : Performance metrics for the highest-performing model, ResNet50, on both original and augmented data.

	Accuracy	Precision	Recall	F1-Score
Original	0.81	0.70	0.59	0.64
Augmented	**0.92**	**0.93**	**0.92**	**0.92**

2.6.3 YOLO

YOLOv5 (You Only Look Once version 5) is an advance deep learning model designed for object detection. The goal of the artificial intelligence field of object detection is to identify both the types and locations of things in a picture.The reason for the term "You Only Look Once" is that it can analyze an image in a single step, meaning that it doesn't require passing through the gateway more than once. This approach contrasts with traditional methods that might require multiple passes over the image, making YOLOv5 highly effective for quick and accurate analysis. In the medical field, YOLOv5's capabilities extend to diagnosing pathological conditions.

2.6.3.1 YOLO Architecture

The YOLO architecture is divided into three main components: the backbone, the neck, and the head. The backbone is responsible for feature extraction through convolutional operations. The neck, which may include layers such as Feature Pyramid Networks (FPN) or Path Aggregation Networks (PAN), refines and enhances these features at multiple scales. Finally, the head consists of fully connected layers that predict the bounding boxes, objectness scores, and class probabilities for each detected object.

A key feature of YOLO is its use of a grid system, where the input image is divided into a grid of S × S cells. Each cell is responsible for predicting bounding boxes and class probabilities for objects whose center falls within the cell. This design allows YOLO to process images in a single forward pass, enabling real-time detection.

Additionally, YOLO employs techniques like anchor boxes and non-maximum suppression (NMS). Anchor boxes help the model predict multiple objects in a single cell by providing predefined shapes and sizes, while NMS reduces the number of overlapping bounding boxes by keeping the one with the highest confidence score [38].
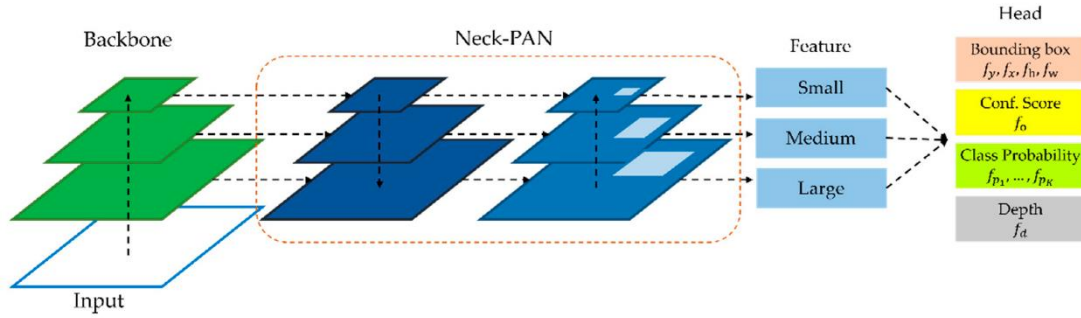
Figure 2.21 : YOLO architecture [38]

Figure 2.22 shows the pre-trained data and how accurately they were detected. The shape and colour distribution of the lesions can cause errors in the predicted labels.
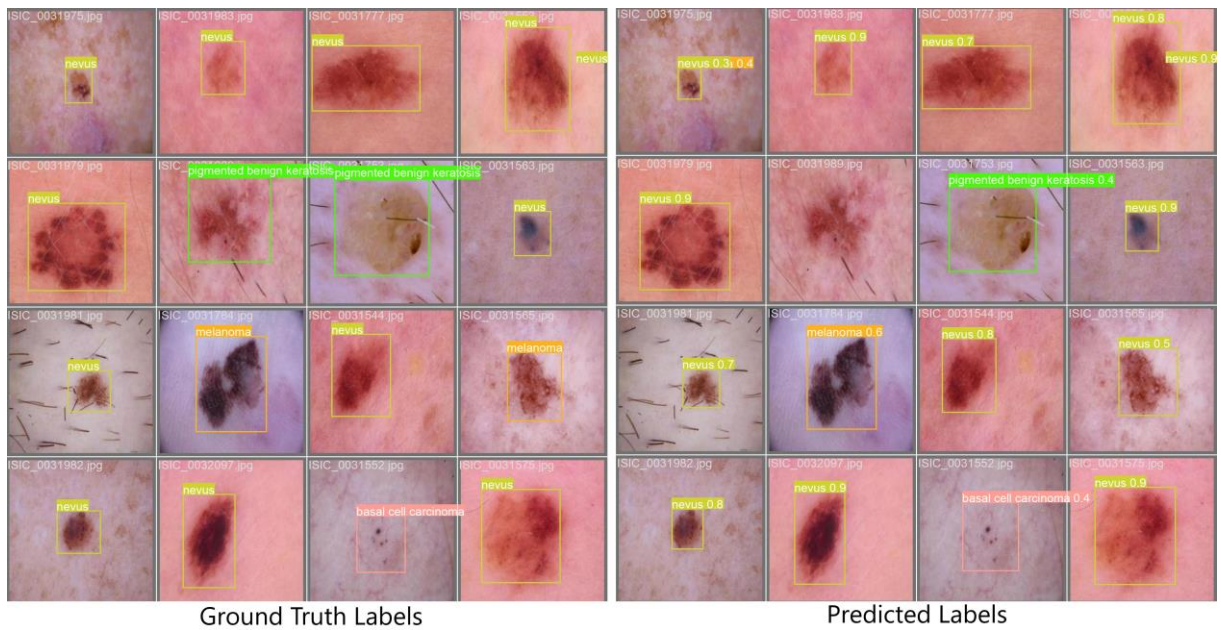
Figure 2.22 : Real and predicted labels by Yolov5

2.6.3.2 YOLOv5 results
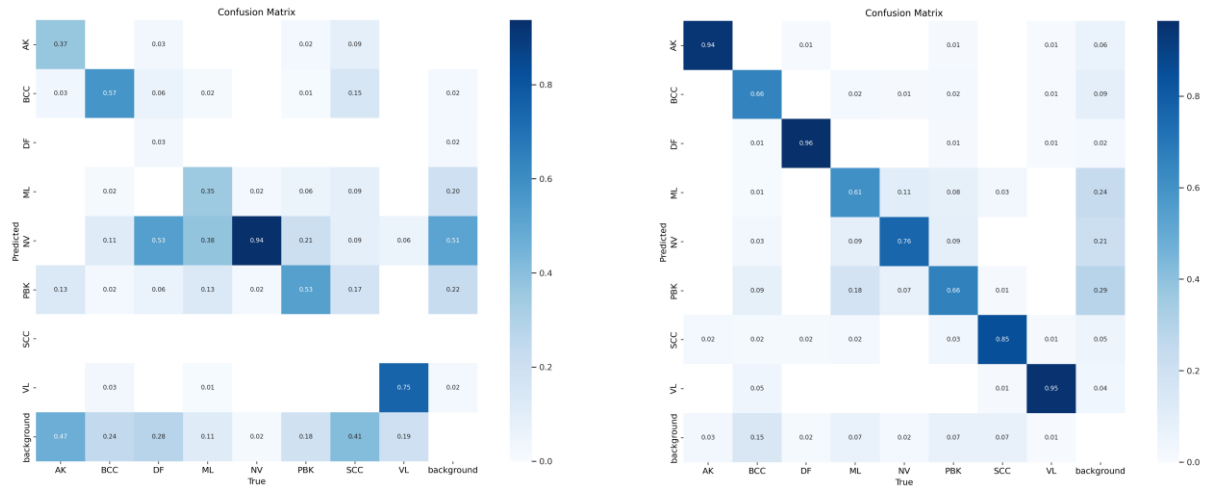
Figure 2.23 : YOLOv5 confusion matrices for a) original data, b) augmented data

The image in Figure 2.23 shows two confusion matrices regarding the performance of the YOLOv5 model on a multi-class classification task, which is likely related to skin lesion detection given labels like AK, BCC, DF, etc. After data augmentation, the model performance improved significantly and most classes were predicted correctly as indicated by the high values along the diagonal. It is quite clear from Figure 2.23 that there was a significant reduction in misclassifications, especially in difficult categories like NV and background. Data augmentation probably provided the model with a more diverse and representativedataset, resulting in better generalization and accuracy.
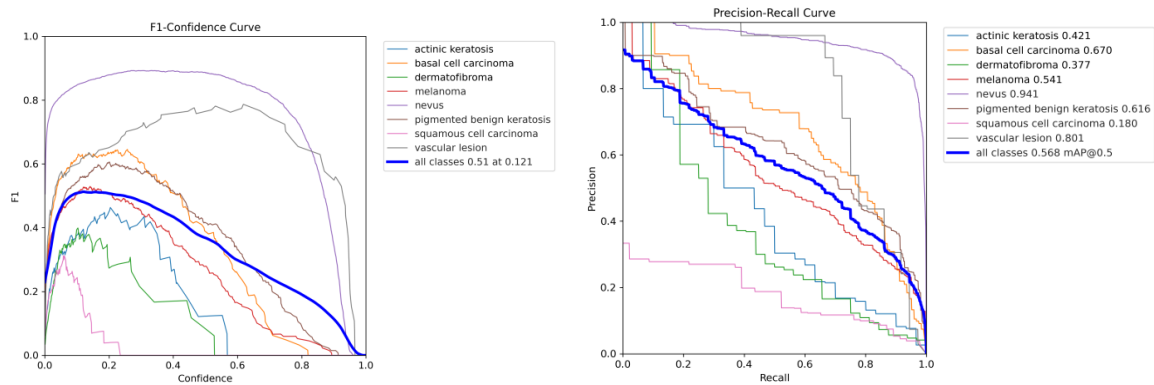
Figure 2.24 : Orjinal data's a) F1 Score and b) PR Curve

Nevus (purple line) and vascular lesion (black line) in Figure 2.24 have the highest peak F1 scores, indicating that the model performs better in these classes, especially at higher

confidence levels. Other classes such as actinic keratosis (blue), basal cell carcinoma (orange), melanoma (red), and dermatofibroma (green) show lower F1 scores at different confidence levels, indicating that these classes are more difficult for the model to correctly classify. This may be due to the unbalanced class distribution in the original data. The model's appropriate operating point is indicated by the thick blue line, which represents the average performance across all classes and has an F1 score of 0.51 at a confidence level of 0.121. Overall, there is variability, especially for classes with lower F1 scores, indicating possible areas for improvement, even though the model performs strongly for some classes.

The Precision-Recall (PR) curve in Figure 2.24 represents the performance of a YOLOv5 model across various classes in a multi-class classification task. Each colored line corresponds to a different class and shows how precision varies with recall for that particular class. The model's appropriate operating point is indicated by the thick blue line, which represents the average performance across all classes and has an F1 score of 0.51 at a confidence level of 0.121. Overall, there is variability, especially for classes with lower F1 scores, indicating possible areas for improvement, even though the model performs strongly for some classes.

,
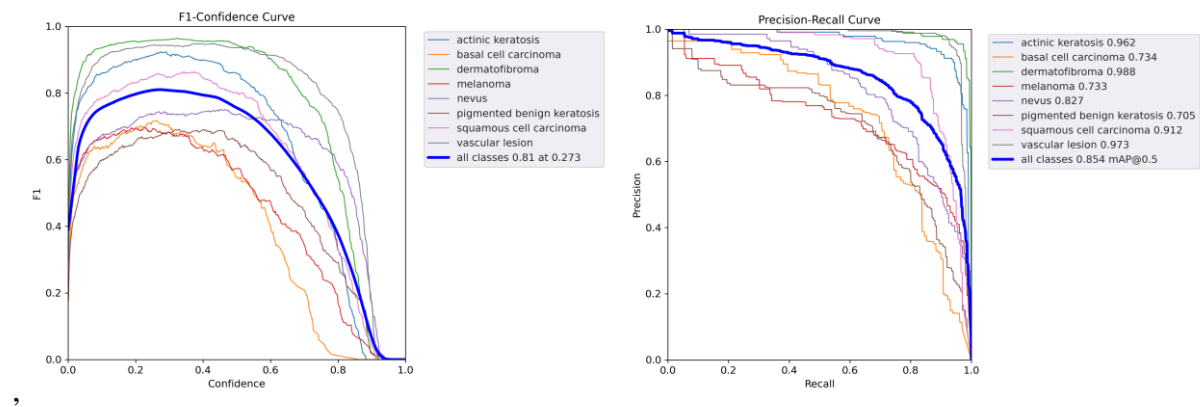
Figure 2.25 : Augmented data's a) F-1 Score and b) PR curve.

The model achieves a maximum F1 score of 0.81 at a confidence level of 0.273. This indicates the optimal threshold where precision and recall are best balanced for the overall model. Classes such as vascular lesion (grey) and dermatofibroma (green) maintain higher F1 scores across a wide range of confidence levels, indicating consistent model accuracy. Melanoma (red) and pigmented benign keratosis (brown), in particular, show lower F1 scores

as the confidence level increases, indicating that these classes are more difficult for the model to correctly classify. However, it can still be observed that the classification accuracy increases when looking at the original data plots in Figure 2.25.

Precision-Recall (PR) curve shows how precision (the proportion of true positive predictions out of all positive predictions) varies with recall (the proportion of true positive predictions out of all actual positives) for each class, as well as the overall performance across all classes. It can be shown that dermatofibrama (cyan curve) has the highest PR performance with an area under the curve (AUC) of 0.988, indicating the model is highly effective at distinguishing this class. More fine-tuning of the YOLOv5 model might help improve the overall mAP and the precision-recall balance across different classes.

Table 2.12 :   Performance metric results for YOLOv5 on both the original and augmented datasets

	Precision	Recall	F1	mAP50
Original	0.466	0.581	0.51	0.568
Augmented	0.8	0.824	0.81	0.854

2.6.3.3 YOLOv5n

YOLOv5n (YOLOv5 Nano) is designed to be a smaller, more efficient variant of the YOLOv5 family. In medical settings where real-time analysis or limited computational resources are critica nano model can be advantageous.In many medical environments, particularly in remote or resource-limited settings, there may not be access to powerful GPUs or extensive computational infrastructure. YOLOv5n, being a lighter model, can be deployed on less powerful hardware, such as mobile devices or embedded systems, without compromising too much on performance.While YOLOv5n offers several advantages, it's important to consider that its smaller size might come at the cost of reduced accuracy, particularly in detecting subtle features in medical images, which are often critical for accurate diagnosis.

2.6.3.4 Yolov5n Results
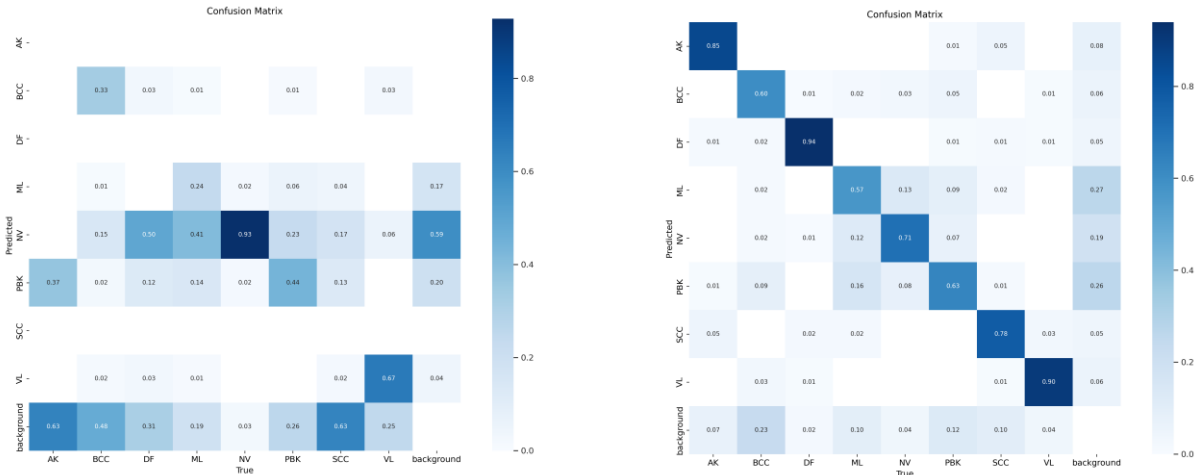
Figure 2.26 : YOLOv5n confusion matrices for a) original data, b) augmented data

Light-YOLOv5 achieves detection precision that permits lesion discrimination while drastically reducing the number of parameters and calculations. Regarding the algorithm's accuracy, there is still space for development. In matrix (b) with augmented data, there is a noticeable improvement in correct classifications (diagonal elements are darker blue) compared to matrix (a) with original data. This indicates that data augmentation helped the model learn better and classify more accurately. In comparison to matrix (a) (0.67), certain classes, such as SCC, exhibit a notable improvement in classification accuracy in matrix (b) (0.78). Similar gains are also seen in other classes, such as AK and VL.Overall, data augmentation positively impacts the model's ability to generalize and improve its classification accuracy. Future research endeavors will delve deeper into exploring ways to enhance the algorithm's accuracy in lesion identification applications.

Table 2.13 : Performance metric results of YOLOv5n for original and augmented data

	Precision	Recall	F1-Score	mAP50
Original	0.45	0.507	0.39	0.462
Augmented	0.754	0.783	0.77	0.821
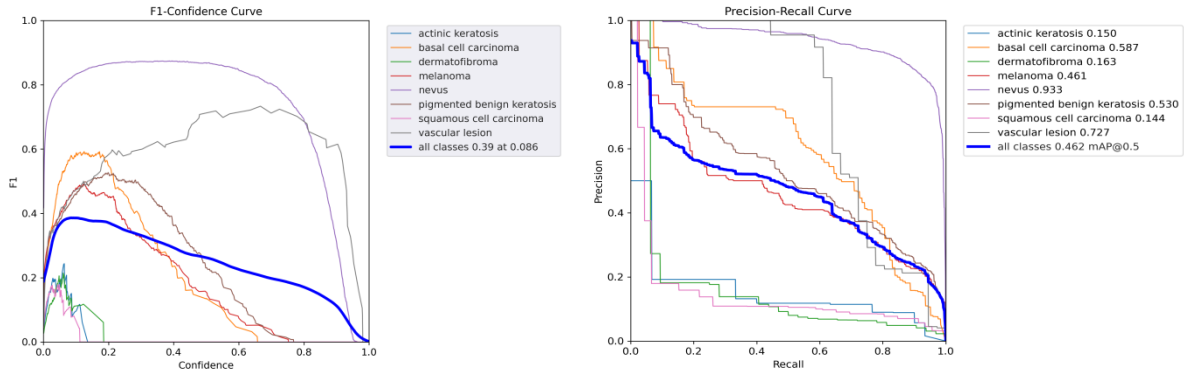
Figure 2.27 : F1-Curve and PR-Curve for orijinal

The F1 curve in Figure 2.27 shows that Nevus (purple line) has the highest peak F1 score, indicating that the model performs best in detecting this class at a given confidence level. Other classes such as Basal Cell Carcinoma (orange) and Pigmented Benign Keratosis (brown) have intermediate peaks but have lower overall performance compared to Nevus. The thick blue line represents the overall performance across all classes. The F1 score is relatively low (0.39) at a confidence threshold of 0.086, indicating that the model has average precision and recall across all classes, with significant variability depending on the class. The overall performance seen for the original data (across all classes) suggests that there may be room for improvement, particularly in balancing detection capabilities across all lesion types.
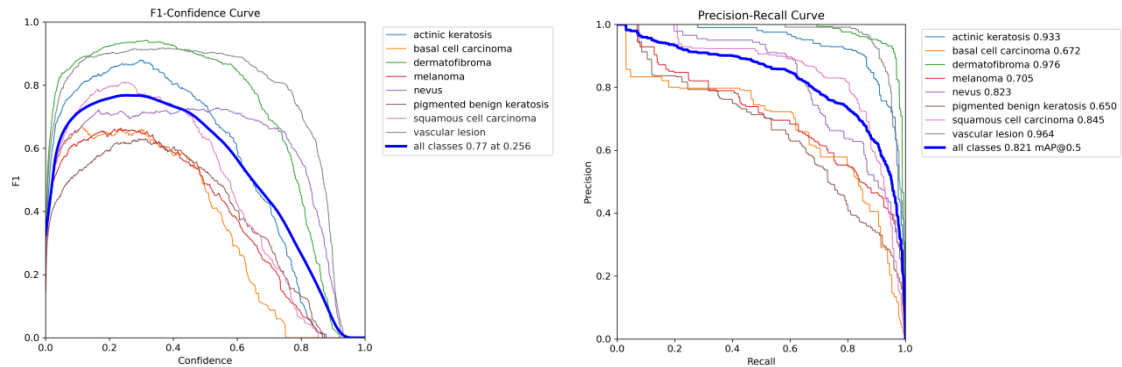
Figure 2.28 : YOLOv5n augmented data: a) F1 score, b) PR Curve

Each class's F1 score at various confidence thresholds is represented as a curve on the graph; classes like "melanoma," "nevus," and "vascular lesion" have comparatively larger F1 scores, indicating greater performance at particular confidence levels. Conversely, classes with lower

F1 scores—such as "pigmented benign keratosis" and "basal cell carcinoma"—indicate that the model has greater difficulty with these classes. In other words, there is potential for improving the performance of the model, especially in the identification of "basal cell carcinoma" and "pigmented benign keratosis," and it is crucial to carefully choose the confidence threshold depending on the specific class and the intended balance between sensitivity and recall.

The Precision-Recall (PR) curve in Figure 2.28 shows the performance of a YOLOv5n model on various skin lesion classes after data augmentation. The model achieves a mean Average Precision (mAP) of 0.821 at 0.5 IoU threshold across all classes. The highest precision-recall performance is seen for "dermatofibroma" with 0.976, while the lowest performance is seen for "basal cell carcinoma" with 0.672.

3. RESULTS

3.1 Comparing Results

The development and application of the machine learning model for skin cancer diagnosis yielded promising results, demonstrating its potential as an effective tool in distinguishing between different classes of skin lesions.The results demonstrate the strengths and weaknesses of different approaches and highlight the potential of advanced machine learning techniques in improving diagnostic accuracy for skin cancer.

Before we begin a performance metric comparison for models it's important to understand that "accuracy" in object detection tasks, such as those typically handled by, is a bit different from accuracy in traditional classification tasks. YOLO models are evaluated based on how well they detect objects and localize them within an image, not just on whether they correctly classify an image.While accuracy is not typically used as a standalone metric in object detection, you can calculate a form of accuracy by considering the proportion of correctly detected and classified objects over the total number of objects.However, this approach is less informative than mAP because it does not account for the localization aspect. mAP is the most relevant and standard metric for YOLO and other object detection models that can be used in place of accuracy score for comparison with other models, like classification models.

Table 2.13and Table 2.14 present the performance metrics for various machine learning models and deep learning architectures tested on the original datatset and the augmented dataset. The models evaluated include traditional classifiers like K-Nearest Neighbors (KNN), Support Vector Machines (SVM), and Logistic Regression (LR), Random Forest as well as deep learning models such as Xception, VGG16, ResNet50, YOLOv5, and YOLOv5n. Each model's performance is measured in terms of accuracy/mAP, precision, recall, and F1-score.

ResNet50 is the top performer across both datasets. For the original dataset, it achieves an accuracy of 81.93%, with strong precision and recall, leading to an impressive F1-Score of 64%. With the balanced dataset, its performance further improves to an accuracy of 92%, precision of 93%, recall of 92%, and an F1-Score of 92%, demonstrating exceptional robustness and balance.

VGG16 also shows significant improvement with the balanced data, reaching an accuracy and F1-Score of 85%, indicating strong performance across all metrics.Random Forest benefits from the balanced dataset, with improved accuracy and F1-Score of 78%, reflecting its enhanced ability to classify effectively when data is balanced.YOLOv5 and YOLOv5n, designed for object detection, show better results with the balanced dataset compared to the original. YOLOv5n, in particular, achieves an accuracy of 82% and an F1-Score of 77%, making it competitive with some traditional and deep learning models.

Overall, data augmentation significantly enhances the performance of most models, especially deep learning models like ResNet50 and VGG16, and also improves YOLO models in the classification context.

For future research considering YOLOv5n's results for augmented data, lesion classification can be accomplished through a mobile application, and these models can be operated on mobile devices due to the simplicity of light networks in comparison to other more sophisticated techniques. This makes it possible for health services to reach more people and boosts the likelihood of an early diagnosis.In essence, the successful development of machine learning models for skin cancer diagnosis, particularly the notable performance of light YOLO model YOLOv5n in this context, coupled with their compatibility with mobile applications, signifies a significant stride forward in healthcare accessibility and early detection efforts.

Table 3.1 :   Comparison of all models in the paper for original data, mAP50 score will be used instead of accuracy for YOLO models.

	Accuracy	Precision	Recall	F1-Score
KNN	0.67	0.38	0.28	0.30
SVM	0.71	0.43	0.27	0.31
Logistic Regression	0.70	0.44	0.28	0.32
Random Forest	0.71	0.47	0.25	0.28
Xception	0.61	0.37	0.37	0.33
VGG16	0.76	0.61	0.43	0.48
RestNet50	**0.81**	**0.70**	**0.59**	**0.64**

	Accuracy	Precision	Recall	F1-Score
YOLOv5	0.57	0.47	0.58	0.51
YOLOv5n	0.46	0.45	0.51	0.39

Table 2.14 in the paper compares various models using augmented data. For evaluating the performance of YOLO models, the metric used is mAP50 (mean Average Precision at IoU threshold 0.50) instead of the typical accuracy measure.

Table 2.14 :   Comparison of all models in the paper for augmented data, mAP50 score will be used instead of accuracy for YOLO models.

	Accuracy	Precision	Recall	F1-Score
KNN	0.72	0.72	0.72	0.71
SVM	0.53	0.53	0.53	0.53
Logistic Regression	0.51	0.51	0.51	0.51
Random Forest	0.78	0.78	0.78	0.78
Xception	0.71	0.71	0.71	0.70
VGG16	0.85	0.85	0.85	0.85
RestNet50	**0.92**	**0.93**	**0.92**	**0.92**
YOLOv5	0.854	0.8	0.824	0.81
YOLOv5n	0.82	0.75	0.78	0.77

4. REALISTIC CONSTRAINTS AND CONCLUSIONS

Skin cancer, including melanoma, is one of the most prevalent forms of cancer worldwide, with its early detection being crucial for effective treatment. Dermatologists play a pivotal role in diagnosing skin lesions, but the increasing demand for their expertise highlights the need for automated tools to aid in diagnosis. Leveraging advancements in deep learning, our project focuses on developing a skin lesion classification system capable of distinguishing between eight classes of skin lesions.

4.1 Practical Application of this Project

To ensure the reliability and diversity of our dataset, we used a renowned and reliable skin lesion data library to gather a comprehensive collection of skin lesion images. These images were meticulously annotated by experts, providing ground truth labels essential for model training and evaluation. The collaboration facilitated a robust dataset that encompasses a wide range of skin conditions and variations. Once validated, the skin lesion classification model can be seamlessly integrated into clinical workflows to assist dermatologists in real-time diagnosis. The model can serve as a valuable decision support tool, providing complementary information to assist in lesion assessment and treatment planning. The system contributes to improving patient outcomes and reducing healthcare burden by streamlining the diagnostic process and enabling early detection of malignant lesions.

4.2 Realistic Constraints

Although our skin lesion classification project shows promise in improving diagnostic accuracy and patient outcomes, some realistic limitations need to be considered. First of all limited access to certain classes of leasions or differences arising from dermoscope images not being taken under the same conditions may affect the generalization abilities of the model. Additionally, integration of AI-powered tools into existing clinical workflows may encounter

resistance or skepticism from healthcare professionals and patients alike, requiring effective communication, training, and user education initiatives to encourage acceptance and adoption.

4.2.1 Social, environmental and economic impact

Early detection of skin cancer holds significant potential for improving medical outcomes and saving lives. The widespread availability of technology for early detection could lead to cost reductions by eliminating the need for certain procedures, such as biopsies, making skin cancer screening more accessible. This accessibility is particularly beneficial for individuals who may not have easy access to traditional healthcare services. The impact extends beyond cost savings; early-stage treatments are often more effective and less expensive than advanced cancer treatments. By reducing the overall healthcare costs associated with late-stage cancer, early diagnosis can contribute to more sustainable and efficient healthcare systems. Moreover, timely treatment not only improves health outcomes but also plays a role in maintaining a productive workforce. Individuals who receive successful early treatment can return to work sooner, minimizing the economic impact associated with prolonged medical leaves and disability. This dual benefit of improving health and economic productivity underscores the importance of advancements in early detection technologies for skin cancer.

4.2.2 Cost analysis

The cost of this project, in which two engineers worked for 8 months, is approximately 480,000 TL, considering the salaries. In order to carry out the training process, the researchers needed to use a computer with sufficient GPU capacity. The cost of this computer is 50,000 TL. Additionally, various hardware expenses as well as costs related to software licenses for applications such as MATLAB and Python must also be taken into account. It should also be included that using Google Colab Pro pays for faster code execution.

4.2.3 Standards

The relevant standards for this project are outlined here.

- IEEE (Institute of Electrical and Electronics Engineers)
- BTK (Information and Communication Technologies Authority)

4.2.4 Health and safety concerns

This algorithm will be used for medical diagnosis. If used as the main diagnostic tool the false negative diagnosis may cause possibly fatal outcomes. On the other hand false positive results may cause unnecessary panic, loss of time due and economic loss due to the test that will be done afterwards. Additionally, the accuracy percentage achieved in various deep learning networks may not be sufficient to make a definitive diagnosis in detecting malignant lesions. Therefore, misdiagnosis of malignant cancer can lead to serious consequences. In particular, the accuracy, sensitivity, and specificity of deep learning algorithms must be rigorously evaluated before they can be used reliably in clinical practice. Misdiagnosis of malignant lesions can lead patients to experience unnecessary worry and anxiety, lead to additional medical tests, and even unnecessary surgical interventions. Therefore, prior to clinical application of deep learning algorithms, it is critical to robustly validate their reliability and safety. Therefore, before implementing the model in clinical settings, it is essential to validate its performance and integration with existing diagnostic workflows.

4.3 Future Work and Recommendations

To maximize the practical utility of the model, it is essential to integrate it seamlessly into clinical workflows. Developing user-friendly software or mobile applications that can assist dermatologists and general practitioners in real-time could facilitate widespread adoption. Ensuring the tool's compatibility with existing medical systems and databases will be crucial for smooth integration. Also, enhancing the model to provide real-time processing and instant feedback to clinicians and patients will be a significant advancement. This can be achieved by optimizing the algorithm for faster inference and implementing efficient computational techniques. Real-time processing can aid in immediate decision-making and improve the overall diagnostic process. Moreover, future work should focus on expanding the dataset with more diverse and comprehensive dermoscopic images. This includes collecting data from different demographics, skin types, and various stages of skin lesions. A larger and more varied dataset will improve the model's generalization capabilities and robustness.

REFERENCES

[1] **American Cancer Society,** *"Cancer Facts and Figures 2023". Atlanta: American Cancer Society; 2023.* [Online].Available: https://www.aimatmelanoma.org/facts-statistics/

[2] **H. Kittler, H. Pehamberger, K. Wolff, and M. Binder**, "Diagnostic accuracy of dermoscopy," *The Lancet. Oncology*, vol. 3, no. 3, pp. 159–165, 2002.

[3] **D. S. Charan, H. Nadipineni, S. Sahayam, and U. Jayaraman,** "Method to classify skin lesions using dermoscopic images," 2020.

[4] **T. J. Brinker *et al.*,** "Deep learning outperformed 136 of 157 dermatologists in a head-to-head dermoscopic melanoma image classification task," *European journal of cancer (Oxford, England: 1990), 113*, pp. 47–54, 2019.

[5] **B. Shetty, R. Fernandes, A. P. Rodrigues, *et al.*,** "Skin lesion classification of dermoscopic images using machine learning and convolutional neural network," *Sci. Rep.*, vol. 12, 2022, doi: 10.1038/s41598-022-22644-9.

[6] **A. Adebiyi, N. Abdalnabi, E. Hoffman, J. Hirner, E. Simoes, M. Becevic, and P. Rao,** "Accurate skin lesion classification using multimodal learning on the HAM10000 dataset," *medRxiv,* 2024. [Online]. Available: https://www.medrxiv.org/content/early/2024/05/31/2024.05.30.24308213. doi: 10.1101/2024.05.30.24308213.

[7] **M. Deif, R. E. Hammam, and A. Solyman,** "Skin Lesions Classification Based on Deep Learning Approach," *Journal of Clinical Engineering*, vol. 45, 2020. doi: 10.1097/JCE.0000000000000405.

[8] **B. Taşar, "**SkinCancerNet: Automated Classification of Skin Lesion Using Deep Transfer Learning Method," *Traitement Du Signal,* vol. 40, no. 1*,* pp. 285-295, 2023. Available: https://www.proquest.com/scholarly-journals/skincancernet-automated-classification-skin/docview/2807002674/se-2. doi: 10.18280/ts.400128.

[9] **A. Esteva, B. Kuprel, R. Novoa, et al.,** "Dermatologist-level classification of skin cancer with deep neural networks," *Nature*, vol. 542, pp. 115–118, 2017. [Online].

Available: https://www.nature.com/articles/nature21056. doi: 10.1038/nature21056.

[10] **A. Shahin, A. Kamal, and M. Elattar,** "Deep ensemble learning for skin lesion classification from dermoscopic images," in *Proc. IEEE Conf. Biomed Eng. Comput. Sci. (CIBEC)*, 2018, pp. 150-153, doi: 10.1109/CIBEC.2018.8641815.

[11] **M. Kassem, K. Hosny, and M. Fouad,** "Skin lesions classification into eight classes for ISIC 2019 using deep convolutional neural network and transfer learning," *IEEE Access*, vol. 8, pp. 1-1, 2020. doi: 10.1109/ACCESS.2020.3003890.

[12] **D. Moldovan,** "Transfer learning based method for two-step skin cancer images classification," in *Proc. of the 2019 E-Health and Bioengineering Conference (EHB)*, Nov. 2019, pp. 1–4. IEEE. doi: 10.1109/EHB47216.2019.8970067.

[13] **S. M. Thwin and H. S. Park**, "Skin lesion classification using a deep ensemble model," *Applied Sciences*, vol. 14, no. 35599, 2024. [Online]. Available: https://www.mdpi.com/2076-3417/14/13/5599. doi: 10.3390/app14135599.

[14] **P.Tschandl, N. C. Codella, B. N. Akay, G. Argenziano, R. P. Braun, H. Cabo, D. Gutman, et al.,** "Comparison of the accuracy of human readers versus machine learning algorithms for pigmented skin lesion classification: an open, web-based, international, diagnostic study," *Lancet Oncol.*, vol. 20, no. 7, pp. 938–947, 2019. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8237239/. doi: 10.1016/S1470-2045(19)30333-X

[15]]**H. A. Haenssle, C. Fink, R. Schneiderbauer, F. Toberer, T. Buhl, A. Blum, A. Kalloo, A. B. H. Hassen, L. Thomas, A. Enk, et al.,** "Man against machine: diagnostic performance of a deep learning convolutional neural network for dermoscopic melanoma recognition in comparison to 58 dermatologists," *Ann. Oncol.*, vol. 29, no. 8, pp. 1836-1842, Aug. 2018. doi: 10.1093/annonc/mdy166.

[16] **V. Ranjan, K. Chaurasia, and J. Singh,** "Multi-class skin lesion classification using intelligent techniques," in *Adv. Comput. and Intell. Technol.*, R. N. Shaw, S.

Das, M. Paprzycki, A. Ghosh, and M. Bianchini, Eds. vol. 958, Singapore: Springer, 2024, pp. 41. doi: 10.1007/978-981-97-1961-7_41.

[17] **M. M. Hossain, M. M. Hossain, M. B. Arefin, F. Akhtar, and J. Blake,** "Combining state-of-the-art pre-trained deep learning models: A noble approach for skin cancer detection using max voting ensemble," *Diagnostics*, vol. 14, no. 1, p. 89, 2024. [Online]. Available: https://pubmed.ncbi.nlm.nih.gov/38201399/. doi: 10.3390/diagnostics14010089.

[18] **N. Elyasi and M. Hosseini Moghadam,** "Classification of skin lesions by TDA alongside Xception neural network," *J. AI Data Min.*, vol. 10, no. 3, pp. 333-344, 2022, doi: 10.22044/jadm.2022.10948.2239.

[19] **M. J. Alam, M. S. Mohammad, M. A. F. Hossain, I. A. Showmik, M. S. Raihan, S. Ahmed, and T. I. Mahmud,** "S2C-DeLeNet: A parameter transfer based segmentation-classification integration for detecting skin cancer lesions from dermoscopic images," *Comput. Biol. Med.,* vol. 150, p. 106148, 2022. doi: 10.1016/j.compbiomed.2022.106148.

[20] **S. S. Mohammed and J. M. Al-Tuwaijari,** "Skin Disease Classification System Based on Machine Learning Technique: A Survey," in *IOP Conf. Ser.: Mater. Sci. Eng.*, vol. 1076, p. 012045, 2021. doi: 0.1088/1757-899X/1076/1/012045

[21] **M. Elshahawy, A. Elnemr, M. Oproescu, A.-G. Schiopu, A. Elgarayhi, M. Elmogy, and M. Sallah,** "Early melanoma detection based on a hybrid YOLOv5 and ResNet technique," *Diagnostics*, vol. 13, no. 2804, 2023. [Online]. Available: https://pubmed.ncbi.nlm.nih.gov/37685342/. doi: 10.3390/diagnostics13172804.

[22] **R. Xu,** "Image classification of skin cancer using deep neural networks with scaling laws," *International Journal of Computer Science and Information Technology*, vol. 3, no. 2, pp. 102-116, 2024. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9554840/. doi: 10.62051/ijcsit.v3n2.12.

[23] **A. A. Marghoob, R. P. Braun, and A. W. Kopf**, Eds., *Atlas of Dermoscopy*. London: Informa Healthcare, 2012.

[24] **M. Alsaidi, M. Jan, A. Altaher, H. Zhuang, and X. Zhu**, "Tackling the class imbalanced dermoscopic image classification using data augmentation and GAN," *Multimedia Tools and Applications*, vol. 83, 2023, doi: 10.1007/s11042-023-17067-1.

[25] **G. Guo, H. Wang, D. Bell, Y. Bi, and K. Greer**, "KNN model-based approach in classification," in OTM Confederated International Conf., 2019, pp. 986-996, [Online]. https://doi.org/10.1007/978-3-540-39964-3_62

[26] **I.-T. Yang and H. Prayogo,** "Efficient Reliability Analysis of Structures Using Symbiotic Organisms Search-Based Active Learning Support Vector Machine," Buildings, vol. 12, no. 4, p. 455, 2022, doi: 10.3390/buildings12040455.

[27] **A. Parmar, R. Katariya, and V. Patel**, "A review on Random Forest: An ensemble classifier," in *Proc. Int. Conf. Intell. Data Commun. Technol. Internet Things (ICICI)*, 2018, Lecture Notes on Data Engineering and Communications Technologies, vol. 26, J. Hemanth, X. Fernando, P. Lafata, and Z. Baig, Eds. Cham, Switzerland: Springer, 2019, pp. 758-763. doi: 10.1007/978-3-030-03146-6_86.

[28] **N. Otsu,** "A threshold selection method from gray-level histograms," in *IEEE Transactions on Systems, Man, and Cybernetics*, vol. 9, no. 1, pp. 62-66, Jan. 1979, doi: 10.1109/TSMC.1979.4310076.

[29] **R. Johr,** "Dermoscopy: Alternative melanocytic algorithms - The ABCD rule of dermatoscopy, Menzies scoring method, and 7-point checklist," *Clinics in Dermatology*, vol. 20, pp. 240-247, 2002, doi: 10.1016/S0738-081X(02)00236-5.

[30] **C. Barata, M. Ruela, M. Francisco, T. Mendonça and J. S. Marques,** "Two Systems for the Detection of Melanomas in Dermoscopy Images Using Texture and Color Features," in *IEEE Systems Journal*, vol. 8, no. 3, pp. 965-979, Sept. 2014, doi: 10.1109/JSYST.2013.2271540.

[31] **M. S. Kartal and Ö. Polat**, "Detection of Benign and Malignant Skin Cancer from Dermoscopic Images using Modified Deep Residual Learning Model," Artificial Intelligence Theory and Applications, vol. 2, no. 2, pp. 10-18, 2022.

[32] **Mathworks.** "What Is Deep Learning? | How It Works, Techniques & Applications." *Mathworks.com*, 2019, www.mathworks.com/discovery/deep-learning.html.

[33] **V. Bansal, "Transfer Learning in Convolution Neural Network," Medium. [Online]. Available: https://medium.com/@vishal025/transfer-learning-in-convolution-neural-network-b69504f1d052. Accessed: Aug. 10, 2024.**

[34] **I.Ritharson, K.Raimond, and A.Mary**,"Novel Deep Learning-based CNN Architecture for Rice Leaf Disease Detection."*Intelligent Decision Technologies,* 2023, doi: 10.13140/RG.2.2.20721.45924.

[35] **C. Chandini,** "CNN family and Business Applications of a Convolutional Neural Network," *Medium*, Sep. 5, 2023. [Online]. Available: https://medium.com/@aschandinip/cnn-family-and-business-applications-of-a-convolutional-neural-network-1349f606abe1.

[36] **P. Mahajan,** "Understanding ResNet Architecture," *Analytics Vidhya*, Sep. 21, 2020. [Online]. Available: https://medium.com/analytics-vidhya/understanding-resnet-architecture-869915cc2a98. Accessed: May 26, 2024.

[37] **M. Lagunas and E. Garces,** "Transfer Learning for Illustration Classification," in *Proc. Spanish Computer Graphics Conference*, 2018. doi: 10.2312/ceig.20171213.

[38] **J. Yu and H. Choi,** "YOLO MDE: Object Detection with Monocular Depth Estimation," *Electronics*, vol. 11, no. 1, p. 76, 2022. [Online]. Available:https://www.mdpi.com/2079-9292/11/1/76.doi: 10.3390/electronics11010076

CURRICULUM VITAE

PHOTO

Name Surname   : Dilge Dağ

Place and Date of Birth   : Istanbul/ 17.07.2000

E-Mail   : dilgedag@gmail.com

CURRICULUM VITAE

Name Surname   **: Rabia Batır**

Place and Date of Birth   **: Van/ 26.03.2000**

E-Mail   **: batir19@itu.edu.tr**