Arab American University

Faculty of Graduate Studies

A Validation Study of the Power and Effectiveness of Machine Learning in Mammogram Interpretation

By

Hanin Mohammad Ahmad Saadah

Supervisor

Dr. Amani Yousef Owda

Co-Supervisor

Dr. Majdi Owda

This thesis was submitted in partial fulfillment of the requirements for the Master's degree in Data Science and Business Analytics

June/ 2024

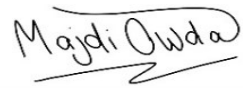
Thesis Approval

A Validation Study of the Power and Effectiveness of Machine Learning in Mammogram Interpretation

By

Hanin Mohammad Ahmad Saadah

This thesis was defended successfully on 29/06/2024 and approved by:

Committee members	Signature
1. Dr. Amani Yousef Owda: Supervisor	
2. Dr. Majdi Owda: Co-Supervisor	
3. Dr. Muath Sabha: Internal Examiner	
4. Dr. Radwan Tahboub: External Examiner	

Declaration

I declare that the thesis titled " A Validation Study of the Power and Effectiveness of Machine Learning in Mammogram Interpretation" is my work, has been completely done by me, does not contain work from other researchers, and has not been submitted for any other degree or scientific work except the reference is made.

Hanin Saadah

202112677

Hanin Mohammad Ahmad Saadah

19/09/2024

Dedication

I dedicate this thesis to my beloved family; my mother, father, sisters, brothers, nieces, and nephews. Their unconditional and unlimited support and encouragement made this achievement much easier and without them, no success would have been possible. To my friends who stood by me in difficult times. To my colleagues who supported me in my career path. I appreciate your love and support and will never forget it. Thank you!

Acknowledgment

Words cannot express my deep gratitude to my supervisors, Dr. Amani and Dr. Majdi for their time, availability, and support. I couldn't have undergone this journey without their knowledge and expertise. I would also like to thank them for their interest in my thesis, as their advice, valuable comments, and following up with me from the beginning have strengthened this thesis and helped me improve it significantly.

Abstract

Breast Cancer is widespread across the globe. It's the primary cause of death in cancer fatalities. According to the Palestinian Ministry of Health (MoH) annual report, it ranked as the third reported death of all reported cancer deaths in the West Bank. Breast cancer has many symptoms like breast pain, discharge, lumps, and masses. It also has risk factors that can be categorized into genetics and lifestyle which helps in developing the cancer. Moreover, mammogram screening is the most common technique to diagnose breast abnormalities, but there is a challenge in the lack of skilled experts able to interpret mammograms accurately. Machine Learning (ML) plays an important role in medical image processing particularly in early detection when the treatment is less expensive and available. This thesis presents two approaches to detecting breast abnormalities based on the Convolutional Neural Network (CNN). The first approach classified images as normal and abnormal, while the second classified them into Breast Imaging-Reporting and Data System scores (BI-RADS). Furthermore, six CNN models were implemented in both approaches, namely VGG16, VGG19, DenseNet121, ResNet50, Xception, and EfficientNetB7. The used dataset is a unique (first-hand) dataset collected from the Palestinian MoH. Based on the results, DenseNet121 outperformed other models in the first approach with 0.83 and 0.85 for testing accuracy and Area Under Curve (AUC) respectively. In contrast, Xception and EfficientNetB7 obtained the best results in the second approach with more than 90% accuracy. As a future work, the outperformed model can be integrated with other patient data like genetic information, medical history, and lifestyle factors to evaluate the risk of developing specific diseases. This would increase the survival rate and enable proactive

measures. Finally, a larger dataset of mammogram images should be collected to improve results and further generalize the models.

Table of Contents

List of Tables

List of Figures

List of Appendices

List of Abbreviations

CNN	Convolutional Neural Network
BI-RADS	Breast Imaging-Reporting and Data System
AUC	Area Under Curve
ML	Machine Learning
MoH	Ministry of Health
WHO	World Health Organization
MRI	Magnetic Resonance Imaging
CT	Computed Tomography
GP	General Practitioner
PCA	Principal Component Analysis
GLCM	Gray-level co-occurrence matrix
LBPs	Local Binary Patterns
DL	Deep learning
RCNN	Recurrent CNN
IG	Information Gain
CHI	Chi-squared test
FCBF	Correlation-Based Filter
AI	Artificial Intelligence
SVM	Support Vector Machine
RF	Random Forest
DT	Decision Trees
LR	Logistic Regression
KNN	K-Nearest Neighbors
NB	Naïve Bayes

DICOM	Digital Imaging and Communications in Medicine
MLO	Medio-lateral Oblique
CC	Cranio Caudal
DenseNet121	Densely Connected Convolutional Network 121
ResNet50	Residual Network 50
Xception	The Extreme Inception
VGG16	Visual Geometry Group 16
VGG19	Visual Geometry Group 19
ReLU	Rectified Linear Unit
TP	True Positive
TN	True Negative
FP	False Positive
FN	False Negative
ROC	Receiver Operating Characteristic
TPR	True Positive Rate
FPR	False Positive Rate

Chapter One: Introduction

1.1 Breast Cancer Introduction

Breast cancer is prevalent globally, and it is the primary cause of cancer-related fatalities in women, particularly impacting those in low- and middle-income communities (World Health Organization, 2023). In the West Bank, Palestine, breast cancer was the third reported death of all reported cancer deaths as shown in Figure (1.1), and it accounted for 30% of deaths among women of reproductive age (Ministry of Health, 2022). While breast cancer is typically diagnosed more frequently in women aged fifty years or older, there is a concerning increase in its occurrence among younger women. Currently, breast cancer ranks as the second leading cause of cancer-related deaths in women aged less than 40 globally (Zhu et al., 2023). Approximately 30% of breast cancer cases are preventable through modifiable risk factors like excess body weight, physical inactivity, and alcohol intake, and can be further reduced through mammography screening and advances in treatment (Giaquinto et al., 2022).

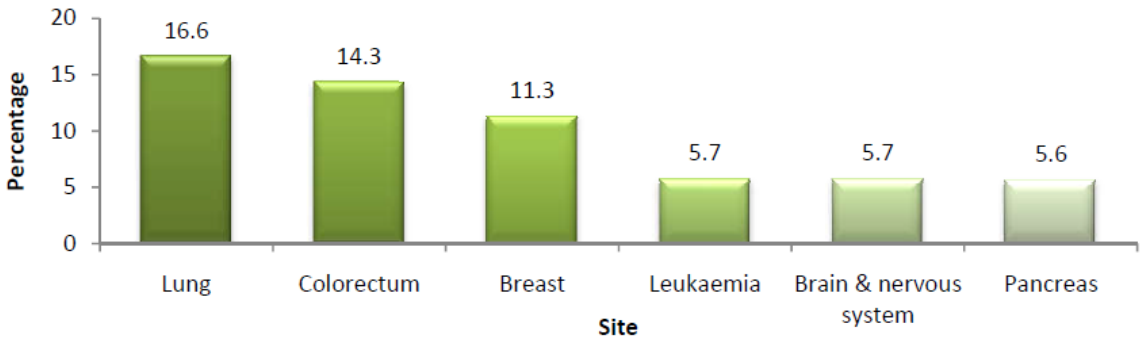


Figure 1.1: Proportional Distribution of the Most Reported Cancer Deaths of all Reported Cancer Deaths, West Bank, 2021 (Ministry of Health, 2022).

Moreover, breast cancer has many symptoms such as breast pain, discharge, or the presence of lumps (Kumari et al., 2022). It has also risk factors that can be divided into two groups, namely genetics, and lifestyle (Zhu et al., 2023). Furthermore, these risk

factors alongside other diseases like hypertension and diabetes help in developing breast cancer (İz & Tümer, 2016). In addition, there are various ways to screen breast abnormalities, but the most common is mammogram screening (World Health Organization, 2016). A mammogram image is a type of X-ray image that requires a specialist to interpret it and detect abnormalities (Geras et al., 2019). Additionally, the Breast Imaging-Reporting and Data System (BI-RADS) is a global standard scale for mammogram image classification according to the scores based on the results (Eberl et al., 2006). Besides, the shortage of specialists responsible for interpreting images and the increasing number of women seeking mammogram screening poses a challenge in providing this service with accuracy and high quality (AlWaheidi et al., 2020). Thus, delay in providing a proper treatment to control the cancer.

ML plays a pivotal role in the healthcare field, especially in medical imaging. Researchers developed various models and applications to detect and diagnose diseases like cancer at early stages when the treatment is available and affordable (World Health Organization, 2007). These models empower care providers to take a second opinion or decision which will improve the provided services at the clinical level. This will help in increasing the survival rates worldwide.

1.2 Problem Statement

Computer-aided detection systems based on ML can enhance the performance and accuracy of breast cancer screening. However, the Palestinian MoH faces a shortage of resources, including well-trained doctors to interpret mammograms. Moreover, a shortage of doctors means that interpreting the mammograms can take days or weeks, consequently delaying diagnosis. Therefore, there is a need to introduce a valid and accurate approach to reading mammograms using ML in the Palestinian health system.

From this problem, the idea of this thesis was formulated to introduce a solution to be integrated into the Palestinian health system to overcome existing challenges and limitations.

1.3 Research Objectives

The main objective of this thesis is to implement a ML model based on the Convolutional Neural Network (CNN) for the early detection of the changes and abnormalities in breast mammography images. In addition, there are other objectives to cover the following:

- Collecting a local dataset of mammogram images in addition to medical reports from the MoH Centers in West Bank, Palestine.
- Understanding the existing mammogram data in MoH centers, and identifying the characteristics of normal and abnormal mammograms.
- Implementing a ML model to interpret and classify mammography images.
- Helping the MoH to interpret their data using the implemented models.
- Using different CNN models in finding the optimal model and comparing results.
- Following different approaches to detect breast abnormalities in mammogram images.

1.4 Contribution

The main contribution to the knowledge in this thesis:

Manually collecting a first-hand dataset from the MoH in the West Bank, Palestine.

Implementing different ML models to detect breast abnormalities in mammographies.

Classifying mammogram images into BI-RADS according to the standard BI-RADS scale.

Extracting Insightful Analysis from the medical history data.

1.5 Thesis Structure

The rest of the thesis is organized as illustrated in Figure (1.2).

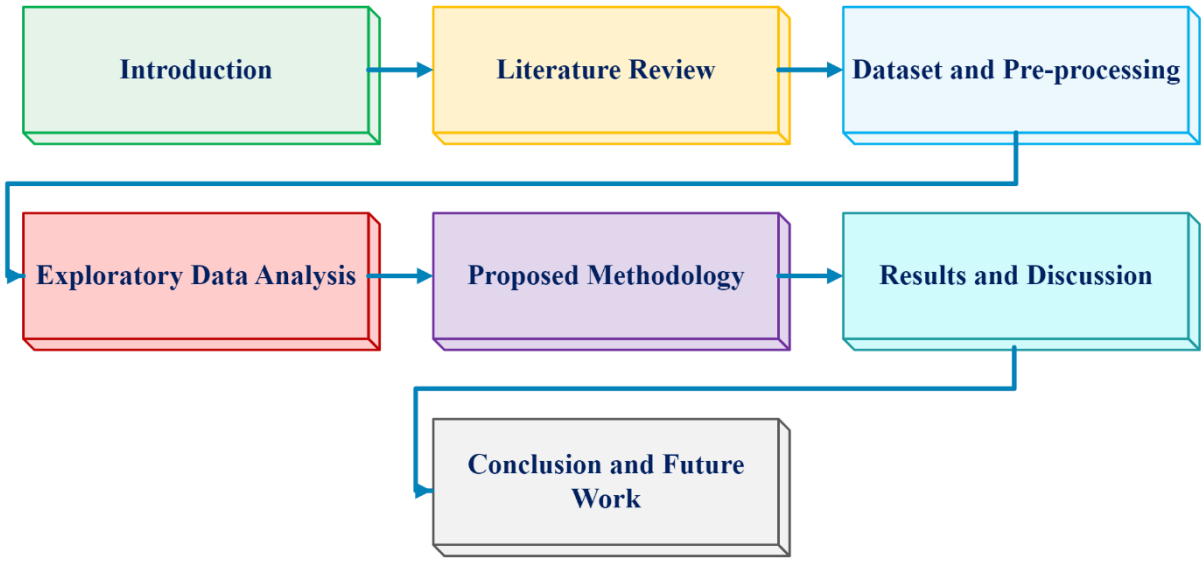


Figure 1.2: Thesis Structure.

1.6 Publications

A paper formulated from this thesis focused on the first approach was submitted to the Indonesian Journal of Electrical Engineering and Computer Science (ijeecs) Scopus Q2. The paper status is under review and it is titled "**Convolutional Neural Networks Breast Cancer Classification Using Palestinian Mammogram Dataset.**" In the next chapter, a comprehensive literature review related to the breast cancer and ML approaches used will be presented.

Chapter Two: Literature Review

2.1 Introduction

This chapter provides an overview of the previous studies in the field of breast cancer. This includes an illustration of the breast anatomy, symptoms, and associated risk factors, early detection, and diagnosis. In addition, the following sections present common techniques for breast screening, the global mammogram evaluation scale, and how it's interpreted. Furthermore, later sections talk about ML applications in medical imaging and interpretation. Finally, illustrates studies conducted to detect breast cancer using ML and deep learning algorithms.

2.2 Breast Cancer Overview

2.2.1 Breast Anatomy

Throughout a woman's life, the breast undergoes numerous transformations, spanning from infancy, adolescence, and motherhood including pregnancy and breastfeeding, all the way to the menopausal stage. It is essential to have a thorough grasp of breast anatomy and its representation in imaging studies to ensure comprehensive knowledge in this field.

The breast is a glandular organ located on the chest wall of females. Though not anatomically categorized as part of the body's upper limb, it resides in the pectoral region, with its blood supply and lymphatic drainage primarily directed toward the armpit. It's a specialized accessory gland of the skin, that secretes milk and exists in both males and females (Snell, 2011). In males and immature females, their structure is similar.

The breasts are composed of glandular and supportive fibrous tissue, situated within a fatty matrix and accompanied by blood vessels, lymphatics, and nerves. The mammary glands reside in the subcutaneous tissue, overlaying the pectoralis major and minor muscles. The highest point of the breast features the nipple, encircled by a pigmented circular region of skin known as the areola (Moore et al., 2018).

During puberty in females, the breasts undergo a gradual enlargement, primarily influenced by ovarian hormones. The ducts elongate, but the increased size of the glands mainly results from the accumulation of fat. The base of each breast extends from the 2nd to 6th rib and from the lateral margin of the sternum to the midaxillary line. While most of the gland is located in the superficial fascia, a smaller portion, known as the axillary tail, extends upward and laterally. It pierces the deep fascia at the lower border of the pectoralis major muscle, extending into the axilla. Each breast is composed of 15 to 20 lobes radiating outward from the nipple, and each lobe's main duct opens independently at the summit of the nipple, featuring a dilated ampulla just before its termination. Surrounding the base of the nipple is the areola, which exhibits tiny tubercles resulting from the underlying areolar glands. Fibrous septa, acting as suspensory ligaments, separate the lobes of the gland. Behind the breasts lies a space filled with loose connective tissue known as the retro mammary space (Snell, 2011).

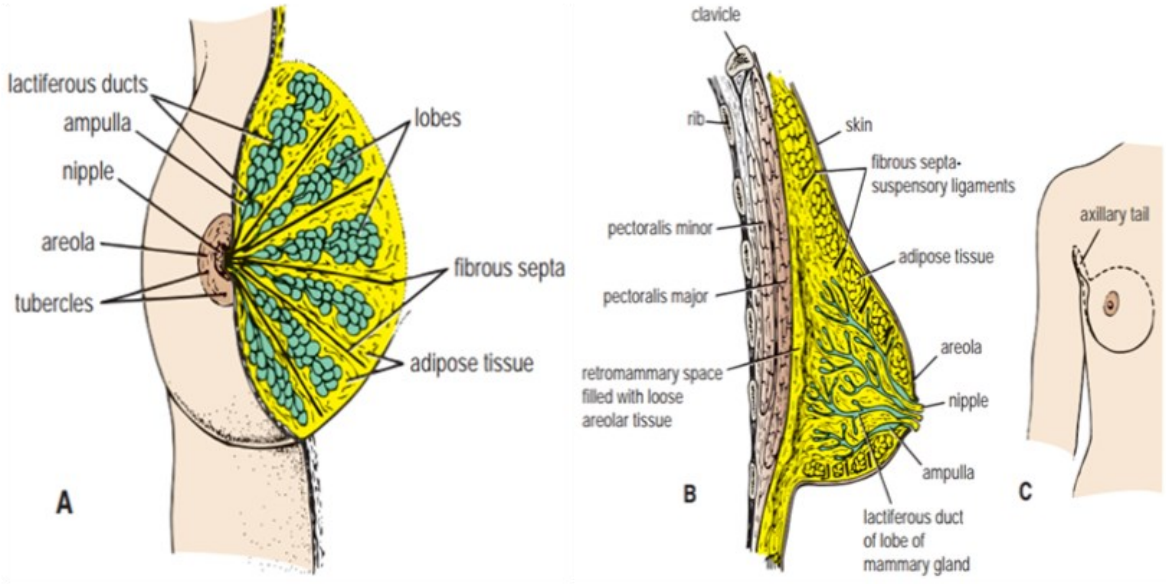


Figure 2.1: Front, and side view of the breast (Snell, 2011)

The mammary gland comprises two primary components. Firstly, the parenchyma, responsible for milk production as shown in Figure (2.1), is made up of lobes divided into 15 to 20 lobules. These lobules contain grapelike clusters of alveoli, which are hollow spheres lined with secretory cells. Within the parenchyma, lactiferous ducts drain each lobule and open at the nipple. Just beneath the areola, these ducts have a small, dilated portion called the lactiferous sinus, which stores a small amount of milk in lactating females. Secondly, the stroma acts as the fibrous framework of the gland, separating the lobules and supporting the lobes. This stroma attaches to the overlying dermis of the skin through suspensory (Cooper) ligaments, which are particularly sturdy on the superior surface of the breast (Gilroy, 2021).

The most common form of breast cancer, invasive ductal carcinoma, originates from the lining of the lactiferous ducts. It primarily spreads through lymphatic channels, often affecting axillary nodes from the supralateral quadrant, but it can also reach supraclavicular nodes, the opposite breast, and the abdomen. Lymphatic obstruction leads to edema, while fibrosis (shortening) of suspensory ligaments can cause a pitted

skin appearance. Additionally, breast cancer can metastasize to the vertebrae, cranium, and brain through venous communication with the azygos system and vertebral venous plexus (Gilroy, 2021).

2.2.2 Symptoms and Risk Factors

The initial indication of breast cancer is the presence of a novel lump or mass within the breast that can be detected through touch. Such lumps are generally rigid, painless, and exhibit irregular borders, making them more potential to be cancerous. However, there are instances where tumors may be soft, tender, and have rounded edges (Cechanoviciute & Cechanovičienė, 2021; Kumari et al., 2022). There are many symptoms associated with breast cancer, some of which are physically obvious or tangible, while others may be subtle or intangible. Some obvious symptoms that can be detected in the breast are changes in the breast shape or size or color, pain or swelling in the breast or nipple, nipple inversion turning inward or discharge other than milk, and redness or thickness in the underarm area (Kumari et al., 2022).

Women who are diagnosed with breast cancer may have other symptoms that are not related to the breast or underarm area. Many of these symptoms may occur at the same time. The cancer clinicians called this case symptom clusters (Kwekkeboom, 2016). A study in Jordan showed that there are five main symptom clusters among breast cancer women in Jordan. The clusters are fatigue, pain, treatment side effects, psychological, nausea, and vomiting (Al Qadire et al., 2021).

Breast cancer has many associated risk factors that affect increased risk (Martín-Payo et al., 2023). A cross-sectional study conducted in Turkey identified multiple births, a short breastfeeding period, overweight, low socioeconomic level, and low level of education

as the most significant factors (Karadag Arli et al., 2019). In (Zhu et al., 2023), risk factors were categorized into two categories, Lifestyle and Genetic. Lifestyle included physical fitness, smoking, and alcohol consumption, social and economic situation, and body mass index. BReast CAncer gene 1 (BRCA1) and BReast CAncer gene 2 (BRCA2) the most common susceptibility genes were in the genetic category. The study in (İz & Tümer, 2016), found that there is a significantly strong correlation between developing breast cancer and older ages, early menarche, late menopause, and family history.

It's important to note that having these symptoms or risk factors does not necessarily mean that the woman has or will develop breast cancer. However, any concerning symptom or risk factor should be reported to a healthcare provider for further screening and investigation.

2.2.3 Early Detection and Diagnosis

Early detection is the cornerstone of controlling breast cancer and improving outcomes and survival. World Health Organization (WHO) defined the early diagnosis of cancers as the detection of initial phases in women with symptoms, as this facilitates straightforward and cost-effective treatment, leading to elevated rates of recovery (Ginsburg et al., 2020). Therefore, the WHO advised care providers to perform a comprehensive situational analysis to facilitate planning, encourage early diagnosis, deliberate mammography screening, establish a nationwide breast cancer control program, and conduct routine monitoring and evaluation (World Health Organization, 2016). Pan American Health Organization (PAHO) considered raising awareness through programs and education about breast cancer as the most important key element that improves outcomes, and women should be empowered to access cancer services

timely (Pan American Health Organization, 2016). When breast cancer is detected early, the treatment of breast cancer has the potential to be very successful, with survival rates of 90% or greater being achievable. In addition, radiation therapy can prevent the necessity of a mastectomy for women (WHO, 2023).

2.2.4 Breast Cancer Screening

Screening refers to the identification of high-risk people who appear healthy and do not show symptoms but have a high likelihood of having early-stage disease that cannot be detected through clinical examination (World Health Organization, 2016). Currently, there are multiple breast screening techniques available, such as mammography, ultrasound imaging, Positron-Emission Tomography (PET), computed tomography (CT), magnetic resonance imaging (MRI), and microwave imaging (Bagchi et al., 2020). By far, studies have proven that mammograms are the most sufficient evidence and effective in reducing mortality (Nazzal et al., 2018; World Health Organization, 2016). Mammography is a recommended initial screening imaging method that uses a breast-specific X-ray imaging modality to produce images of the breast in various positions, helping to detect abnormalities and lesions (Geras et al., 2019).

Improved survival rates and reduced mortality can be achieved by conducting breast cancer screening in a timely manner. A cross-sectional study was carried out in Palestine in 2016 to assess mammogram screening among women at risk. The study revealed that 50% of women had undergone at least one mammogram, but only 21% had received timely mammograms (Nazzal et al., 2018). According to the 2021 health annual report released by the Palestinian MoH, 39.6% of all cases examined across the governorates, totalling 5864, were found to be abnormal, with a total of 2322 cases identified as such (Ministry of Health, 2022). Breast cancer mortality can be reduced by up to 40% if 70%

of target individuals participate in preventive mammography screening programs (Ciuba et al., 2022).

2.2.5 Breast Imaging Reporting and Data System (BI-RADS)

The Breast Imaging Reporting and Data System (BI-RADS) was introduced by the American College of Radiology (ACR) in 1993 (Spak et al., 2017). It's a standard scale for mammogram reporting worldwide, aimed at enhancing communication among healthcare providers, minimizing ambiguity surrounding mammogram results, facilitating case management, and aiding in the monitoring of outcomes (Eberl et al., 2006).

BI-RADS system is commonly used in mammograms, MRIs, and ultrasounds. It employs a numeric scale that ranges from 0 to 6, with each category denoting distinct levels of suspicion concerning breast cancer. Table (2.1) shows different BI-RADS categories and the assigned description (Banerjee et al., 2019; Eghtedari et al., 2021; Falconí et al., 2020; Obenauer et al., 2005; Prabhala et al., 2022).

Table 2.1: BI-RADS Categories (Scores).

Category	***Findings/Diagnosis***	***Management***	***Likelihood of Developing BC***
0	Incomplete assessment	Additional imaging is required	-
1	Negative or normal findings	Routine follow-up every year	0%
2	Benign	Routine follow-up every year	0%
3	Probably benign	Routine follow-up at short intervals (every 6 months)	<=2%
4	Suspicious findings for malignancy.	Biopsy should be considered.	>2% to <=95%
5	Highly suggestive of malignancy	-The doctor's decision and appropriate action should be	>=95%

		taken. - Biopsy is required	
6	Proven malignancy	Appropriate therapy/surgery	100%

2.2.6 Mammography Interpretation by Domain Experts

A significant volume of mammograms is gathered through an extensive mammography screening program and necessitates evaluation from proficient radiologists, who are qualified but overloaded with excessive workloads (Tsai et al., 2022). An important challenge in the field of mammography is the shortage of skilled domain experts capable of accurately interpreting mammograms. This scarcity of radiologists and clinicians specialized in breast imaging can lead to delays in diagnoses, increased workloads for existing experts, and potential errors in the interpretation process. The shortage can be attributed to various factors.

One factor might be inadequate training; for instance, in the West Bank, the majority of the mammography centres are run by General Practitioners (GPs) who, despite their training, often lack confidence in mammogram interpretation due to insufficient training. Another factor is the uneven distribution of GPs across facilities, where some GPs might be assigned to multiple centers, leading to delays in mammogram interpretation (AlWaheidi et al., 2020).

By addressing the shortage of domain experts and implementing strategies to improve interpretation processes, healthcare systems can provide more timely and accurate diagnoses, ultimately improving patient outcomes in the realm of mammography.

Therefore, there is a need to develop AI models to assist radiologists with mammographic interpretation, and AI model development requires interdisciplinary research that integrates medical science and engineering (Bi et al., 2019).

2.3 Machine Learning in Healthcare

2.3.1 Machine Learning and Early Detection

Machine Learning plays a significant role in early detection across various fields, including healthcare (Kurian & Jyothi, 2022). It involves training computer algorithms to analyze large amounts of data, identify patterns, and make predictions (Tahmooresi et al., 2018). In the context of early detection, ML algorithms can be trained on diverse datasets to recognize subtle signs or patterns that may indicate the presence of a particular disease or condition (Kopitar et al., 2020).

By utilizing machine learning techniques, medical professionals can develop models that help detect diseases at their early stages, when intervention and treatment are most effective (World Health Organization, 2007). For example, in breast cancer detection, machine learning algorithms can analyze medical imaging data, such as mammograms or MRIs, to identify suspicious lesions or abnormalities that may require further investigation (Tahmooresi et al., 2018)

Machine learning algorithms can also be used in combination with other patient data, such as genetic information, lifestyle factors, and medical history, to assess an individual's risk of developing certain diseases (Kamboj Akanshaand et al., 2021; Tao et al., 2023). This enables proactive measures, such as preventive screenings or lifestyle modifications, to be taken for individuals at higher risk.

The advantage of machine learning in early detection lies in its ability to process vast amounts of data quickly and efficiently, potentially leading to more accurate and timely diagnoses (Benning et al., 2022). However, it is important to note that machine learning models should be validated and integrated into clinical workflows with proper

evaluation and oversight by medical professionals to ensure their reliability and safety (J. Zhang, 2022).

2.3.2 Image Processing in Medical Imaging

Feature extraction and feature selection are methodologies employed in machine learning to pre-process and convert the original input data into a more informative and manageable format before model training (Zebari et al., 2020). In addition, there are widely employed methods for reducing the dimensionality of data in various applications (Bommert et al., 2020). These techniques serve the purpose of optimizing the learning process and elevating the performance of machine learning algorithms.

2.3.2.1 Feature Extraction

Feature extraction involves generating new features that rely on the initial input feature set to reduce the high dimensionality of the feature vector (Zebari et al., 2020). The primary objective of feature extraction techniques for mammogram images is to capture pertinent information relevant to breast tissue attributes and possible irregularities. These techniques play a crucial role in the early detection and diagnosis of breast cancer. Encouraging experimental outcomes has demonstrated the effectiveness of deep learning in detecting breast cancer through mammogram images. The results in (Charan et al., 2018) further endorse the adoption of advanced deep learning-based techniques for feature extraction and classification in diverse medical imaging applications, particularly for breast cancer detection.

Different feature extraction techniques have been used by researchers to extract features from images. Principal component analysis (PCA) is one of the commonly used methods. it transforms the feature variables, which may have correlations, into a fresh

set of linearly uncorrelated variables known as principal components (Yang & Xu, 2019). It is usually used for dimensionality reduction or data visualization (Mohanty et al., 2020). Several studies implemented PCA for dimensionality reduction purposes such as studies in (Dhahri et al., 2019; Odhiambo Omuya et al., 2021; Samee et al., 2022; Sawssen et al., 2021).

The Gray-level co-occurrence matrix (GLCM) is a texture evaluation technique that analyzes the spatial relationship of pixels. It quantifies the occurrence of pixel pairs with specific values and a defined spatial relationship in an image. The spatial relationship typically refers to the pixel of interest and its immediate right neighbor (horizontally adjacent). Statistical measures are then derived from the calculated matrix to characterize the texture (Pratiwi et al., 2015). Many studies used this method to extract features (Djunaidi et al., 2021; Kamil & Jassam, 2020).

The study in (Fanizzi et al., 2019) used GLCM in addition to the wavelet transform technique which enables the division of the original image into various sub-bands of different frequencies. Other texture analysis-based methods like local binary patterns (LBPs) are also used to detect breast masses by investigating the spatial patterns and fluctuations in pixel intensities found within mammograms (Mohamed Abdel-Nasser & Puig, 2016; Ponraj et al., 2017).

Deep learning (DL) architectures also can automatically extract high-level features without the need for manual feature engineering, which sets them apart from traditional machine learning algorithms (Çayir et al., 2018). Convolutional Neural Network (CNN) is one of the DL methods. A convolution layer comprises three operations: convolution, activation, and pooling. In a CNN architecture, multiple convolution layers can be

present, with each layer progressively extracting higher-level features compared to the preceding layers (Zeiler & Fergus, 2013). In (Hoque et al., 2021), a Recurrent CNN (RCNN) was utilized to extract actual blood vessels from regions of interest within retinal images.

2.3.2.2 Feature Selection

Feature selection involves the act of selecting a compact subset of features from the original feature set without making any modifications, thereby preserving their original interpretation. This selected subset is then carefully evaluated and verified in alignment with the desired analysis goal (Jovic et al., 2015). The process of Feature Selection involves removing irrelevant and/or redundant attributes from a dataset to prevent over-fitting, simplify the learning process, or enhance the quality of outcomes (Bommert et al., 2020). Feature Selection has proven its efficacy in medical applications, playing a vital role not only in reducing dimensionality but also in enhancing our comprehension of disease causation. Its successful implementation has enabled us to gain valuable insights into the underlying factors contributing to the development and progression of diseases (Remeseiro & Bolon-Canedo, 2019). Feature Selection techniques are classified into three groups: Filter, Wrapper, and Embedded (Mafarja et al., 2020; Venkatesh & Anuradha, 2019). Filter technique is a feature ranking approach that assigns a score to each feature based on its data characteristics, independent of any specific model. This method doesn't rely on classification algorithms to evaluate features. The highest-scoring features or those surpassing a defined threshold are chosen for selection (Bommert et al., 2022). The study in (Cherrington et al., 2019) reviewed and concluded the most common ranking methods for filtering and grouped them into two groups: univariate and multivariate. The Univariate included the Information Gain

(IG), Chi-squared test (CHI), Fisher score, and v-score ranking. The multivariate methods are Fast Correlation-Based Filter (FCBF), ReliefF, and minimum redundancy maximum relevance mRMR. Filter methods in feature selection face challenges in determining the appropriate threshold to distinguish between important and redundant features, leading to inconsistent results and reliance on domain experts. This issue not only impacts the quality of the final feature set but also introduces inefficiencies when manual verification is needed (Cherrington et al., 2019). The wrapper technique utilizes a classification algorithm to assess feature subsets generated by a search algorithm. Although it often yields better outcomes, it is crucial to exercise caution to prevent overfitting, as the classifier within the wrapper procedure evaluates solutions based on their performance on the training data (H. Das et al., 2022; González et al., 2019). Wrapper techniques have three types, classified based on their feature selection logic, including Sequential Forward Selection (SFS), Backward Elimination (BE), and Recursive Feature Elimination (RFE) (Srinivas et al., 2023). However, these wrapper methods are typically considered computationally expensive due to the computational cost involved in training and testing (Chen et al., 2020). Lastly, the Embedded technique is a combination of filter and wrapper (Ghosh et al., 2020). This technique shares similarities with wrapper methods as they also associate feature selection with classification algorithms. However, in the embedded technique, this link is stronger compared to wrapper methods. Embedded technique utilizes classification algorithms that possess inherent capabilities for feature selection (Liu et al., 2019). Regularization models, extensively employed and frequently proposed by researchers, are among the most widely used embedded methods (Lu, 2019; H. Zhang et al., 2020). Embedded

approaches often offer a favorable balance between achieving optimal predictive performance and managing computational costs effectively (Pes, 2020).

2.3.3 Machine Learning in Mammography Interpretation

2.3.3.1 Machine Learning

Machine learning is a computer science subfield focused on pattern recognition and computational learning theory. It enables computers to learn and improve without explicit programming (Ongsulee, 2017). Machine learning has different types of methods: supervised learning, unsupervised learning, Semi-supervised learning, and Reinforcement learning.

Machine learning has emerged as a powerful tool in mammography interpretation, revolutionizing the field of breast cancer screening and diagnosis (Yoon & Kim, 2021). By leveraging the capabilities of machine learning algorithms, radiologists can improve the accuracy and efficiency of mammogram analysis, leading to earlier detection and improved patient outcomes (Wan et al., 2022).

Currently, a lot of computer-aided detection systems rely on machine learning without the need for instructions to detect breast cancer (Batchu et al., 2021).

While machine learning has demonstrated its efficiency, the prevailing approach among researchers is deep learning, which offers more potent techniques, particularly in the domain of medical imaging, notably in mammography. Deep learning will be the focal point of discussion in the forthcoming chapter.

Machine learning is a sub-section of artificial intelligence (AI) focusing on developing and training models to enable computers to make accurate data-driven decisions and predictions (Sarker, 2021). Moreover, machine learning has shown advanced results in

processing medical images, specifically in diagnosis tasks (Sistaninejhad et al., 2023). It has been heavily utilized in recent years by researchers to analyze medical images and potentially improve treatment outcomes in the healthcare field. Furthermore, machine learning algorithms play an important role in breast cancer detection. They can accurately analyze images to detect tumors, density, and subtle breast changes that occur in tissues and classify them into benign or cancerous (Manjunathan et al., 2023). Additionally, these algorithms can analyze high-complexity images thus providing high consistency and accuracy interpretations (Hossin et al., 2023; Sasirekha et al., 2023). Detecting breast cancer using machine algorithms can uncover patterns and extract insights (Jain, 2022). In general, these algorithms are meant to assist clinicians in planning the treatment process by providing them with insights to make informative decisions to improve the survival rate and decrease the mortality rate.

Machine learning has common approaches employed for detection and recognition which are supervised learning, unsupervised learning, and reinforcement learning and these approaches are used in different tasks depending on the input and output type (Nassif et al., 2019). Supervised learning requires a labeled training dataset to build models to make predictions (Narula, 2023). This approach reduces time, is flexible, optimizes workflows, discovers patterns, and provides high predictive and diagnostic accuracy. However, model complexity could be a limitation in addition to the risk of overfitting, interpretation challenges, labeled data collection challenges, and the lack of standard methods for labeling data (Biswas et al., 2021; P. Das & Mazumder, 2023; Nakayama et al., 2022; Tiwari, 2022). Unlike supervised learning, unsupervised learning doesn’t require a labeled dataset and the model trains itself by finding hidden structures and identifying features in the data (Pattanayak, 2017). Unsupervised learning

is flexible, can detect anomalies, reduces dimensionality, has a low computational cost, and the learning process is structured without labels because it's independent of data (Bodepudi, 2021; Montavon et al., 2022). On the other hand, this type of learning provides less accurate predictions, has interpretation challenges, the potential for false positives, and requires extra features to enhance performance (Abukmeil et al., 2020; Palo et al., 2022). Likewise, reinforcement learning uses sequential decisions based on interactions with the environment to maximize rewards (Epperlein et al., 2022). Although this approach is adaptable to dynamic environments and sensitive to hyperparameters, it suffers from high computational complexity and blind spots (Jones & Headley, 2022; A. R. Mahmood et al., 2018).

Support Vector Machine (SVM) is one of the most commonly used algorithms for diagnosing breast cancer. SVM showed a high-accuracy prediction when employed on the Wisconsin dataset in addition to other classification algorithms like decision trees (DT), logistic regression (LR), and random forest (RF) (Maanav et al., 2023). Moreover, multiclass SVM outperformed decision trees in terms of accuracy in detecting breast lesions when implemented on a dataset of mammogram images with a lot of noise in the background (Wajeed et al., 2023). A study proved that the accuracy and performance of SVM can be significantly improved by using a radial kernel with 5-fold cross-validation (Girik Allo et al., 2023). Similarly, decision tree algorithms are widely used for diagnosing breast tumors and abnormalities. Researchers found that combining decision trees for feature selection with SVM results in obtaining better accuracy (Alfian et al., 2022). Decision trees provide high-accuracy outcomes when utilized in classification tasks compared to other ML algorithms (Bokhare & Jha, 2023). Feature selection can

significantly influence decision tree accuracy and get effective results (Ebrahim et al., 2023).

K-Nearest Neighbors (KNN) has shown promising results for detecting breast cancer in recent years. In (Shafique et al., 2023), KNN was implemented with many feature selection methods such as chi-square and principal component analysis (PCA) to increase the accuracy. This algorithm can efficiently classify patients into malignant and benign which leads to enhancing the predictive performance in clinical use (Wassim et al., 2022). KNN can also offer high accuracy in diagnosing breast cancer in real-world applications and outperform other traditional algorithms (Wassim et al., 2022). Among ML algorithms, random forest (RF) also has an important implementation in this field. It can be used for recognizing patterns of cancer in the breast and giving effective predictions (Saravanakumar & Kannan, 2023). Furthermore, this algorithm can be combined with linear discriminant analysis to enhance the diagnostic systems (Adebiyi et al., 2022). Additionally, grid search is one of the ways to improve the performance when optimized with the random forest (Vinutha et al., 2022).

Similarly, logistic regression (LR) and Naïve Bayes (NB) have shown an improvement in detecting breast cancer in real-world cases. Studies showed that logistic regression combined with other ML algorithms has a high capability in predicting the disease (Poornajaf & Yosefi, 2023). Moreover, a hybrid discriminant logistic model along with a feature selection technique presented in (Tahir Iqbal et al., 2022) has outperformed SVM and Naïve Bayes in reducing computational time. In like manner, logistic regression produced the best accuracy in classifying cancer when it was employed with LASSO feature selection (Mohd Ali et al., 2020). On the other hand, Naïve Bayes plays a crucial role in breast cancer prediction applications. It can be implemented with KNN

to enhance the results of classification (Ceylan, 2020). Bayesian networks in general can perform an accuracy of 90% and above if it was utilized along with feature selection and extraction techniques (Ibeni et al., 2019). A study was conducted to compare LR with NB and they found that LR with a gradient ascent algorithm can achieve better results than NB, and both LR and NB perform similarly if implemented under the Gaussian NB assumption (Kumar Bhowmik, 2015).

The performance of machine learning algorithms used for diagnosing breast cancer is important to make sure that it can be reliable when implemented. For this reason, researchers have employed several performance metrics to evaluate machine learning models. Accuracy, Recall, f1-score, specificity, sensitivity, and area under the curve (AUC) are commonly used metrics for evaluation. The researchers in (Singh et al., 2023) used accuracy, AUC, precision, and recall to evaluate their presented model for breast cancer which was based on a multi-objective hyperparameter optimization. The obtained results of 94.4%, 98.16%, 95.77%, and 99.29% for accuracy, AUC, precision, and recall respectively (Singh et al., 2023). Another study resulted in 97.19% accuracy and 97.1% AUC when they evaluated their random forest model (T. Wang & Fan, 2022). Moreover, accuracy and AUC were used to evaluate the performance of a deep learning model and the results were 97.19% and 99.61% for accuracy and AUC in order (Prastyo et al., 2020). Furthermore, the rotation forest model for detecting lung cancer was evaluated using AUC, precision, accuracy, and recall and the results were 99.3% for the AUC, and 97.1% for the rest (Dritsas & Trigka, 2022).

Although machine learning models have proven their efficiency in detecting breast cancer, researchers face some challenges and limitations when using these models. The lack of standardized methods for the data collection process is one of the challenges

(Manjunathan et al., 2023). Furthermore, the data required for deep learning models to obtain a high accuracy is limited. To address this issue, researchers in (Zakareya et al., 2023) suggested using granular computing and utilizing learnable activation functions. In general, limited annotated data, data quality, risk of overfitting, inaccurate classification, model complexity, and results interpretability are the most common challenges in this manner.

2.3.3.2 Deep Learning

Deep learning, also referred to as deep structured learning, hierarchical learning, or deep machine learning, is a field that explores artificial neural networks and associated machine learning algorithms with multiple hidden layers (Ongsulee, 2017). Deep learning is experiencing a surge in popularity as it finds widespread application in three fundamental areas: detection, prediction, and generation.

Different studies and models were developed to detect breast cancer using deep learning techniques. The researchers in (Yala et al., 2019) developed three models. The first one used logistic regression embedded with the Tyrer-Cuzick risk assessment tool. The second model used the deep convolutional neural network (CNN), ResNet18 in particular, along with PyTorch. The last model was a hybrid deep learning model that merged the previous two models. They evaluated the models using Area Under Curve (AUC). The hybrid model outperformed all models with an AUC of 0.70 and a 95% confidence interval to be between 0.66-0.75. The study in (Shen et al., 2019) proposed another approach of training which is end-to-end and used Resnet50 and VGG16 as classifiers. The result in this study was improved to an AUC of 0.98. Other researchers

used mass lesion mammography to detect breast cancer using CNN algorithms: AlexNet, VGG, ResNet, GoogLeNet, and Inception (Tsochatzidis et al., 2019).

Deep learning offers a valuable approach for classifying mammography based on the BI-RADS categories. Distinguishing between scattered density and heterogeneously dense categories can pose challenges for physicians. In a study (Mohamed, Berg, et al., 2018), a Convolutional Neural Network (CNN) model was developed to effectively differentiate between these two classes by leveraging breast density information. The researchers utilized a substantial mammography database and achieved promising results, with an AUC of 0.98, after excluding potentially inaccurately labeled images. In another study (Lehman et al., 2019), a CNN model was introduced, leveraging both breast density and BI-RADS information. The researchers employed the ResNet-18 architecture implemented in PyTorch and actively engaged radiologists in evaluating the model's performance. The results were deemed satisfactory by the radiologists, suggesting its potential suitability for clinical practice.

In the majority of cases, mammography interpreters strive to identify breast microcalcifications alongside areas of density. In a study (J. Wang et al., 2016), researchers successfully created a deep-learning model specifically designed for detecting breast masses and microcalcifications. The model achieved an accuracy of 87.3%, demonstrating its effectiveness in this context. Ribli et al. (Ribli et al., 2018) employed the Faster R-CNN framework in their model and integrated a dedicated layer known as the Region Proposal Network (RPN) to detect objects in mammograms and classify them as malignant, benign, or normal. The model demonstrated excellent performance with an AUC of 0.95.

Mammography screening often encounters a high prevalence of false positives, which subsequently leads to elevated recall rates. Consequently, this increases the likelihood of unnecessary medical procedures for patients and incurs additional costs. In this regard, the study (Aboutalib et al., 2018) presented a CNN model with an end-to-end approach based on the AlexNet framework. The model can differentiate recalled-benign images from both malignant cases and negative instances. Another study in (T. Mahmood et al., 2021) aimed to mitigate false positives by creating a hybrid model that combined CNN (ConvNet) with Support Vector Machine (SVM). The model achieved an impressive accuracy of 97.8% and an AUC (Area Under the Curve) of 91.4%. Mahmood et al. (T. Mahmood et al., 2022) also employed ConvNet and incorporated five distinct DCNN architectures (ResNet152V2, InceptionResNetV2, InceptionV3, EfficientNetB5, and VGGNet19) in their model. This approach facilitated the diagnosis and classification of breast cancer masses, enabling radiologists to identify even the most minute masses during the early stages. The model achieved an impressive accuracy of 0.97.

Traditionally, much of the research in the literature has focused on classifying mammograms into broader categories, such as normal, benign, or malignant, based on the presence or absence of abnormalities. However, a more comprehensive and widely adopted system for mammogram classification is the BI-RADS which provides a standardized way of reporting mammogram findings, allowing radiologists to communicate their interpretations consistently.

To this end, researchers have created numerous deep-learning models to detect breast lesions and masses, effectively categorizing them according to the BI-RADS classification system. In 2020, (Chang et al., 2020) introduced three models designed to detect breast lesions and tumors from ultrasound images and subsequently classify them

into BI-RADS categories 2 to 5. The models employed RF, SVM, and CNN techniques. The dataset, comprising 151 samples in JPEG format, was collected from various local hospitals in Taiwan. Before analysis, the images underwent pre-processing, followed by segmentation using K-means clustering. Feature extraction was then incorporated into the model. Remarkably, the CNN model attained the highest accuracy of 85.42%, primarily due to its capacity to leverage pixel structures effectively. Later 2020, (Medeiros et al., 2020) introduced a mammography-based model utilizing the publicly available dataset known as "CBIS-DDSM." The dataset consisted of 10,239 mammography samples. For feature extraction, they employed 11 CNN algorithms, alongside traditional classifiers like Bayes, MLP, SVM, and RF. The researchers implemented transfer learning to extract features in Regions of Interest (ROIs) from the mammograms and then classify them into BI-RADS categories 3 and 4. Notably, the combination of DenseNet201-MLP achieved an accuracy exceeding 63%, surpassing the performance of a human expert by 9.0%.

The INbreast dataset serves as a widely used resource for researchers conducting breast cancer classification experiments. It comprises 410 mammograms in DICOM format, acquired from 115 patients and presented in Medio-lateral Oblique (MLO) and Cranio-caudal (CC) views (Siddeeq et al., 2021). Notably, a significant portion of the mammograms in this dataset belongs to BI-RADS category 2. In (Siddeeq et al., 2021), the researchers employed this dataset to classify mammograms into six categories (1, 2, 3, 4, 5, and 6) using a customized CNN based on ResNet architecture. Given the dataset's imbalance, they effectively utilized elastic deformation data augmentation to address this issue. The dataset was thoughtfully divided into 76% for training, 24% for testing, and 10% of the 76% training set for validation purposes. Remarkably, the study

achieved an impressive accuracy of 85.9% even with the imbalanced data. Authors in (T. M. Khan et al., 2021) presented another study using the same INbreast dataset and expanded the classification into eight categories (1, 2, 3, 4A, 4B, 4C, 5, and 6). To achieve this, the researchers utilized various feature extraction techniques such as GLCM, GLRLM, LBP, LBGLCM, and SFTA. Addressing the imbalance, they employed Adaptive Synthetic (ADASYN) for oversampling. Furthermore, they utilized ReliefF for feature selection and applied a Feedforward Neural Network (FNN) for classification. The study in (T. M. Khan et al., 2021) showcased exceptional results, achieving an accuracy, sensitivity, and specificity of 99.5%, 99.5%, and 99.4%, respectively. Researchers employed transfer learning techniques on the INbreast dataset in (Adedigba et al., 2022) using popular pre-trained models, namely AlexNet, VGG, ResNet, DenseNet, and SqueezeNet, to classify mammograms into six distinct categories (1, 2, 3, 4, 5, and 6). Data augmentation was applied to enhance the model's performance. Remarkably, the study (Adedigba et al., 2022) achieved an impressive highest accuracy of 99.8% with the utilization of DenseNet. However, it's important to note that in this study, some misdiagnoses occurred for BI-RADS classes 2 and 5. The misclassifications were observed when mammograms depicting fatty breasts were occasionally misclassified as cancerous.

While some researchers heavily relied on publicly available datasets, others preferred to create their unique datasets. In (Thang et al., 2021), the authors collected their dataset from 1161 Vietnamese patients, resulting in 4664 DICOM images. Utilizing this dataset, they developed a two-phase model. Initially, transfer learning with different algorithms, such as ResNet-50/ResNet-101 and Inception, was implemented to classify images into BI-RADS categories 0 and 1. In the second phase, Mask-RCNN was employed to

accurately identify, localize, and classify lesions according to BI-RADS categories 2 to 5. Despite achieving precise ROI determination, the model's overall accuracy remained suboptimal, with a result of only 58.44%. The use of scanned images during model building, followed by testing on digital images, was identified as a contributing factor to the lower accuracy.

Similarly, (Ibrokhimov & Kang, 2022) presented a two-stage model to classify mammograms into seven categories (1, 2, 3, 4A, 4B, 4C, 5). Two datasets were utilized, with the first being accurately labeled by senior radiologists and containing 3134 mammograms belonging to BI-RADS categories 2 to 5. The second dataset was the INbreast dataset. The data was divided into 50% for training, 13% for validation, and 37% for testing. In the first stage, ROIs were extracted, and enhanced, and square patch images were generated. The second stage involved using Faster R-CNN for tumor detection, with transfer learning fine-tuning parameters using INbreast. The model achieved an accuracy of 95%, and an AUC of 0.97.

Furthermore, (Tsai et al., 2022) claimed to have introduced a model for the first time in the literature to classify mammograms into eight BI-RADS categories (0, 1, 2, 3, 4A, 4B, 4C, 5). They excluded BI-RADS 6, which indicates a known biopsy-proven malignancy, while including BI-RADS 0, indicating incomplete assessment. The data, obtained from E-Da hospital in Taiwan, comprised 5733 mammograms in CC/MLO views from 1490 patients, and was annotated by eight radiologists. Block-image segmentation of 224x224 pixels and EfficientNet were employed in the model. The model achieved impressive results with an accuracy of 94.22%, sensitivity of 95.31%, and specificity of 99.15%. The AUC was measured at 0.9723. Finally, the least

favourable results they got were observed for BI-RADS category 1 lesions, with sensitivity and precision reaching 81.22% and 85.91%, respectively.

As technology advances, incorporating the BI-RADS scoring system into machine learning and deep learning models can lead to even more accurate and nuanced mammogram classifications, ultimately improving breast cancer detection and patient outcomes. However, the application of deep learning in this domain encounters numerous challenges. Firstly, data collection may involve multiple sources and diverse formats as shown in Figure (2.2) (Smiti, 2020), making it challenging to achieve model generalization across these varied data resources. Additionally, the collection of large annotated datasets can be difficult, posing a challenge to training deep learning models effectively. Another obstacle arises in the form of essential pre-processing tasks, which are required prior to classification to address the potential influence of various factors that may lead to misinterpretation (Murtaza et al., 2020). Moreover, training deep learning models with complex architectures and extensive datasets necessitates substantial computational resources and time.

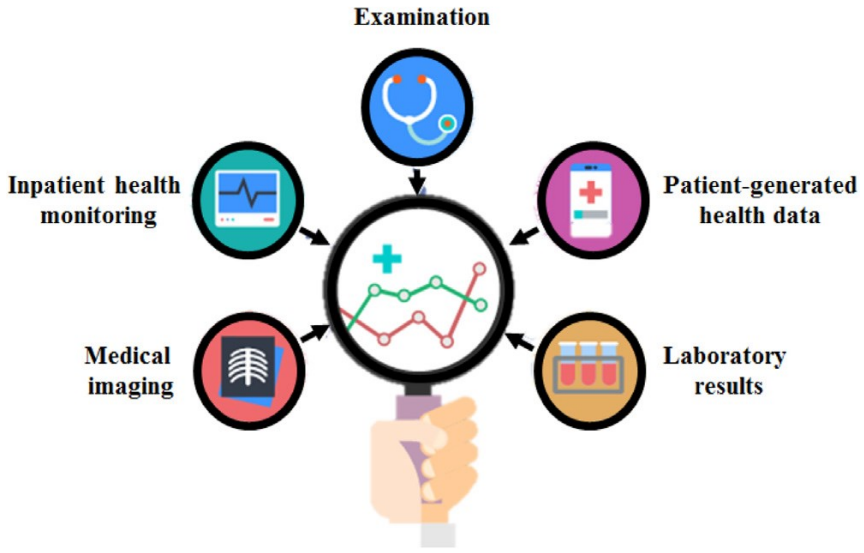


Figure 2.2: Medical data sources (Smiti, 2020)

In summary, deep learning has proven its efficiency in detecting breast cancer. Multiple studies utilized different CNN frameworks to classify mammography into benign, malignant, and normal (Castro-Tapia et al., 2021; Chougrad et al., 2018; Gnanasekaran et al., 2020). Most of the presented works achieved high accuracy which enables the radiologists to use them in the clinical implementation.

2.4 Summary

This chapter provided an overview of breast cancer and the efforts that were made in the previous studies for detecting breast abnormalities. In addition, many machine learning and deep learning models were used in the state of the art. Lastly, the following chapters will present the methodology followed in this thesis to implement the most common deep learning models that are based on the Convolutional Neural Network on a unique dataset.

Chapter Three: Dataset and Pre-processing

3.1 Introduction

In Artificial Intelligence, data represents one of the most important building blocks, the backbone of data-driven applications, and the main pillar for informed decisions across sectors (Kulakli & Osmanaj, 2020). In the era of the information revolution, different sectors produce huge amounts of data in diverse formats. It can be textual, numerical, or even images. This data should be utilized to extract valuable insights, detect patterns, and uncover potential trends (R. Khan et al., 2024). In addition, researchers are using this data for modelling to predict and diagnose various diseases at early stages when the treatment is available and affordable, thus increasing survival rates. In this thesis, the focus will be on the medical image data from the healthcare sector notably mammograms. In this chapter, the dataset used in this thesis will be presented starting from the data collection process of a local mammogram dataset. This includes specifying the source of the data, the size, and the pre-processing phase.

3.2 Data Collection

There are many open-access datasets for mammograms that are available online for researchers. It is worth mentioning that there are common datasets for mammograms that have been widely used in previous works by researchers. Examples of these common datasets are INBreast, MIAS, and CBIS-DDSM. Most of the presented work in the literature utilized these three datasets (Thang et al., 2021), (Ibrokhimov & Kang, 2022).

In this thesis, a contribution to the knowledge was added by collecting a novel brand-new dataset to use for modelling. The aim was to collect data that contained

mammogram images and medical reports from government mammogram centers and this was the starting point.

In the West Bank, there are 14 governmental centers owned and operated by the MoH distributed among all districts providing mammogram screening services (Ministry of Health, 2022). These centers are located in Jerusalem, Ramallah, Jericho, Salfit, Tulkarm, Tubas, Nablus, Jenin, Qalqilya, Bethlehem, North Hebron, Hebron, South Hebron, and Yatta. The data collection process went through several stages illustrated in the following Figure (3.1).

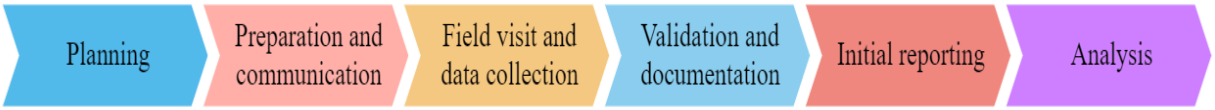


Figure 3.1: Process of data collection.

1. **Planning**: in this stage
    - Defined the objective of this process which is to collect the largest possible number of mammogram images available at the center, in addition to the mammogram reports and medical history accompanying each image.
    - List the number of centers where the required data is available. In this case, the data is only available at one center which is Bethlehem Mammogram Center.
    - Prepared an official request letter to be sent to MoH to facilitate my mission as a researcher to collect the required data from the center.
2. **Preparation and communication**: in this stage

- Sent the official request letter to MoH and got consent to collect and use the dataset in my research (the approval letter is attached in the annexes section).
- Communicated and coordinated with the focal point in the center in advance to set a suitable appointment for the field visit.
- Explained the purpose of the mission to the staff in the center.
- Kept a paper copy of the consent with me to provide if needed during the field visit.
- Prepared a large storage capacity memory to store data.

3. **Field visit and data collection**: in this stage
    - Visited Bethlehem mammogram center and met the responsible doctor and her team. They store the data in simple techniques with their own effort because there is no storage tool provided by the MoH.
    - Spent Two working days in the center to extract and collect the required dataset that is displayed in Table (3.1).

Table 3.1: Collected Data.

BI-RAD	***Number of ladies***	***Total No. of Images***	***Mammogram Reports***	***Medical History Reports***
1	55	297	74	71
2	28	133	42	40
3	26	132	39	39
4	24	89	43	41
5	17	34	17	17
6	24	46	23	23
Total	**174**	**731**	**238**	**231**

4. **Validation and documentation**: in this stage
    - Went through all mammogram images to double-check accuracy and reliability and to make sure that it's useable for modeling.

- Labelled the images into Six categories according to the BI-RADS scale and the mammogram report.
- Grouped medical reports and medical history into six groups as well.

5. **Initial reporting**: in this stage
    - Created a ***Metadata*** sheet about the image dataset in an Excel sheet to summarize the general findings in the dataset in terms of the BI-RAD score, the year of birth, age, and the breast side (left, right, or both) where the abnormalities can occur.
6. **Analysis**: in this stage
    - Conducted different analysis types on the collected dataset.
    - Extracted insights and conclusions based on the analysis, and this will be presented in detail in chapter four.

3.2.1 Challenges:

- Lack of mammogram data resources in the West Bank generally.
- Lack of the availability of mammogram data in the MoH particularly despite the hard work in screening mammograms on the ground. This is attributed to unavailable storage tools.

In summary, despite the MoH mammogram centers being overwhelmed with the number of screening appointments compounded by shortages of dedicated staff, the lack of data stored posed a significant challenge in collecting a comprehensive dataset. Consequently, the research ended up collecting the dataset exclusively from Bethlehem Mammogram Center.

3.3 Data Description

3.3.1 Size

In general, women are examined using a digital mammogram device. Four images of the breast are taken per woman, two images for each side. Sometimes more than this is taken based on certain indications. Furthermore, women older than 40 are encouraged to repeat the screening every year. The collected dataset contained different data types. Firstly, a total of 731 mammogram images with a standard dimension of 4710×5844 pixels were gathered from 174 ladies who were screened in Bethlehem Mammogram Center before July 2023. Secondly, I have collected also a total of 238 mammogram reports which included the BI-RADS score resulting from the examination for the same ladies. These reports were used in the labeling process later. Lastly, 231 medical history files have been also collected. Moreover, the number of mammogram reports and medical history files is greater than the number of women. This is because some women have undergone the examination several times over the years, and each time a medical history is taken in addition to a new mammogram report being issued.

3.3.2 Data Type, Format, and Views

The dataset includes mammogram images in jpg and DICOM format. DICOM stands for Digital Imaging and Communication in Medicine (Clunie, 2021). It's an international standard presented by the National Electrical Manufacturers Association (NEMA) and the American College of Radiology (ACR) to facilitate harmonization among different medical imaging systems (Mantri et al., 2022). In addition, the images in the dataset are taken from both breast sides. Each side is imaged in different orientations with two different views namely Mediolateral Oblique (MLO) and Cranial Caudal (CC). The MLO view is taken from the center of the chest outward, whereas the

CC view is taken from above the breast. The MLO view offers a more extensive perspective of the upper-outer quadrant, providing an optimal visualization of the breast's lateral side. In contrast, the CC view presents a comprehensive depiction of the entire breast (Mohamed, Luo, et al., 2018). Figure (3.2) presents an example of MLO and CC views of the left side for a lady in the BI-RAD 6 category.

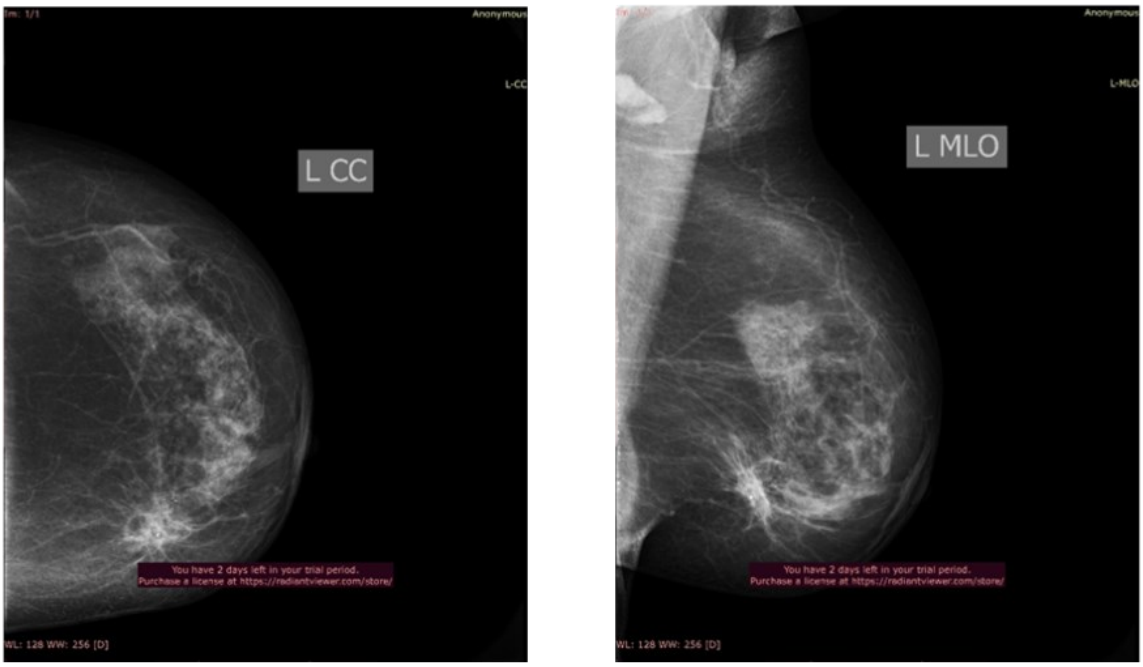


Figure 3.2: Examples of the mediolateral oblique (MLO) view and Cranial Caudal (CC).

In addition to images, there are other types of data in this dataset, which are textual files representing the mammogram report and medical history. The mammogram report is a text document in which doctors record their analysis, observation, and interpretation of the mammogram image screening examination. This includes diagnoses for both the left and right sides of the breast, the BI-RAD score, impressions, and specifying the assessment for either one or both sides of the breast. Additionally, the medical history file includes demographic details and (yes/no) questions concerning breast health, family planning, family history of breast cancer, and pregnancy. It also incorporates free-text responses and concluding comments (example attached in annexes).

3.3.3 Features

The mammogram reports within this dataset indicate and mention various features observed in the images, including lesions, masses, asymmetry, microcalcifications, clusters, density, and lymph nodes. These features play a significant role in assessing mammograms and assigning the BI-RAD score. The following is a brief about each feature.

- **Lesions**: refers to the abnormal areas in the breast tissues that could be malignant or benign (Nwadike et al., 2018). It's one of the important mammographic features that may appear in the screened images.
- **Masses**: also called breast lump, it's the abnormal development in the breast tissues where the shape and edges look different than the rest of the tissues (National Cancer Institute, 2023). Breast masses have different sizes and textures. It may or may not be cancerous, as well as potentially causing pain. Detection through self-examination can be challenging, making image screening a valuable method for identification. Furthermore, it's considered a sign of breast cancer.
- **Asymmetry**: Asymmetric breast tissue denotes a variation in volume or density, where one breast exhibits a greater amount of tissue than its counterpart in the opposite breast (Samardar et al., 2002).
- **Microcalcifications**: Breast calcification in general is the small calcium deposits in the breast tissues. It shows as small white dots or flecks in the mammogram. Microcalcification in particular appears as small flecks less than 0.5 mm similar to the size of small grains of salt. It's usually non-cancerous but some patterns require further investigation (Mayo Clinic, 2023).

- **Density**: measures the amount of fibrous or glandular tissues compared to the fatty tissues in the breast. Dense breasts are at higher risk of developing breast cancer (American Cancer Society, 2022).
- **Lymph nodes**: Intramammary lymph nodes (IMLN) are characterized as lymph nodes enclosed by breast tissue on all sides. This represents a frequently observed benign occurrence in mammograms (Bitencourt et al., 2019).
- **Clusters**: refer to groups of abnormal features, such as microcalcifications or masses, that appear closely grouped in a mammogram.

Typically, the presence of these features in a mammogram image does not automatically indicate the presence of breast cancer; instead, it may suggest a potential development of cancer. Many of these features are more likely to be non-cancerous. However, for certain patterns, further investigation is advised to facilitate early detection of the disease.

3.4 Data Pre-processing

This section provides details about the image pre-processing phase. It's an important step in the modelling cycle to prepare the data to be fed into the used models. In this thesis, two approaches were conducted in the modelling stage. However, the same pre-processing steps were applied in both. The collected images included a lot of annotations, text, and unwanted areas like the Pectoral muscle. Likewise, medical reports had some useless data that would not lead to useful information. Several pre-processing methodologies were applied to the data to clean and prepare it to be suitable for use in the later stages. The following sections present the pre-processing techniques for the images and medical history files.

3.4.1 Mammogram Images Pre-processing

The mammogram images went through seven stages shown in Figure (3.3) to get rid of the annotations and clean the images to become ready for exploration and modelling.

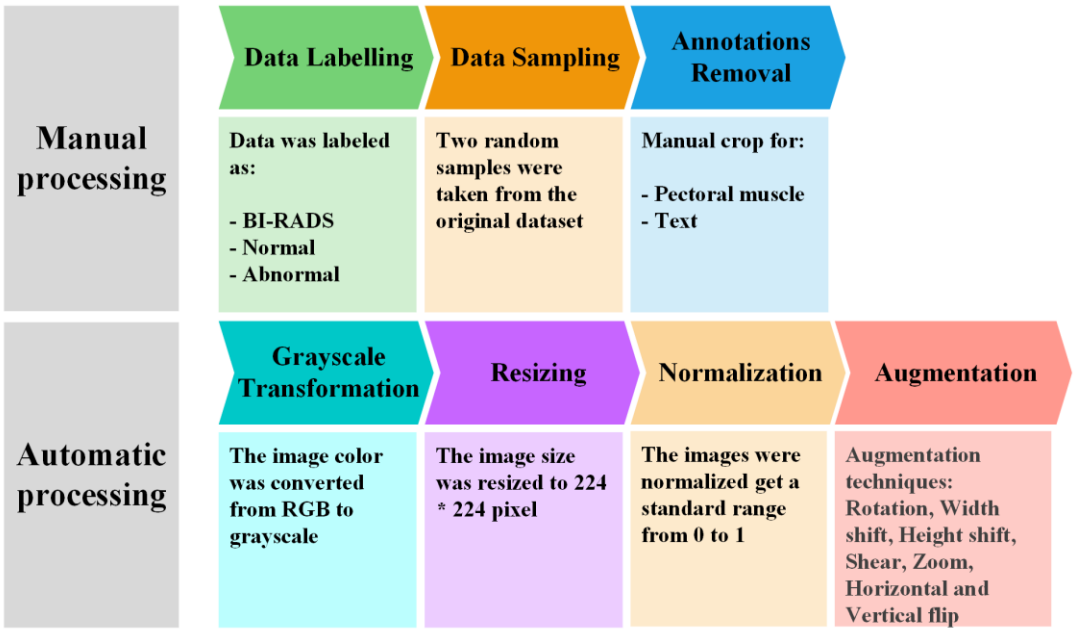


Figure 3.3: Image pre-processing workflow.

First of all, the collected mammogram images were categorized into 6 groups based on the BI-RADS score by the center staff that we gathered the data from. Moreover, the images themselves did not include the BI-RADS score result but were documented in the mammogram report accompanying each image. Although the images were categorized by the doctor and technicians, image labels and results were checked again and mapped together to avoid human errors. In addition, the images were categorized into normal and abnormal to be used in one of the modelling approaches. Based on the BI-RADS score and the likelihood of abnormalities being malignant, the images of scores 1 to 3 were labeled as normal, and the images of scores 4 to 6 were labeled as abnormal.

Moreover, two balanced random samples were taken from the original dataset. The first sample included 200 mammogram images, where 100 images were taken from the normal category (BI-RADS 1,2, and 3) and 100 from the abnormal category (BI-RADS 4,5, and 6), while the second sample included 40 images, 10 images from each of BI-RADS groups 3, 4, 5, and 6. Neither sample included many images because some of the original categories contained very few images, such as BI-RADS five and six, which are the focus of this study. In addition, some images contain texts covering important parts, and scars of previous breast surgeries that may confuse or mislead the results.

Furthermore, the images included in the sample contain some sensitive texts, such as the patient's data, in addition to other texts that indicate the type of view whether it's MLO or CC, the time and the date the image was taken, and information about the screening device. These texts have been manually cropped to ensure that other important areas of the breast are preserved from being affected. The Pectoral muscle was also cropped in the MLO view so as not to confuse the classification process because it appears in white, and abnormal areas also appear in white.

Additionally, the image colors were converted from RGB to Grayscale to reduce channels and dimensions. The images also were resized from their original size to 224*224 to be compatible with the input size of the models used later. Then, normalization was applied to the images to get a standard range between 0 to 1. Finally, several techniques were used to enlarge the number of images in both samples while keeping it balanced. These techniques included rotation, width and height shift, shear, zoom, and horizontal and vertical flip. As a result of the augmentation, 10 pre-processed images were created from each original image totaling 2000 images in the first sample of normal and abnormal categories and 400 images in the second sample of BI-RADS 3

to 6. Moreover, Figure (3.4) shows an image from the random dataset before and after preprocessing. The image located in the first row on the left is the original image before doing anything. The image included annotations like the patient's details, the view mode (MLO), other unwanted text, and the pectoral muscle. In addition, the image on the right side is the same but after cropping the text and the pectoral muscle, converting the color to grayscale, resizing it to 224*224, and implementing normalization. Finally, the ten images in the second and third rows are the augmented images that were created from the same image in the first row by using the techniques mentioned previously.

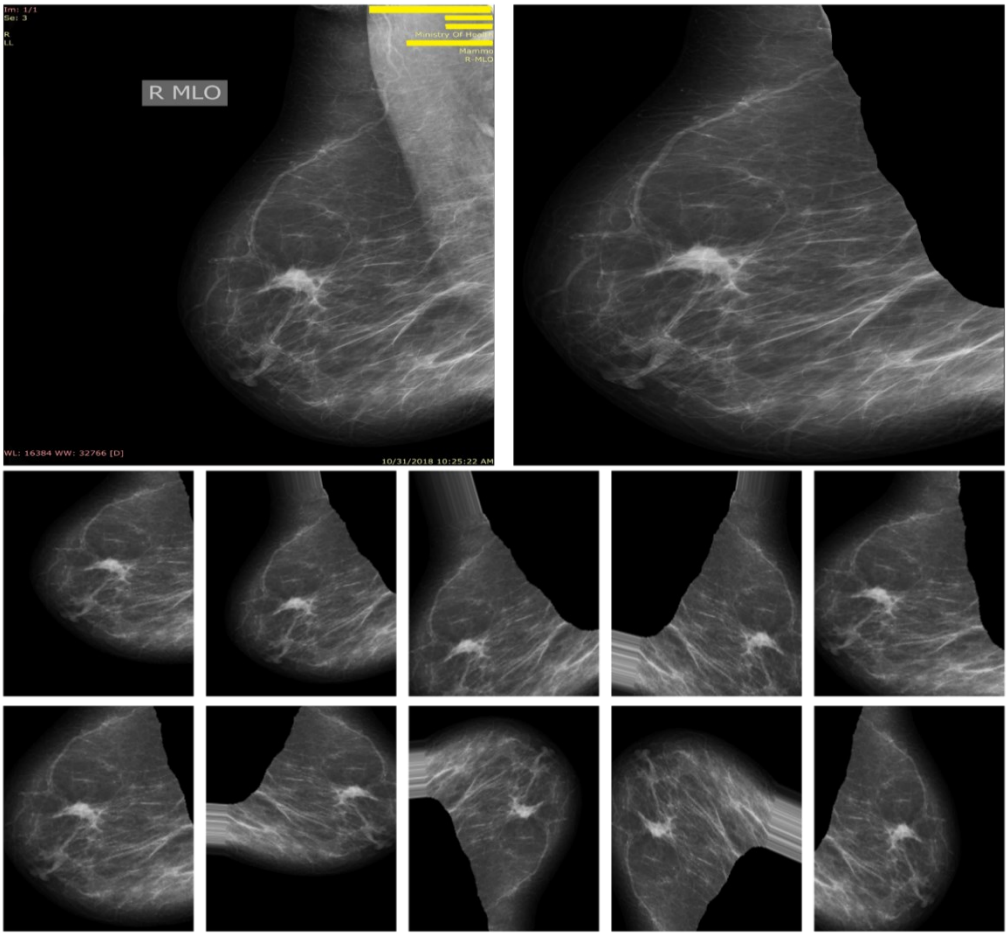


Figure 3.4: Mammogram image before and after pre-processing.

3.4.2 Medical History Reports Pre-processing

Medical history is invaluable for cancer patients, especially breast cancer. it's considered the cornerstone for accurate diagnosis and treatment. It provides care providers with baseline information about the patient's situation before any kind of assessment. From the patient's history, doctors can identify associated risk factors and family history of breast cancer or other diseases like diabetes and hypertension. Hence, guiding the doctors to put a personalized and suitable treatment plan. In addition, the collected medical history reports in the form of a Word document include many questions concerning the following:

-Age	-Other diseases
-Referral status	-Family history of breast cancer/other cancer disease
-Marital status	-Breastfeeding now or not
-Previous problems	-Pregnancy status
-Previous mammography	-Age at first menses
-Contraceptive Pills / Hormones	-Age at marriage
-Breast pain/tenderness	-Age at the first delivery
-Breast discharge/color	-Number of children
-Breast lump/mass	-Menses' status
-Previous operations/biopsy	-Weight

The questions and answers for the 231 reports have been transcribed into an Excel sheet. Some questions were neither dropped nor excluded due to their low completeness rate. Many questions like the date of the previous mammogram were not filled out and the completeness rate was less than 10%. Moreover, null values in the Excel were filled with column mode. In Figure (3.5), a snapshot of the Excel sheet is presented for the first four rows with the first row being the header.

BIRAD	Report date	Age	Referred by	Learned by activities	Marital Status	Previous problems	Previous Mammogr
1	2023	45	MOH	Medical	Married	0	0
1	2023	53	MOH	Medical	Married	0	1
1	2023	48	Herself	Medical	Married	0	0

Contraceptive Pill/Hormones	Breast pain / ter	Breast discharge /color	Breast lump / Mass	Previous breast OP / Biopsy	Other disease / Operation
0	0	0	0	0	1
0	0	0	0	0	1
0	0	0	1	0	1

Family history of breast cancer /other Cancer disease	Breastfeeding now	Pregnant	Age first menses	Age married	first delivery
0	0	0	14	24	25
0	0	0	11	18	19
1	0	0	15	26	28

No pregnancies	Previous menses	Weight
5	Normal menstrual	69
6	Menopause	125
7	Normal menstrual	70

Figure 3.5: Snapshot from the Excel sheet of medical history reports.

3.5 Summary

This chapter presented the collected mammogram dataset. In addition, a description of this dataset regarding the size, format, data type, and features was provided. Finally, this chapter included an overview of the pre-processing phase. Insights and explorations from this data will be presented in the next chapter.

Chapter Four: Exploratory Data Analysis

4.1 Introduction

Exploratory Data Analysis (EDA) is an important stage in the data science cycle. In this stage, different analysis and visualization techniques are utilized to transform data into useful information, uncover patterns, and extract insights. In this study, statistical analysis and visual analysis which includes graphical representation will be employed to explore the data. This stage is important to make an informed decision later and will ensure a comprehensive understanding of the data before going to the modelling stage. This section is divided into subsections to explore the data from different perspectives.

4.2 Original Dataset Exploratory

In the data collection section above, I've mentioned that I've created an initial report (Metadata) which is an Excel sheet containing information about images in the mammogram dataset describing the BI-RAD score, the side of the breast, year of birth, and age in years. Figure (4.1) is a snapshot from the Metadata sheet.

BIRAD	Side	Year of birth	Age in years
3	Left	1963	60
3	Right	1963	60
3	Left	1948	75
3	Left	1963	60
3	Both	1974	49
3	Both	1977	46
3	Both	1977	46
3	Right	1973	50

Figure 4.1: Snapshot form metadata sheet.

Using Python and Jupyter Notebook, I created a new column categorizing individuals into six age groups based on their existing age values in the dataset.

The following table lists the columns with their type and value.

Table 4.1: Metadata Features.

Column (feature)	*Type*	*Value*
BIRAD	Integer	1,2,3,4,5,6
Side	Text (object)	**Left**: the image is taken of a woman who has abnormalities on one side of the breast, in this case, the left side. **Right**: the image is taken of a woman who has abnormalities on one side of the breast, in this case, the right side **Both**: the image is taken of a woman who has abnormalities on both sides of the breast.
Year of birth	Integer	[1939-1987]
Age in years	Integer	[36-84]
Age_group	Text (Object)	30-39, 40-49, 50-59, 60-69, 70-79, 80-89

4.2.1 Statistical Analysis for The Original Dataset

In this metadata sheet, some statistical measures were used focusing mainly on age.

Table (4.2) summarizes these measures.

Table 4.2: Statistical Measures.

Measures	*Value*
Total number of values	731
Min	36
Max	84
Mean	54.9
Median	54
Mode	50
Variance	74.5
Standard deviation	8.6
Skewness	0.74
Mean Absolute Deviation (MAD)	6.8

As Table (4.2) shows, the minimum age in the dataset is 36 while the maximum is 84. In addition, the average age in the original dataset is 54.9 which is close to the median value of 54, while the most frequent age value is 50. Moreover, the variance that assesses the average squared difference between data values and the mean is high with a value of 74.5 which indicates the wide range and the variability of age values. Furthermore, both standard deviation and mean absolute error compare the values to the

mean but using different formulas. The values were 8.6 and 6.8 for the standard deviation and MAD respectively. Additionally, the skewness measures the symmetrical distribution in the dataset. Although the data follows the normal distribution, we can see that the mean is slightly greater than the median which indicates that the data values are positively skewed or in other words right skewed as shown in Figure (4.2) below. These measures will be clear through the visuals displayed in the following section.

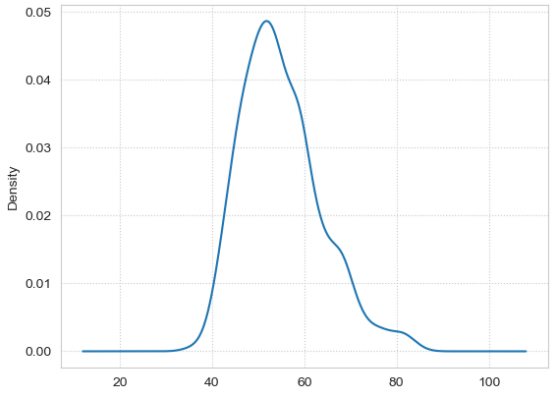


Figure 4.2: Age Skewness.

4.2.2 Data Visualization for The Original Dataset

In this section, several visualizations will be presented and interpreted to explore the data in detail and showcase the distributions across the dataset. The first visual is Figure (4.3) which is a bar chart illustrating the number of mammogram images in the dataset distributed by BI-RADS scores (categories). I chose to use the bar chart because we have a set of categories (BI-RADS) and a quantitative value (No. of mammogram images). In the x-axis, the BI-RADS categories are arranged according to the natural ordering from 1 to 6, while the y-axis represents the number of screened mammogram images. Each bar has a different color because the feature in the x-axis is a discrete categorical variable.

By looking at the chart, it's obvious that the dataset is imbalanced. BI-RAD 1 ranks first with 297 images that resulted in normal findings while BI-RAD 5 ranks last with 34 images that are highly suggestive of malignancy. The number of images in BI-RAD 2 and 3 is almost the same with benign findings. 89 images got suspicious findings for malignancy and were categorized as BI-RAD 4. Lastly, 46 images got proven malignancy findings and were categorized as BI-RAD 6.

We can conclude that the likelihood of getting normal results in Bethlehem Center is greater than the suspected or proven malignancy results and the number of detected women with proven breast cancer is very low in this center. Proven breast cancer percentage may differ from one district to another and it is attributed to many reasons like daily lifestyle, family history, genetics, and other reasons mentioned in the literature review chapter.

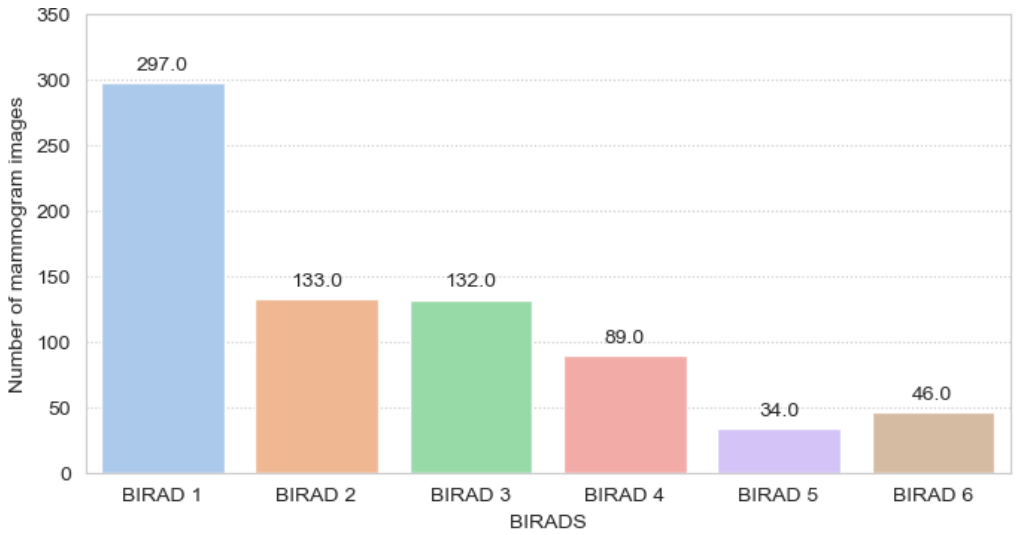


Figure 4.3: Number of mammogram images distributed by BI-RADS.

The collected images in the dataset were taken from women of different ages. Figure (4.4) illustrates the distribution of age groups among the dataset. The x-axis represents the age groups arranged according to the natural ordering and the y-axis represents the

number of mammogram images. We can see that the majority of images are related to women aged between 50-59. Age groups 40-49 and 60-69 took second and third place respectively with 195 and 137 images. In addition, there are a few images in the older age groups 70-79 and 80-89, and only 2 images lie in the youngest age group 30-39. Logically, there is a clear increase in the number of images in age groups older than 40 years because there are awareness campaigns and encouragement for women in these groups to undergo mammography examinations. Women aged 40 years and older are at higher risk of developing breast cancer than those younger ages.

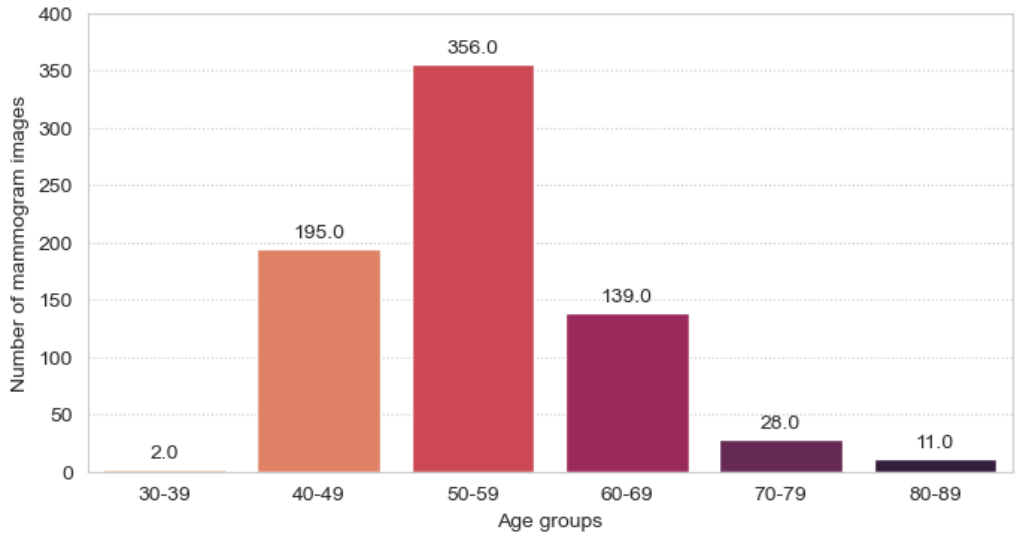


Figure 4.4: Number of mammogram images distributed by age groups.

Women usually experience breast abnormalities on either the left or right side. However, some women have abnormalities on both sides of the breast and it's not necessarily that both sides are classified with the same BI-RADS category. For example, the left side might be diagnosed as BI-RAD 3 while the right side as BI-RAD 5. In this dataset, some women are experiencing abnormalities on both sides. This is explained in Figure (4.5) below. Here, the x-axis depicts the breast side, and the y-axis depicts the number of images. As we can see, there is no significant difference between left and right. There

are 298 images with abnormalities on the left side, 286 on the right side, and 147 images on both sides.

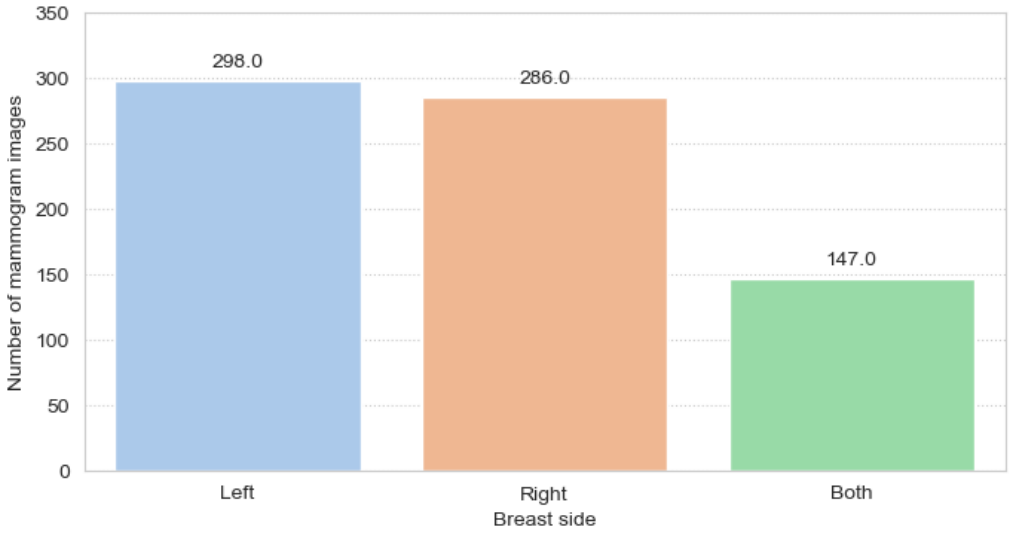


Figure 4.5: Number of mammogram images distributed by breast side.

The following pie chart shown in Figure (4.6) breaks down the screened images into 3 proportions: left, right, and both. 41% of the dataset is for the left side, and 39% for the right side, and together they constitute more than two-thirds of the sample. Images related to women experiencing abnormalities on both sides represent only 20% of the collected dataset.

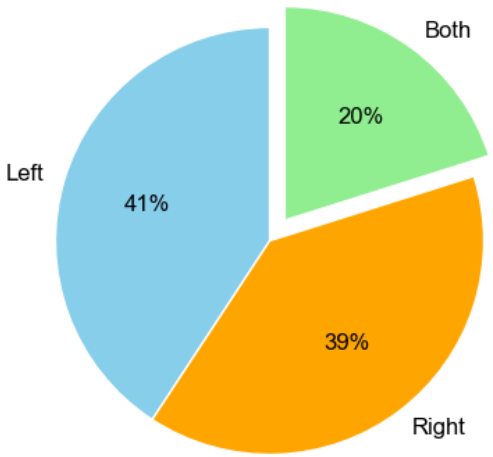


Figure 4.6: Breast side percentages.

In Figure (4.7), there are six bar charts related to the BI-RADS categories. These charts illustrate the number of mammogram images versus BI-RAD and age group. The x-axis signifies the age group, the y-axis signifies the number of images, and each chart is related to one of the BI-RADS categories. In BI-RAD 1, the number of images increases significantly until it reaches a peak at the age group of 50-59 and then decreases significantly until it reaches zero at the age group of 80-89. It can also be noted that the number of images in the 50-59 age group constitutes more than half of the images in this BI-RAD. Moreover, In BI-RAD 2, the number of images increases gradually as age increases, starting from the age group of 40-49 until it reaches a peak in the age group of 50-59 and then begins to decrease gradually. It can be seen that there are no images in the 30-39 age group.

In BI-RAD 3, it can be noted that most of the images in the dataset are concentrated in the age groups 40-49 and 50-59, while there is a noticeable difference in the other age groups. For example, there is a small number of images in the age groups 60-69 and 70-79, while there are no images at all in the age groups under 40 or over 80. Furthermore, In BI-RAD 4, images are centered around the 50-59 age group while there are few images in the age groups greater than 70 and no images in the age group less than 40. In this BI-RAD, the likelihood of having malignant results is less than 95%.

Additionally, BI-RAD 5 indicates highly suggestive results of malignancy with greater than 95% being malignant. It’s obvious in this BI-RAD that the occurrence of this category is in ages older than 50 and under 70. No images were found in age groups under 40, and only 2 images in the age groups older than 70. Finally, BI-RAD 6 means that the screening examination results in proven malignancy with a 100% chance of being malignant. In the chart below, we can see that the image number increases starting

from the age group 40-49 to reach its highest value of 18 in the age group 50-59 then decreases significantly in ages older than 70. As mentioned earlier, there is a small number of images in BI-RADS 5 and 6, and those images are concentrated around the 50-59 age group as well. The under 40 and over 70 age groups have few or no images.

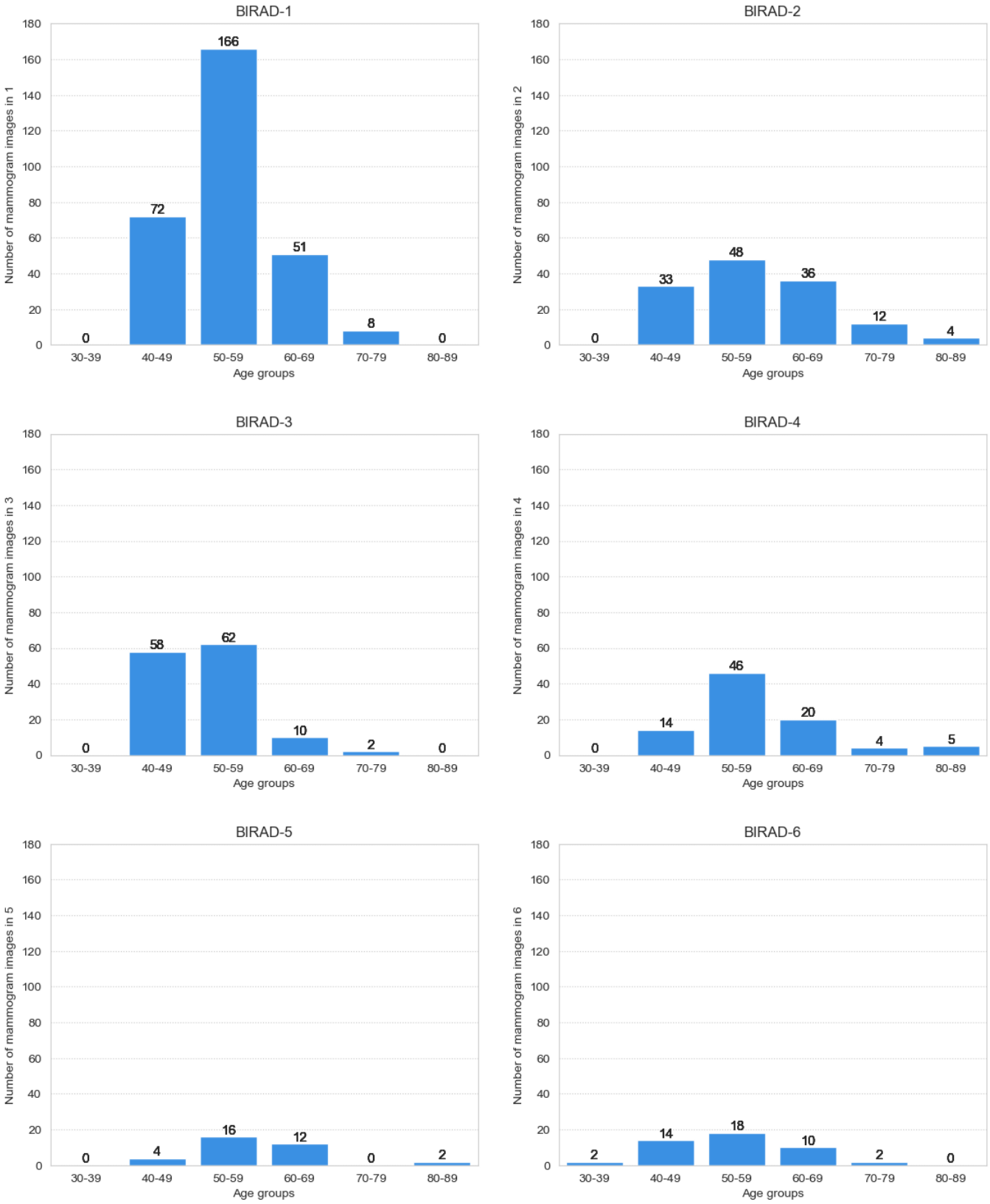


Figure 4.7: Number of mammogram images distributed by age groups and BI-RADS.

Delving into a more in-depth analysis, the distribution of images versus BI-RADS and the breast side have been explored. In Figure (4.8), the x-axis is used to represent BI-RADS categories, and the y-axis to represent the number of images. Moreover, I used colors as an additional dimension to represent the breast side as noted in the legend. Blue is for the left side, orange is for the right side, and green is for both sides.

It can be seen that the images in BI-RADS 1 and 5 are distributed almost equally on the left and right sides of the breast, while there are no images from cases that experience abnormalities on both sides at the same time. “Both” sides images are seen in BI-RADS 2,3,4, and 6. In BI-RAD 3, the “both” images represent more than the half. BI-RAD 3 is the threshold between benign and suspected malignant results. It seems from the results that women in BI-RAD 3 are at higher risk of developing abnormalities on both sides of the breast. The same case is also seen in BI-RAD 6 as well. This may be because the cancer is more likely to spread to both sides in advanced stages.

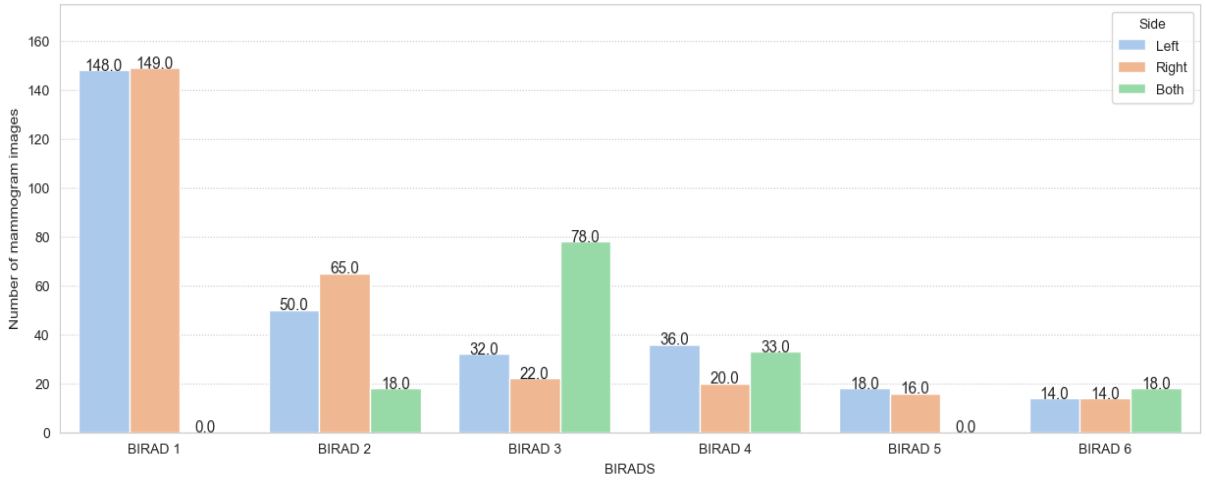


Figure 4.8: Number of mammogram images versus side and BI-RADS.

Subsequently, the analysis shifts to explore the distribution of the number of images versus age group and breast side. In Figure (4.9), the x-axis represents the age groups,

the y-axis represents the number of screened images. The color dimension for the breast sides is used as in the previous chart as well. Following what has been presented before, this chart confirms what was mentioned previously. First, the number of images is distributed among age groups between 40 and 80, where it begins to increase gradually until it reaches a peak in the age group 50-59 and then gradually decreases. Second, the number of images is distributed almost equally on the left and right sides of the breast. Finally, images of women who are experiencing abnormalities on both sides are found in the age groups older than 40 and younger than 70, with the largest number in the age group 50-59.

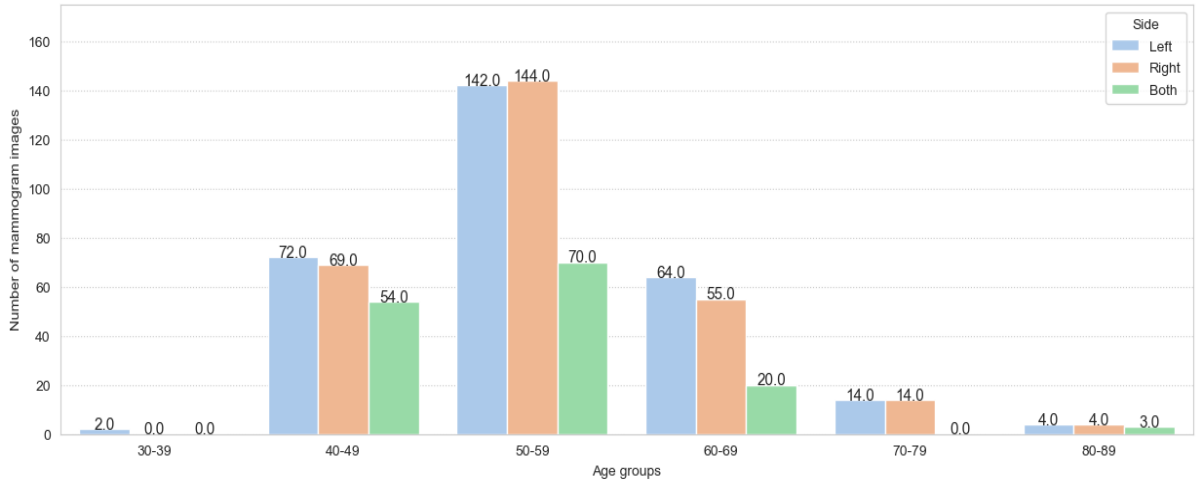


Figure 4.9: Number of mammogram images versus side and age group.

Finally, Figure (4.10) illustrates the age distribution of BI-RADS categories among the dataset. I used the overlapping density plot to achieve this aim. In this figure, age is represented in the x-axis while density is represented in the y-axis. We can see that age values in general follow the normal distribution with positive skewness on the right side and that was proven with the statistical measures provided in the previous section. Age values are concentrated in ages around 55. Moreover, Ages older than 70 are presented in BI-RAD 2. In the West Bank, women older than 40 are targeted for mammogram

screening while women under 40 are usually referred by their care provider when there are suspicious risk factors or the presence of specific symptoms.

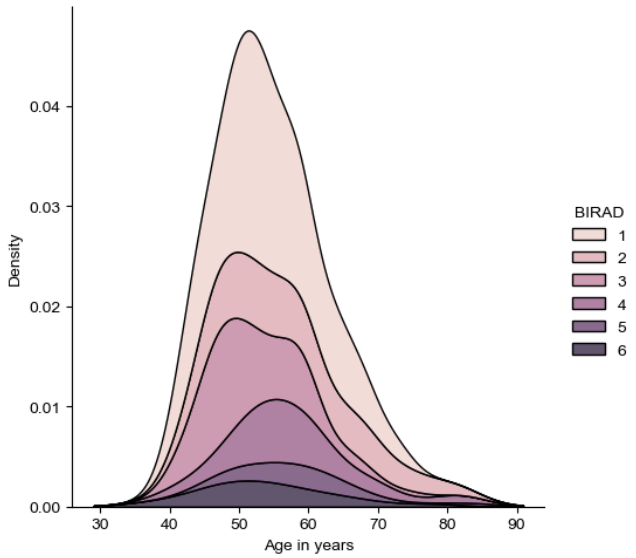


Figure 4.10: Age distribution versus BI-RADS.

4.3 Sample Dataset Exploratory (Normal and Abnormal categories)

In this section, several statistical and visual analyses will be utilized to analyze the sample dataset that was taken from the original dataset. This sample as mentioned earlier classifies BI-RADS categories into two classes; normal and abnormal. The normal class includes BI-RADS 1, 2, and 3, while the abnormal class includes BI-RADS 4, 5, and 6. The following sections present the exploratory analysis for the sample.

4.3.1 Statistical Analysis for The Sample Dataset

In this section, many statistical measures were conducted such as central tendency measures. This sample dataset has 200 images distributed equally between the two classes; 100 in the normal and 100 in the abnormal class. Figure (4.11) showcases a snippet from the Excel sheet showing the features used to describe the images in this sample dataset.

	BIRAD	Side	Year of birth	Age in years	View	Class	Age_group
0	1	Right	1969	54	cc	Normal	50-59
1	1	Right	1970	53	cc	Normal	50-59
2	1	Left	1969	54	cc	Normal	50-59
3	1	Left	1964	59	cc	Normal	50-59
4	1	Left	1970	53	cc	Normal	50-59

Figure 4.11: Sample dataset snippets.

This sheet has the same features used to describe images in the previous sections in addition to other features such as “View” and “Class”. These two new features are described in Table (4.3) below.

Table 4.3: Additional Features.

Column (feature)	***Type***	***Value***
View	Text (object)	*MLO*: Mediolateral Oblique view *CC*: Cranial Caudal view
Class	Text (object)	*Normal*: BI-RADS 1, 2, and 3 *Abnormal*: BI-RADS 4, 5, and 6

This statistical analysis focused on analyzing the age values in the sample dataset. Table (4.4) below lists the measures used and the results.

Table 4.4: Statistical Measures.

Measure	***Value***
Total number of values	200
Min	36
Max	84
Mean	57.1
Median	56
Mode	55
Variance	88.6
Standard deviation	9.4
Skewness	0.52
Mean Absolute Deviation (MAD)	7.4

From the results in the table, we can notice that the minimum age in the sample dataset is 36 while the maximum is 84. In addition, the average age in the original dataset is

57.1 while the median value is 56, and the most frequent age value is 55. Moreover, the variance is higher with a value of 88.6, indicating that the data is more spread around the mean. Furthermore, the standard deviation value is 9.4 and the mean absolute error is 7.4. Additionally, the skewness in this sample is less with the value of 0.52, and also the mean is greater than the median which indicates that the data is still considered positively skewed as shown in Figure (4.12). These measures will be displayed through the visuals in the following section.

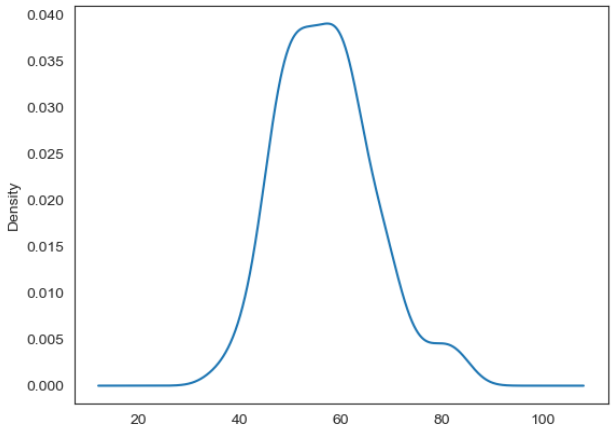


Figure 4.12: Age Skewness.

4.3.2 Data Visualization for The Sample Dataset

This section provides various charts and figures used to describe the data visually. Different types were utilized such as bar charts, grouped bars, and pie charts. These visuals will enrich this study with insights from the collected data. To start with BI-RADS, Figure (4.13) below shows the number of mammogram images distributed by BI-RADS scores. It is clear from the figure that the highest number of images is in BI-RAD 6 and the smallest number is in BI-RADS 2. Moreover, BI-RADS 1, 3, and 6 have almost similar numbers of 37, 38, and 42 in order. Similarly, BI-RADS 2, 4, and 5 lie within the same range.

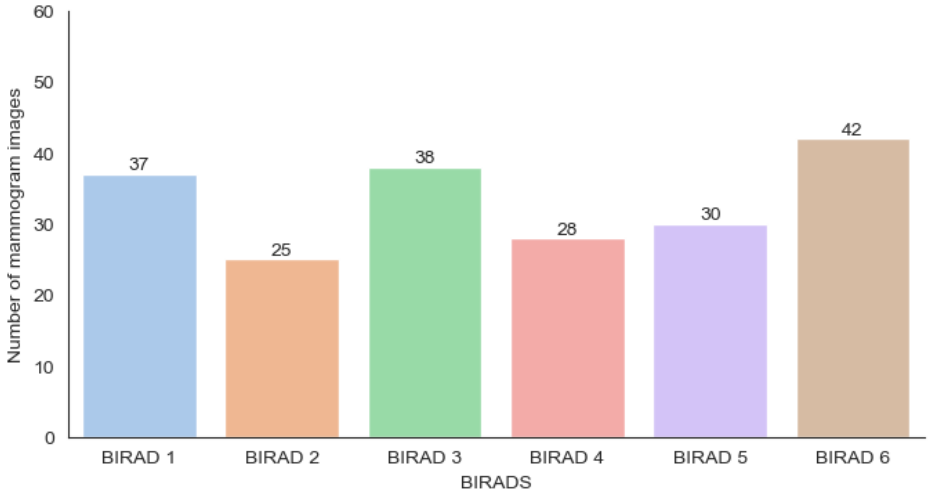


Figure 4.13: Number of mammogram images distributed by BI-RADS.

BI-RADS 1, 2, and 3 represent the normal class in blue, while 4, 5, and 6 represent the abnormal class in orange as shown in Figure (4.14). In the normal class, the least number of images is in BI-RAD 2, while the number of images in BI-RADS 1 and 2 represents 75% of the class. On the other hand, BI-RADS 4 and 5 together represent 58% of the class while BI-RAD 6 represents less than half with a percentage of 42%. The images within the same class are not equal because the images were selected randomly from the original dataset. However, the total of images inside each class is equal as shown in Figure (4.15).

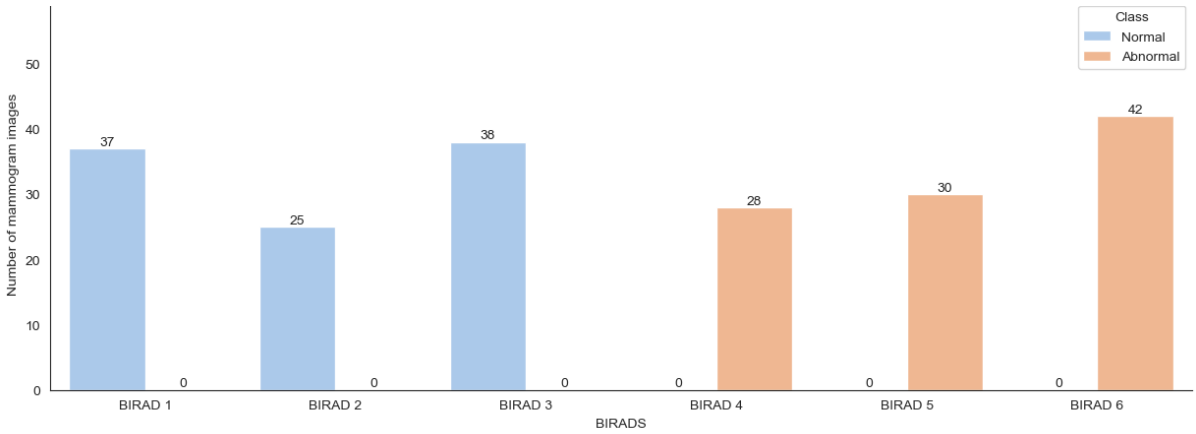


Figure 4.14: Number of mammogram images versus class and BI-RADS.

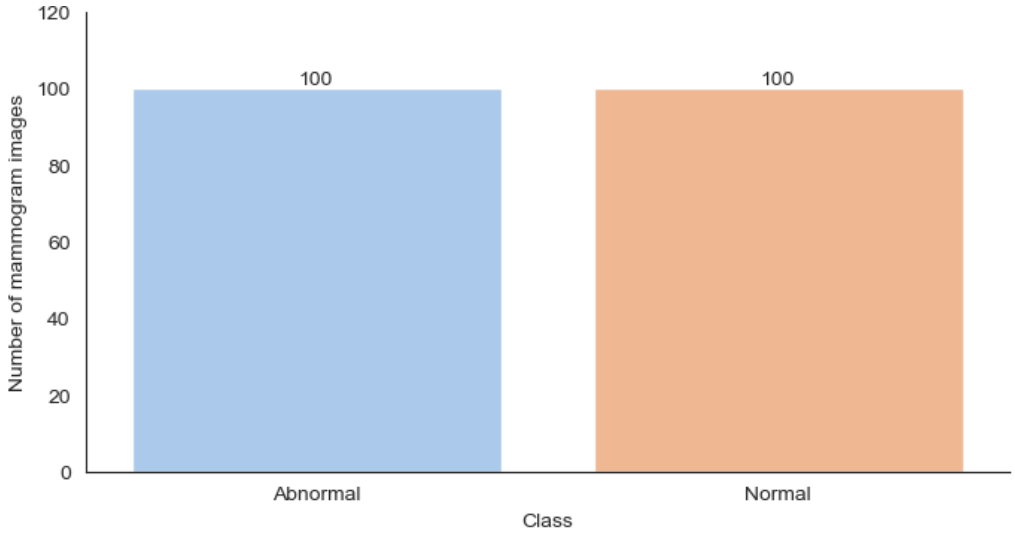


Figure 4.15: Number of mammogram images distributed by class.

Age group is an important aspect to be considered in such diseases. As one gets older, the probability of developing various diseases such as breast cancer increases. The most important reasons attributed to it are immunodeficiency with advancing age and the spread of chronic diseases among the elderly, such as diabetes and high blood pressure, which are considered among the most important risk factors that increase the rate of breast cancer. Figure (4.16) illustrates how the age groups are distributed among the sample. it can be seen that the dominant age group in the sample is 50-59, and this means that the largest group of women who have undergone a mammogram are between the ages of fifty and fifty-nine. On the other hand, there are a small number of women in the sample who are under the age of thirty-nine and those who are over seventy.

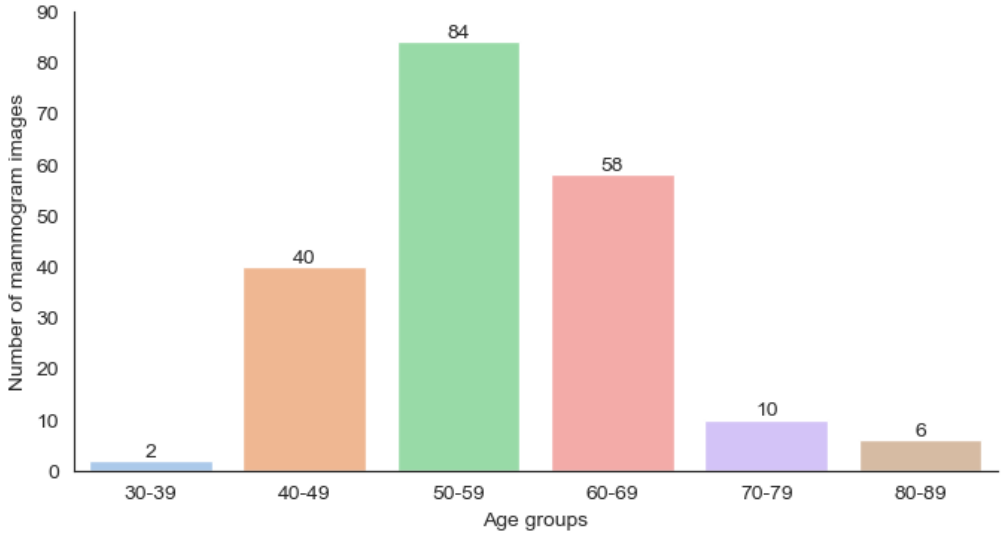


Figure 4.16: Number of mammogram images distributed by age groups.

Additionally, Figure (4.17) shows the distribution of the age groups among the classes normal and abnormal. In general, there is a slight difference in the number of images in each age group in the class. The number of images increases as age increases, gradually increasing until it reaches the highest value in the age group from 50 to 59, and then gradually decreases after that. Moreover, there are no images in the normal class for the ages under 40 and older than 80.

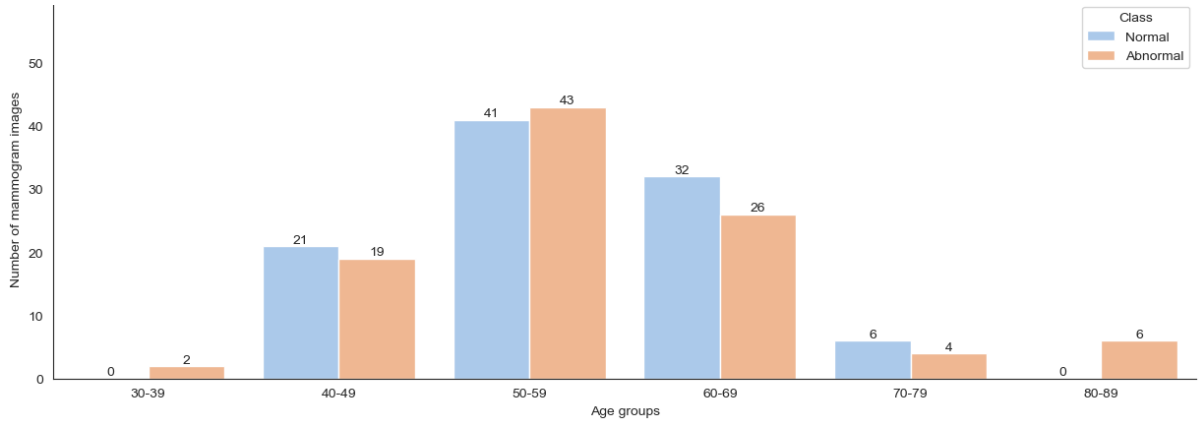


Figure 4.17: Number of mammogram images versus class and age group 1.

Figure (4.18) confirms the previous insight where we can clearly distinguish between the two classes. Both classes follow the normal distribution with slight differences.

Here, we can see that there is more variation in ages in the abnormal category where ages are distributed over six age groups while in the normal category, they are distributed over only four categories.

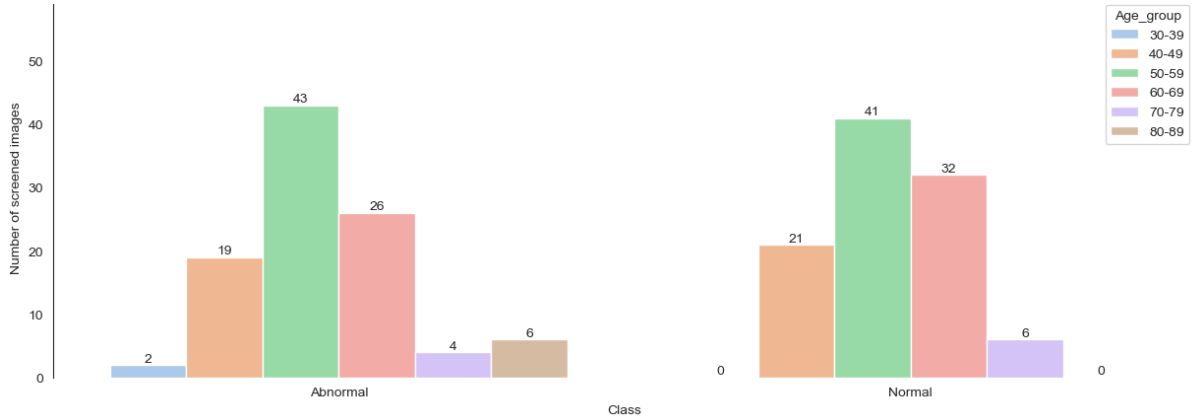


Figure 4.18: Number of mammogram images versus class and age group 2.

Another important aspect is the location of the abnormalities that appear in the breast. Women can experience cancer on one or both sides of the breast. To explore how it is distributed among the sample, different charts were used to discover that. It can be noted in Figure (4.19) that the largest number of images were taken from women with the possibility of cancer cells on the left side of the breast, and this group constitutes 41% of the sample which is almost half, while the number of images taken from women with the possibility of cancer cells on the right side is less with percentage of 34%. Also, there are 25% of the images in the sample are taken from women suffering from possible cancer cells on both sides of the breast. The “Right” side together with the “Both” sides constitutes the other half of the sample.

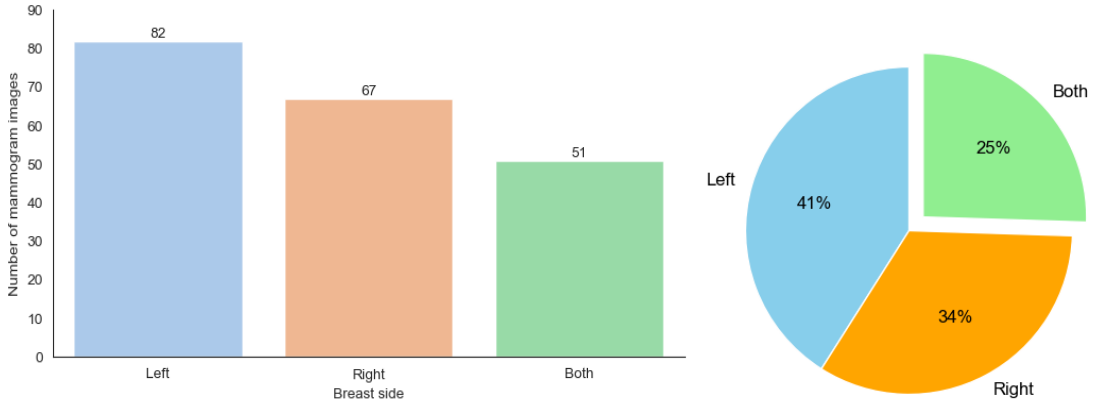


Figure 4.19: Number and percentage of the 'side' values in the dataset.

In the same way, it can be observed in Figure (4.20) that the number of images is similar on the left, right, and both sides of the breast between the normal and abnormal categories. Furthermore, the number of images in the “left” group is the highest and decreases gradually to reach the “both” group with the least number of images in both categories.

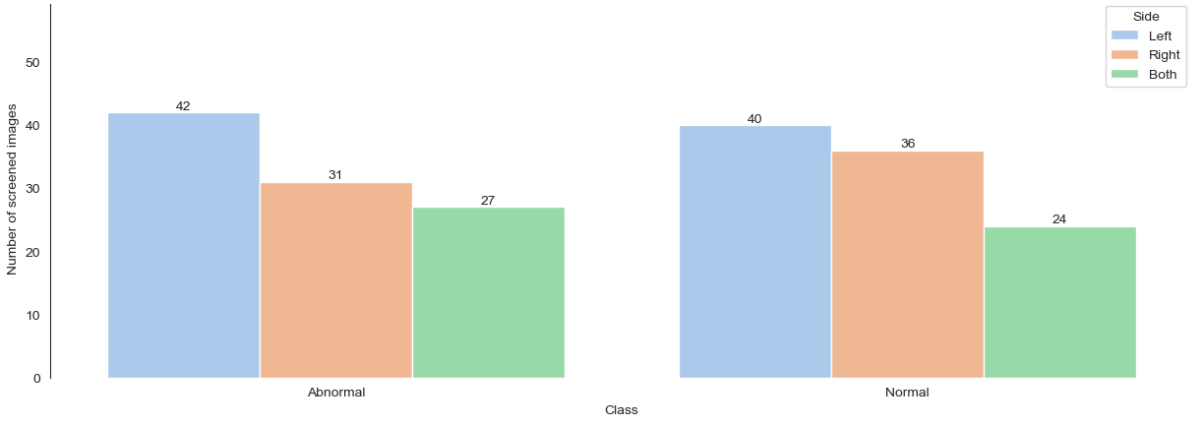


Figure 4.20: Number of mammogram images versus side and class.

Figure (4.21) shows the distribution of images versus Side and BI-RADS. There is a significant difference in the distribution of the Side across BI-RADS categories. BI-RADs 1 and 5 have almost an equal number of images of women that have abnormalities either on the left or the right side of the breast but not on both. BI-RAD 6 includes a close range of images of women who experience abnormalities on the left,

right, or both sides of the breast. Additionally, women who have abnormalities on both sides of the breast are significantly shown in BI-RAD 3. Less number of women who have abnormalities on the right side are found in BI-RAD 4. In general, the left and right sides are distributed throughout the BI-RADS scores with similar numbers for the left and right groups inside each BI-RAD.

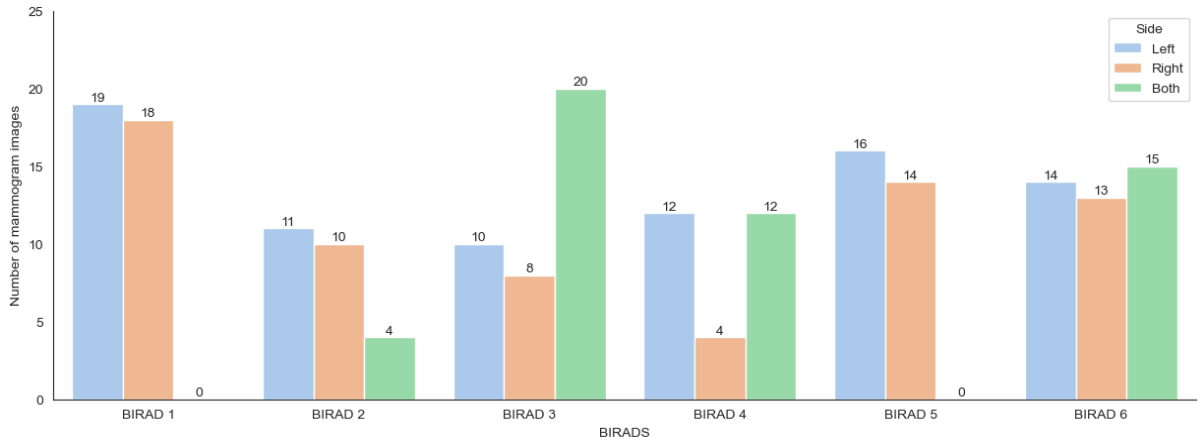


Figure 4.21: Number of mammogram images versus side and BI-RADS.

Figure (4.22) below shows that the number of women who have abnormalities on both sides of the breast tends to increase in the age groups older than 40 and under 70, while there are few or no images in the ages under 40 and older than 70.

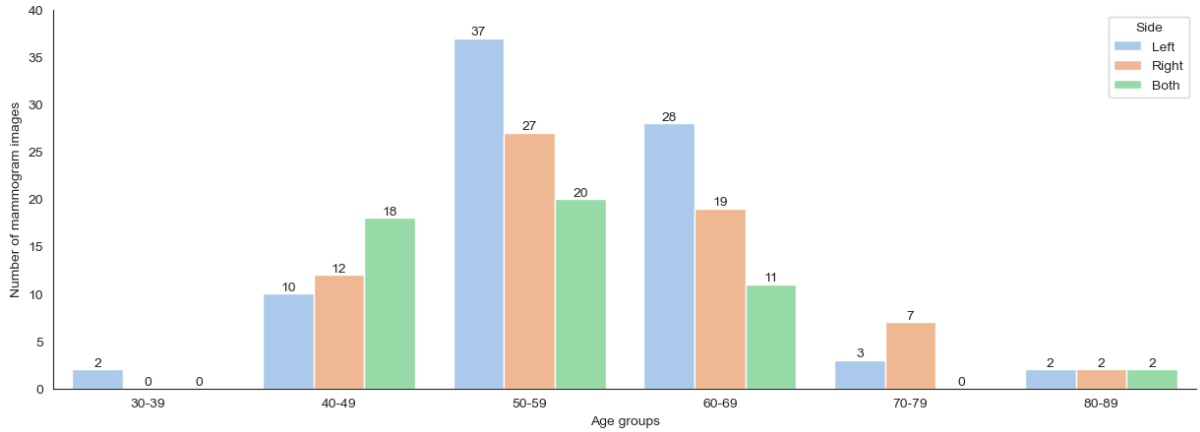


Figure 4.22: Number of mammogram images versus side and age group.

Usually, the mammogram screening device takes images of the breast in different orientations with two different views as mentioned earlier. These views are namely Mediolateral Oblique (MLO) and Cranial Caudal (CC). The MLO provides an optimal visualization of the breast's lateral side rather than the CC view. Figure (4.23) displays the number of images in both CC and MLO views in the sample. In this sample, the number of images in the CC views is greater than the numbers in the MLO views. Moreover, the CC view represents 54% of the sample, while the MLO view constitutes 46%.

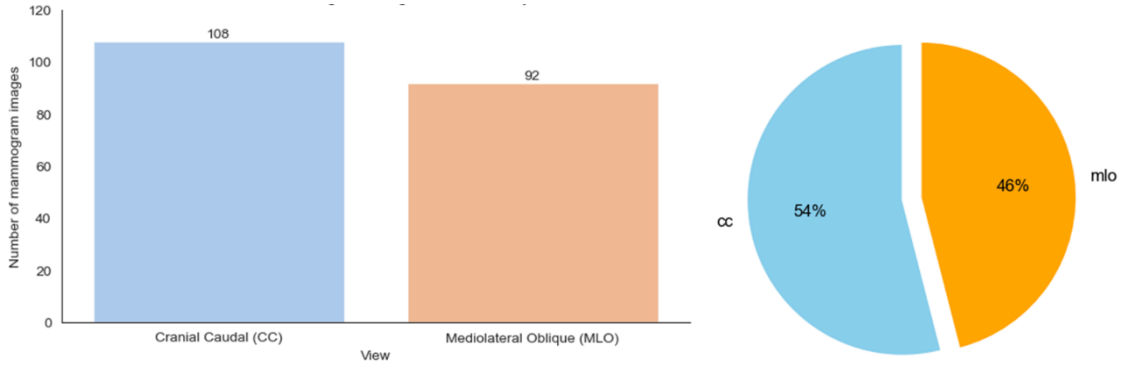


Figure 4.23: Number and percentage of the 'view' feature.

Likewise, Figure (4.24) below shows the number of images taken from the CC and MLO views and how they are distributed between the normal and abnormal classes. In the Abnormal class, the numbers of the CC and MLO images are almost equal where there are 51 and 49 images for the CC and MLO respectively. Comparatively, in the normal sample, there is a noticeable difference in the numbers, as the number of images in the CC view was higher and had 57 images while the MLO view had 43 images.

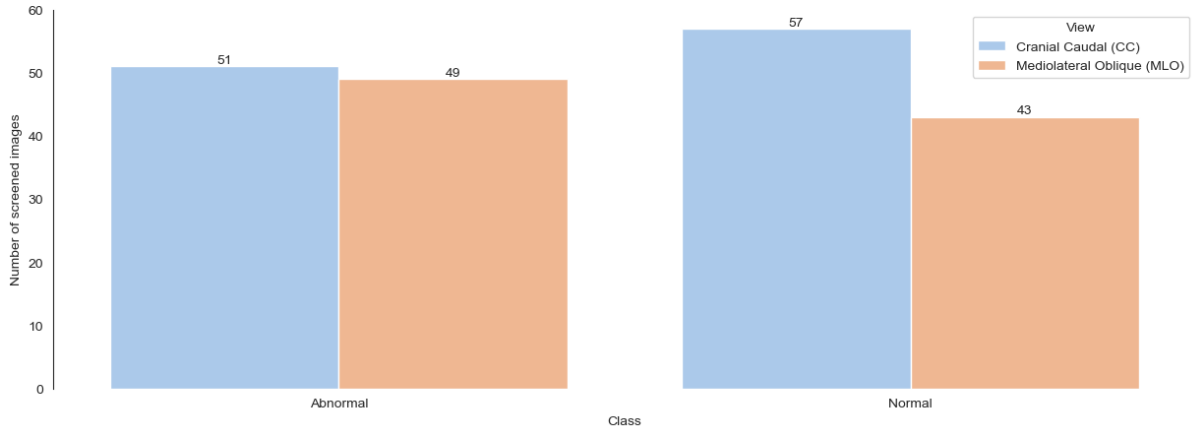


Figure 4.24: Number of screened images versus view and class.

Finally, Figure (4.25) compares the number of mammogram images in CC and MLO views across the age groups. The age group [50-59] has the highest number of images in both views. In addition, the number of images in both views is equal or has a small difference in age groups [30-39], [60-69], [70-79], and [80-89], while there is a significant difference in the numbers in ages between 40 to 59.

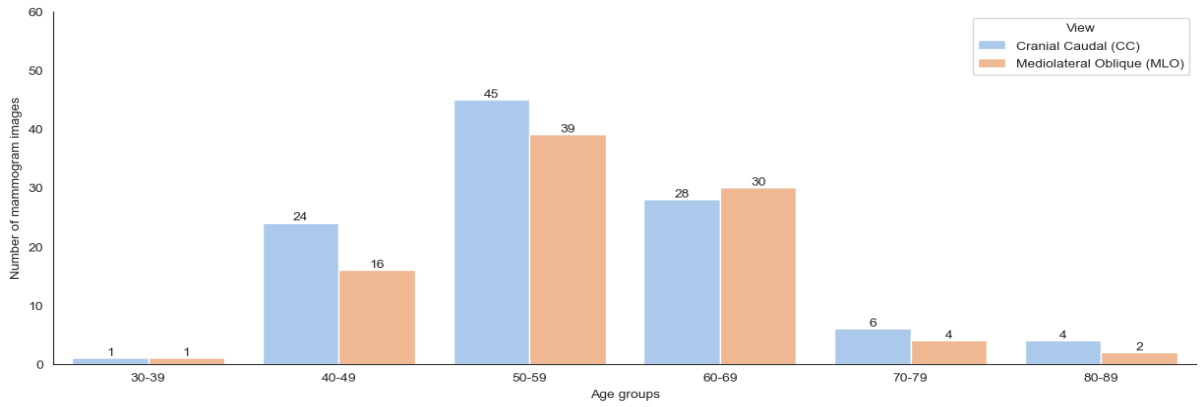


Figure 4.25: Number of mammogram images versus view and age group.

4.4 Medical History Reports Exploratory

This section illustrates some figures about the medical reports that were collected. The total number of medical reports is 231 reports and were all analyzed. Figure (4.26) shows the number of reports distributed by BI-RADS. The x-axis represents the BI-RADS scores and the y-axis represents the number of reports. It can be seen that there

are 71 reports related to BI-RAD 1, while there are 40, 39, and 41 related to BI-RADS 2, 3, and 4 respectively. Moreover, few reports related to BI-RADS 5 and 6. The majority of the reports are related to BI-RADS 1 to 4.

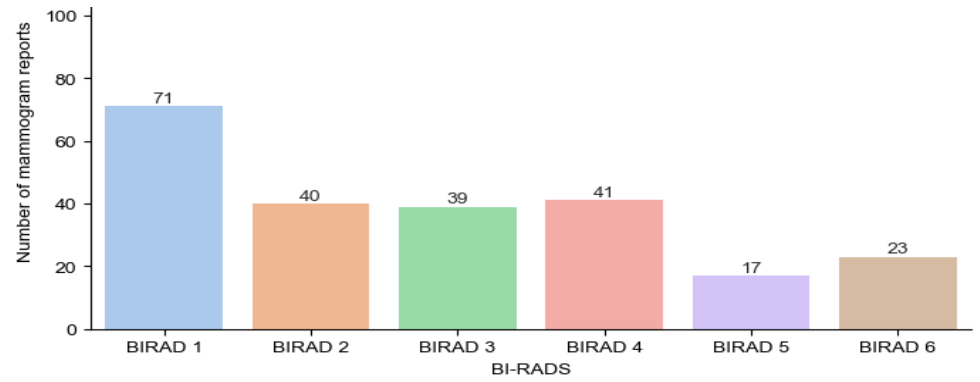


Figure 4.26: Number of medical reports distributed by BI-RADS.

Similarly, the following figure shows the marital status of the women belonging to these reports. In Figure (4.27), the x-axis represents the marital status and the y-axis represents the number of reports. Most of the reports are related to married women which represent approximately 82% of the total. In addition, there are a few reports related to women of other marital status. There are 7% related to single women, 9% to widowed women, and almost 0% to divorced women.

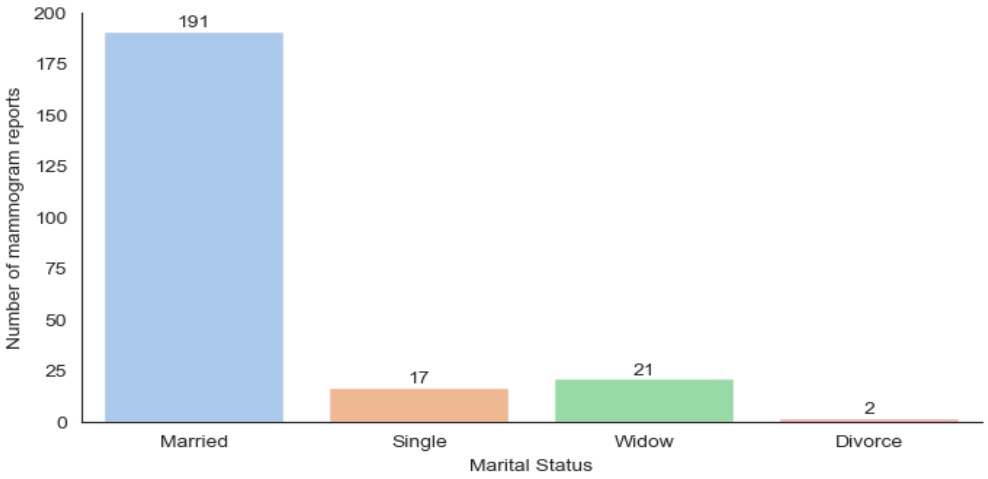


Figure 4.27: Number of medical reports distributed by marital status.

By the same token, Figure (4.28) presents the referral status where the x-axis indicates the source of referral. There are 87 reports related to women who were referred by the MoH to do the screening test. Moreover, there's a good number of reports for women who decided to do the mammogram screening by themselves. Likewise, there are 73 reports for women who were referred by care providers from outside the MOH, it could be from private centers or hospitals. Additionally, 2 women were referred by the UN facilities.

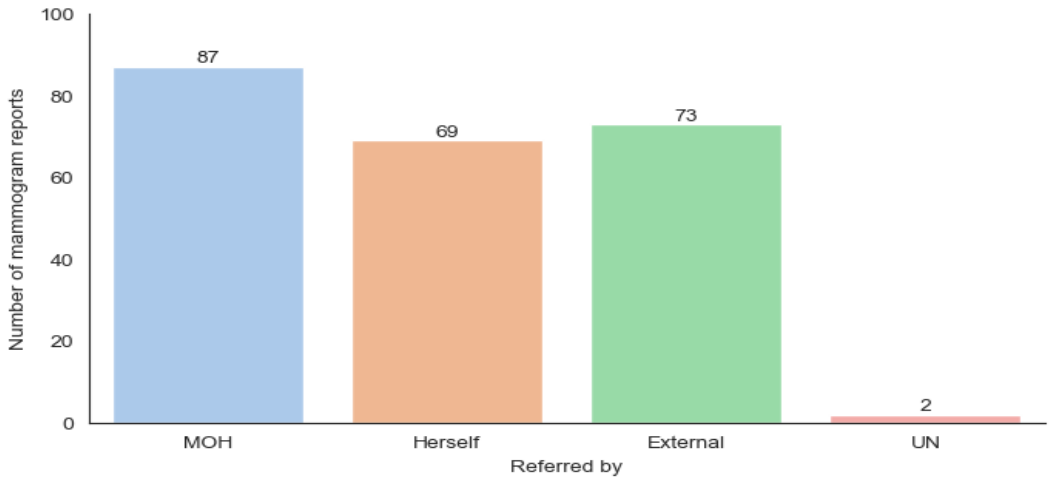


Figure 4.28: Number of medical reports distributed by referral status.

Figure (4.29) below illustrates the answers to some yes/no questions that concern breast cancer symptoms, family history of diseases, previous operations, and previous problems. The x-axis represents the question, while the y-axis represents the number of reports. It can be seen that most of the answers are No, but there are questions with Yes answers on the right side. Furthermore, there are 158 out of 231 women have done mammogram screening previously. In addition, 124 women have other diseases such as diabetes and hypertension which are high-risk factors for developing breast cancer. Besides, 44 women have a family history of breast cancer or other cancers. Also, some

women have symptoms like breast pain, tenderness, breast discharge, lumps, and masses.

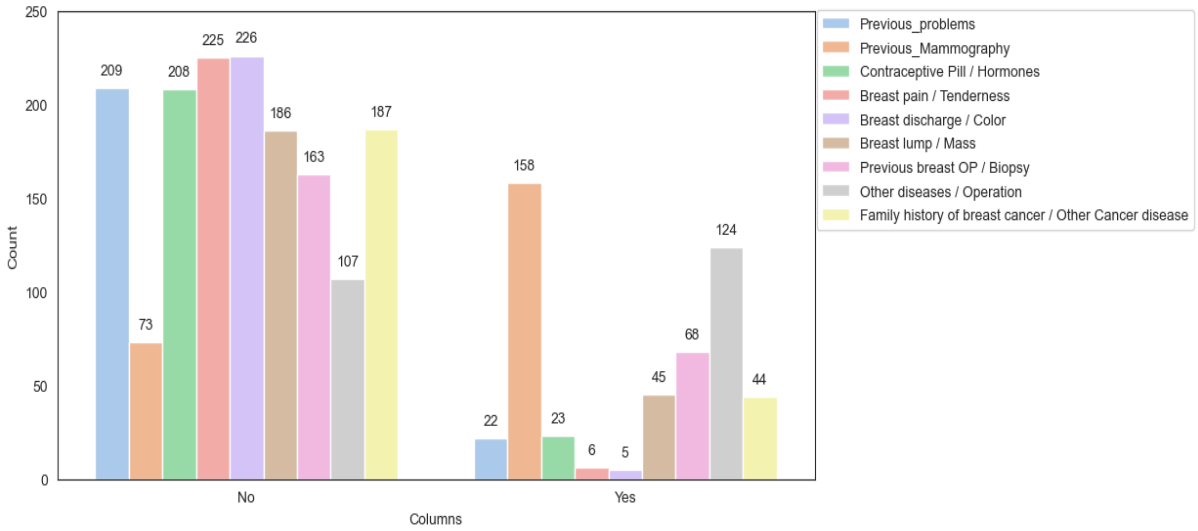


Figure 4.29: Count 'Yes' values across multiple columns.

In the same way, Figure (4.30) shows the answers for age at first menses. The x-axis shows the age in years and the y-axis shows the number of reports. The majority of women had their menses at the age of 13. Moreover, few women had their menses at ages older or younger than 13 years old.

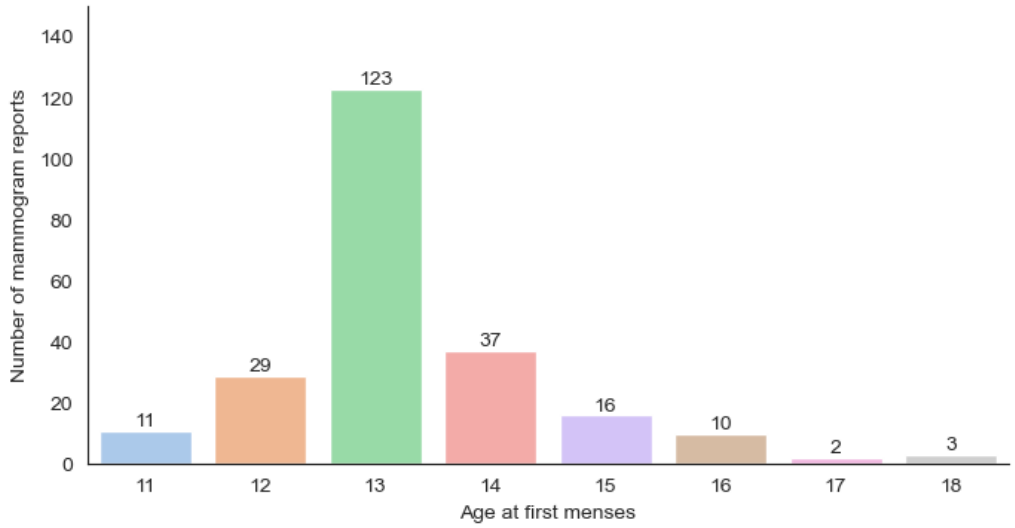


Figure 4.30: Number of medical reports distributed by age at first menses.

Similarly, Figure (4.31) shows the age at marriage in the x-axis and the number of reports in the y-axis. The majority of women got married at ages 15 to 19 years old and represent 58% of the reports. Moreover, 26% got married at ages 20 to 24, and 9% at ages 25 to 29. Also, a few women got married at younger ages less than 15 or older than 29.

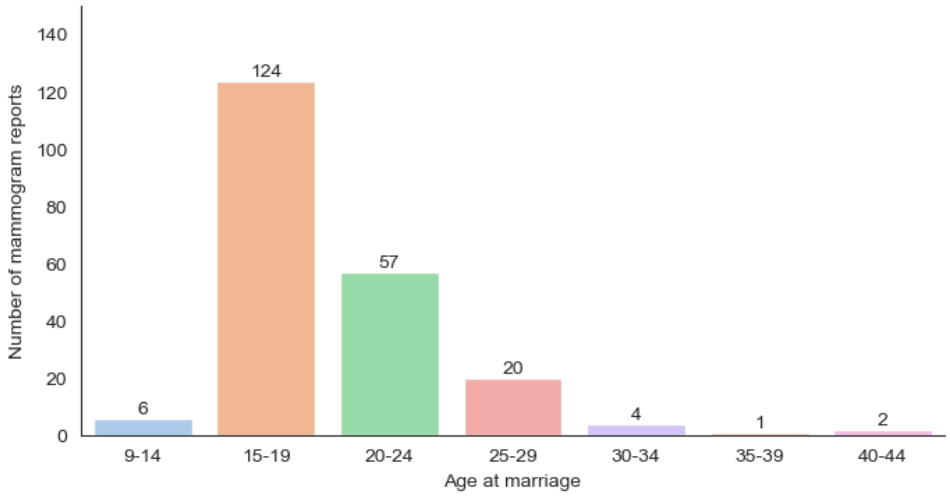


Figure 4.31: Number of medical reports distributed by age at first marriage.

Furthermore, Figure (4.32) is associated with the previous figure where the women had their first baby in the same age group that they got married in. Most of the women had their first delivery at ages 20 to 24, which represents 48% of the reports. In addition, 66 women had their first baby at ages 15 to 19, and 34 women delivered for the first time at ages 25 to 29. Additionally, 11 women had their first child at ages older than 30, and no women had their first born at ages younger than 15.

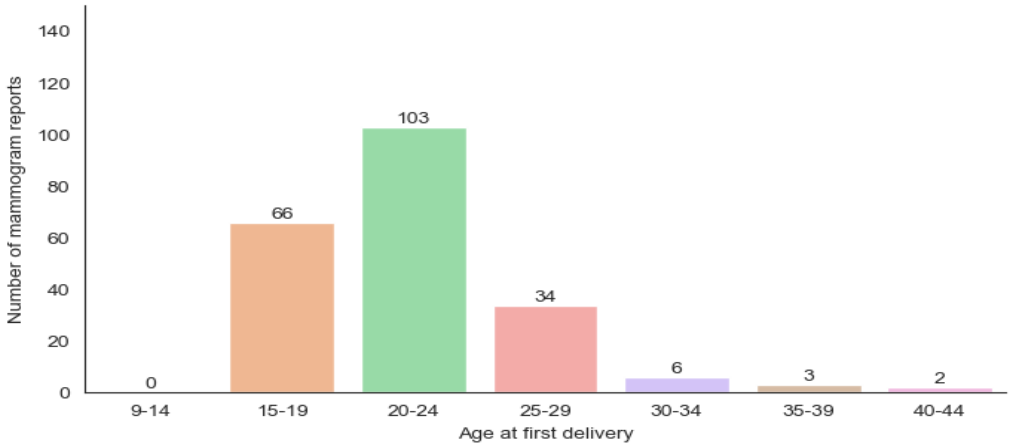


Figure 4.32: Number of medical reports distributed by age at first delivery.

Figure (4.33) shows the number of children born to women who had mammograms and whose reports were collected. 63% of the reports belong to women who have given birth to 5 to 8 children. Also, there are 68 reports belonging to women who have given birth to less than four children, and 11 reports to women who delivered from 9 to 12 children.

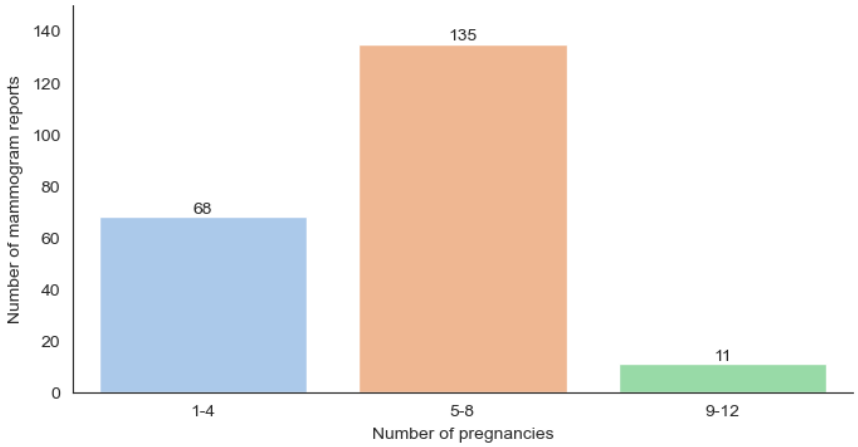


Figure 4.33: Number of medical reports distributed by number of pregnancies.

In the same way, Figure (4.34) depicts the menses status where the x-axis represents the status and the y-axis shows the number of reports. More than half of the reports related to women in the menopausal stage representing 66%. Additionally, 76 out of 231 reports

representing 33% are for women who have regular menses. In contrast, only 3 reports related to women who have irregular menses.

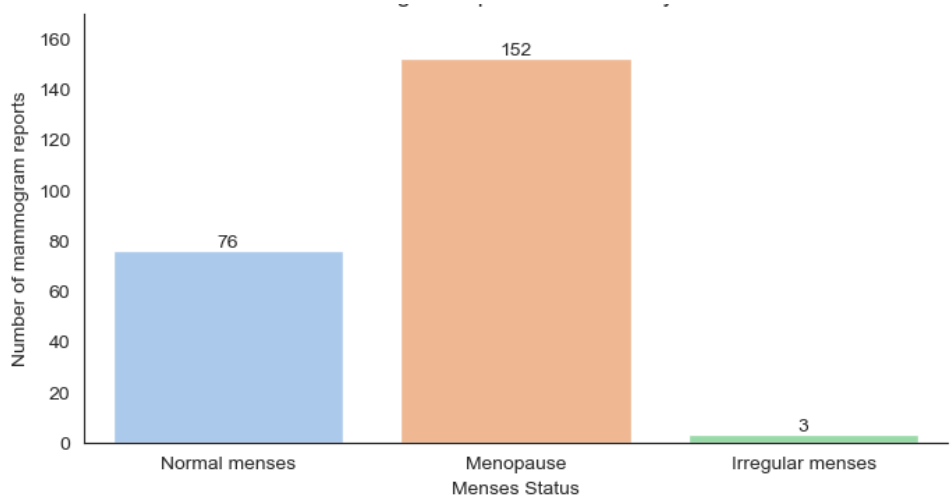


Figure 4.34: Number of medical reports distributed by menses status.

Finally, Figure (4.35) shows women's weights divided into five categories. The weights of women recorded in the reports vary, but most women's weights range between 60 and 79 kilograms, and this category constitutes 58%. There are also 62 reports of women weighing between 80 and 99 kilograms, and this group suffers from excess weight. Furthermore, there are only 16 women whose weight is more than 100 kilograms, and this group suffers from obesity. In addition, 7% of women weigh between 40 and 59 kilograms.

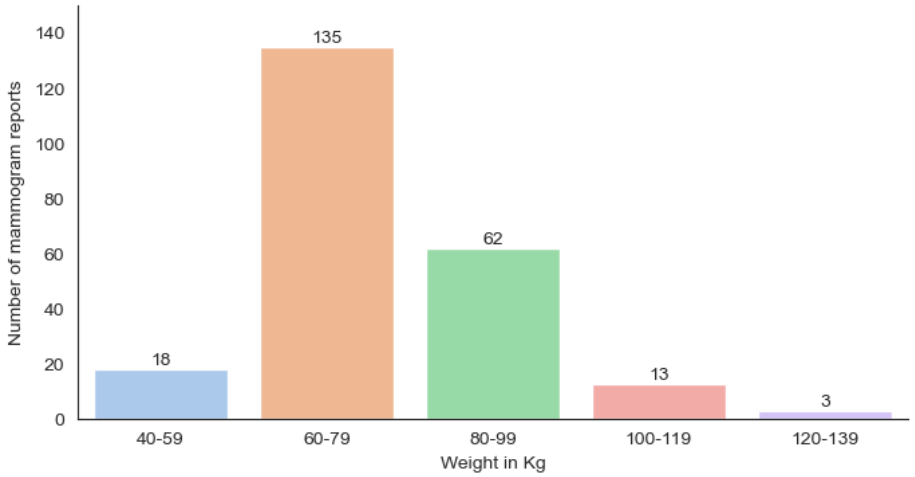


Figure 4.35: Number of medical reports distributed by weight.

4.5 Summary

In conclusion, this chapter provided different visualisation idioms to explore and analyze the data and uncover insights. The original dataset was visualized in addition to the sample dataset of normal and abnormal. Lastly, many visuals were utilized to analyze the medical reports, and the results were illustrated. The next chapter will include the methodology followed in this thesis, selected models, and how they were utilized to address the goals of the thesis.

Chapter Five: Proposed Methodology

5.1 Introduction

This chapter will present the proposed methodology used in this thesis. The modelling phase is when the cleaned and processed data is fed into the models to get the output predictions. In this thesis, two approaches were proposed to detect breast abnormalities in mammogram images. In addition, the selected models and some parameters like learning rate and optimizer were presented. Lastly, the evaluation methods used to measure the models' performance were illustrated. Figure (5.1) illustrates the methodology workflow which will be discussed in detail in the coming sections. The first three phases are discussed in chapters three and four. This chapter focuses on the phases four to six.

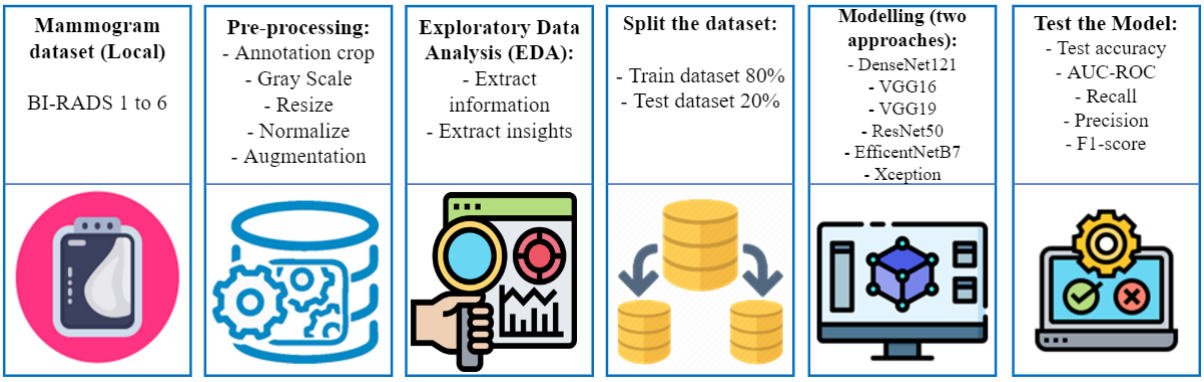


Figure 5.1: Research methodology conducted in this thesis.

5.2 Proposed Approaches

In this thesis, two approaches were presented for detecting breast abnormalities in mammogram images. Both approaches are binary classifications. In the first approach, the sample dataset of normal and abnormal categories was used, while in the second approach, the BI-RADS scores pairs were used. This section will present these approaches in detail.

5.2.1 First Approach: Normal and Abnormal Classification

In the first approach, a mammography dataset of size 200 images was used. The dataset was categorized into two classes, namely 100 images in the normal class and 100 in the abnormal. Moreover, the augmented technique was implemented in the data to enlarge it. After augmentation, the result was creating a dataset of size 2000 (1000 Normal images, 1000 Abnormal images). After that, the augmented dataset was split into an 80% training set and a 20% testing set. Moreover, the training set was fed into the selected models which will be discussed later. After that, the models were trained and tested on the testing set. The models were evaluated using many performance measures like accuracy, precision, and recall. Figure (5.2) illustrates the first approach workflow.

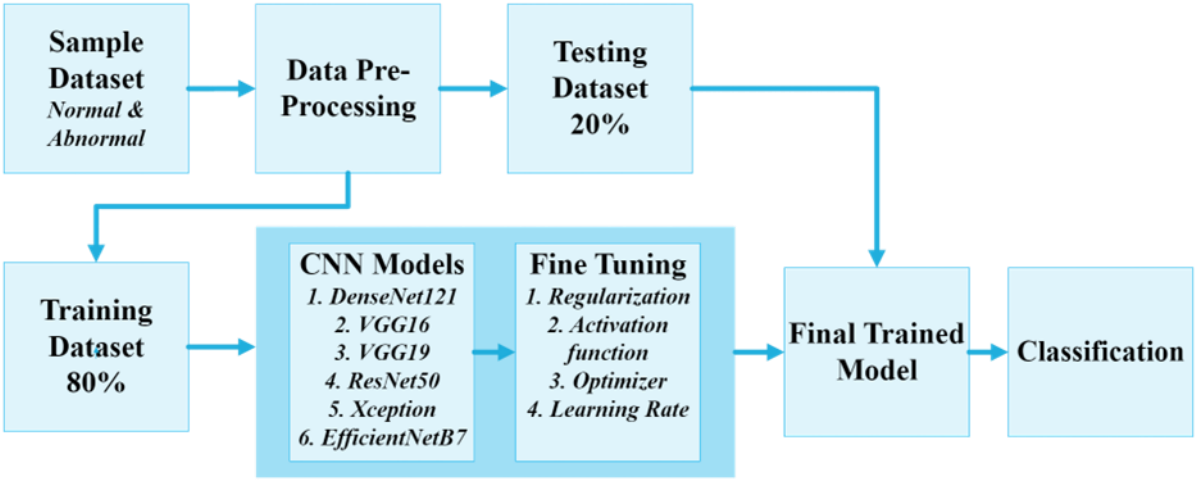


Figure 5.2: First approach workflow.

5.2.3 Second Approach: BI-RADS Classification

In the second approach, the classification was based on the BI-RADS scores from 3 to 6. The BI-RADS 1 and 2 were excluded because there is a poor chance of having abnormalities in the images related to these scores. Figure (5.3) shows the workflow of this approach where there are six subsets to be fed into the models. Each one of the subsets is of size 200 and includes two BI-RADS classes. After splitting each set into

training and testing, the training set is fed into the 6 models, while the test set is used for evaluation. In this approach, the parameters were not tuned because the models were trained using the best values obtained in the previous methodology. Lastly, an evaluation was done using different performance measures such as confusion matrix, f1-score, and many more. The results obtained from this approach will be discussed in chapter six.

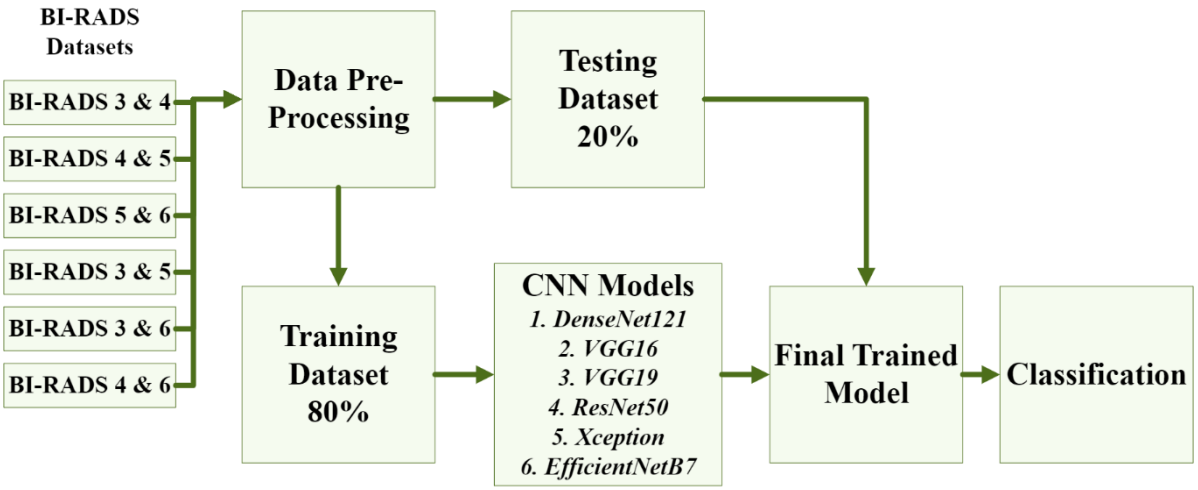


Figure 5.3: Second approach workflow.

5.3 Software and Libraries

In this thesis, the Jupyter Notebook platform and Python language were used to conduct the practical side of this thesis. Several libraries in Python were used in the pre-processing and modelling stages. These libraries include built-in codes and functions to help the users in conducting various experiments without having to write the code from scratch to save time and effort. Moreover, there are a lot of libraries in Python for different tasks such as classification, prediction, regression, image processing, visual analysis, and other tasks.

The following libraries were used in this thesis:

- **OS**: This library allows the users to interact with the operating system. These interactions can be to create, upload, and manage files. In addition, it can be used to manage input, outputs, directories, and set environment variables.
- **OpenCV**: Open-Source Computer Vision Library; it's an open-source library for machine learning. It was originally developed by Intel and has more than 2500 algorithms for computer vision and image processing to recognize, identify, and track objects. This library has a Python interface along with other interfaces and it's supported in many operating systems such as Windows and Linux.
- **NumPy**: Numerical Python; An open-source and fundamental library to handle single and n-dimensional arrays and scientific computing in Python. It's also used to deal with numerical data because it offers comprehensive mathematical functions. Moreover, it contains functions to facilitate statistical analysis. It can also integrate with other libraries like SciKit-Learn and SciPy.
- **Matplotlib**: An open-source library written in Python in addition to other languages, created by John Hunter. It is an essential library that has comprehensive scripts for data visualization. Furthermore, it's considered a utility to create, explore, explain, and analyze different visuals such as bars, charts, histograms, and many more.
- **TensorFlow**: It's an open-source library for AI developed by the Google Brain team. It focuses mainly on Deep Learning and Neural Networks tasks. It's used widely to build and train models. It can also be integrated with other libraries to enhance performance.
- **Keras**: A deep learning and open-source library that is considered the interface of the TensorFlow library. It's distinguished for user friendly and extensibility.

Besides Neural Networks, Recurrent and Convolutional Neural Networks are also supported in this library. Moreover, model layers such as pooling, dropout, and normalization are managed in this package.

- **Sklearn**: Scikit-Learn; an open-source library for machine learning and data modeling. Supports supervised and unsupervised learning. It's built on top of other libraries such as NumPy and Matplotlib. Additionally, it's commonly used in classification, regression, clustering, dimensionality reduction, model selection, and pre-processing applications.

5.4 Data Split

This stage is important in the modelling process. As shown previously in Figure (5.1), this stage precedes the modeling stage. Usually, datasets are split into training and testing datasets. The training dataset should be fed into the model in the training phase and then the model is evaluated using the observations in the testing dataset. The latter dataset is used to make sure that the learning process is successful. Moreover, datasets can be split into three subsets namely training, validation, and testing subsets. The validation subset is usually used to tune model hyperparameters.

In the same way, there is no standard protocol or guideline on how to split the data. However, there are common percentages to split such as 70% for training and 30% for testing, or 80% for the training and 20% for the testing. It depends on many reasons like the size of the original dataset or the sample. In this thesis, the data was split into two main subsets; 80% for the training subset, and 20% for the testing subset. My dataset is considered small so I didn't split it into three subsets. This division was followed in both approaches presented earlier.

5.5 Selected Models

In this thesis, a transfer learning approach was followed and six Deep Learning pre-trained models were used. Transfer learning is a machine-learning concept that utilizes gained knowledge from a task to enhance performance in another different but related task. The chosen models are DenseNet12, VGG16, VGG19, ResNet50, and EfficientNetB7. These models were selected based on their performance and evaluation results in prior studies for similar tasks. The following subsections will provide an overview of each model.

5.5.1 DenseNet121

Densely Connected Convolutional Network (DenseNet) is one of the CNN models that achieved high performance in the literature compared to other pre-trained models. It's constructed from 120 convolutions and 4 AvgPool and includes four components which are Connectivity, DenseBlocks, Growth Rate, and Bottleneck layers. DenseNet takes the input from all previous layers' connected outputs to the dense block layer (Ayad & Abdulmunim, 2023). Figure (5.4) presents the schematic layout of the DenseNet (Huang et al., 2017). The connectivity pattern is the main idea behind the DenseNet structure where the feature maps from previous layers are concatenated onto the inputs of future layers. Continually concatenating results in very deep inputs. There are many activation functions such as ReLU are utilized to increase the non-linearity in the pooling layers when the feature maps are fed into the coming layers. Equation (1) illustrates the mathematical formula for this schema.

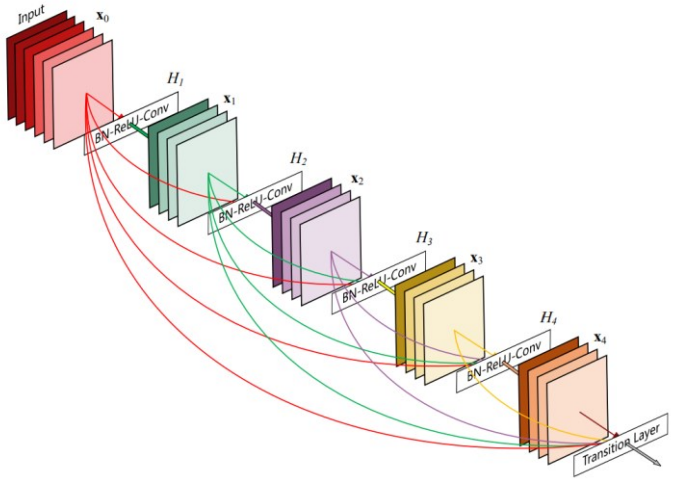


Figure 5.4: DenseNet Schematic layout (Huang et al., 2017)

$$x_l = H_l([x_0, x_1, \dots, x_{l-1}]) \quad (1)$$

Where l^{th} is the layer that receives the concatenated feature maps $x_0,x_1,\dots,x_{l-1}$ from all previous layers. H_l is a composite function of operations like batch normalization (BN), a rectified linear unit (ReLU), and a 3 × 3 convolution (conv) that generates the K levels for the mapping features in the coming layer which determines the growth rate in the network. The growth rate K is calculated in Equation (2).

$$k_l = k_0 + k * (l - 1) \quad (2)$$

Where k_0 is the number of channels in the input layer.

5.5.2 VGG16

VGG stands for Visual Geometry Group. It's a deep-learning model based on CNN architecture used in image processing, detecting objects, and image classification. It has many layers that contain convolutional, pooling, and fully connected layers. In general, it has 13 convolutional layers and 3 fully connected layers as shown in Figures (5.5) and (5.6).

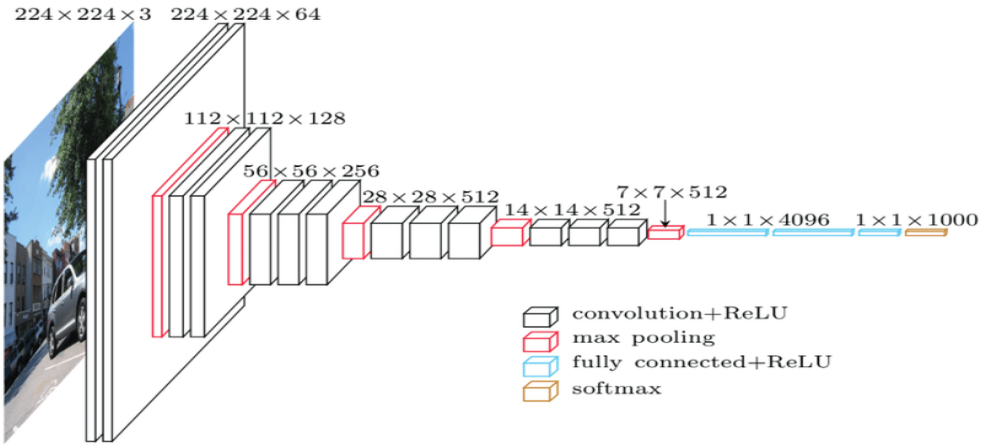


Figure 5.5: VGG16 layers (Sriram et al., 2022)

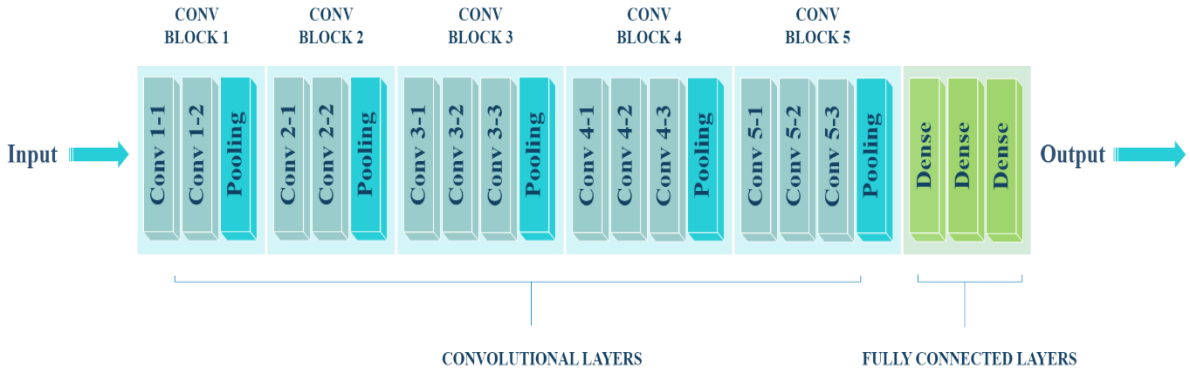


Figure 5.6: VGG16 Architecture.

VGG16 architecture consists of the following layers:

- **Input layer**: the input image size should be 224×224
- **Convolutional layers**: VGG16 has 13 convolutional layers that have filters of size 3×3. The number of filters differs from one layer to the other. The first set of layers has 64 filters, the following set has 128, and 256, 512 filters in the following layers.
- **Max pooling layers**: it follows each set of convolutional set of layers. It's utilized to reduce dimensions created from the previous convolutional layer.

- **Fully connected layers**: there are 3 fully connected layers with the ReLU activation function in the VGG16 model. The first two layers have an output size of 4096, while the last layer's output size is 1000. Moreover, the last layer uses the softmax activation function to the output for classification.

5.5.3 VGG19

VGG19 is also a CNN model that is commonly used with image classification and detection. The architecture is the same and the only difference is the number of layers. VGG16 is constructed from 13 convolutional layers and 3 fully connected layers while in VGG19 there are 16 convolutional layers and 3 fully connected (Hindarto et al., 2023). The image shape should be 224×244 in both models. VGG19 architecture consists of the same layers that exist in VGG16 because they both were built from the same base architecture but with an additional 3 convolutional layers in VGG19. The layers schematic overview in VGG19 is shown in Figure (5.7) below. Note that the extra layers are highlighted in yellow which are part of the convolutional layers.

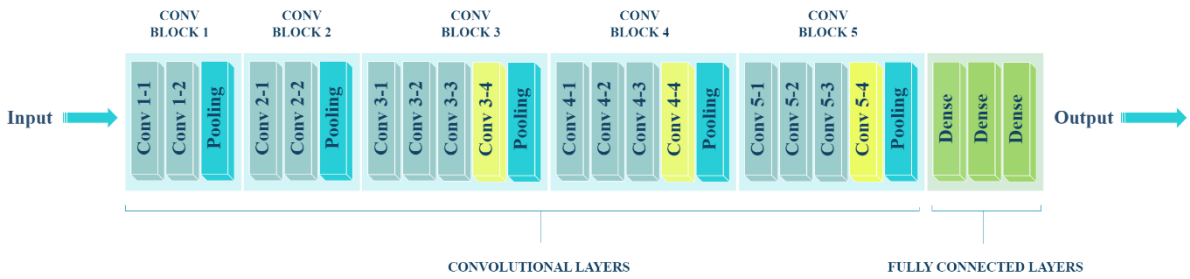


Figure 5.7: VGG19 Architecture.

In conclusion, VGG16 and VGG19 are types of VGG networks named ConvNet. The ConvNet has five potential configurations listed in Figure (5.8) below. The difference between these models is the number of layers.

ConvNet Configuration					
A	A-LRN	B	C	D	E
11 weight layers	11 weight layers	13 weight layers	16 weight layers	16 weight layers	19 weight layers
input (224 × 224 RGB image)					
conv3-64	conv3-64 **LRN**	conv3-64 **conv3-64**	conv3-64 conv3-64	conv3-64 conv3-64	conv3-64 conv3-64
maxpool					
conv3-128	conv3-128	conv3-128 **conv3-128**	conv3-128 conv3-128	conv3-128 conv3-128	conv3-128 conv3-128
maxpool					
conv3-256 conv3-256	conv3-256 conv3-256	conv3-256 conv3-256	conv3-256 conv3-256 **conv1-256**	conv3-256 conv3-256 **conv3-256**	conv3-256 conv3-256 conv3-256 **conv3-256**
maxpool					
conv3-512 conv3-512	conv3-512 conv3-512	conv3-512 conv3-512	conv3-512 conv3-512 **conv1-512**	conv3-512 conv3-512 **conv3-512**	conv3-512 conv3-512 conv3-512 **conv3-512**
maxpool					
conv3-512 conv3-512	conv3-512 conv3-512	conv3-512 conv3-512	conv3-512 conv3-512 **conv1-512**	conv3-512 conv3-512 **conv3-512**	conv3-512 conv3-512 conv3-512 **conv3-512**
maxpool					
FC-4096					
FC-4096					
FC-1000					
soft-max					

Figure 5.8: ConvNet Configuration *(Simonyan & Zisserman, 2014)*

5.5.4 ResNet50

Residual Network; ResNet is a CNN model that is used to detect objects. It was the result of adding residual structure to the CNN to fix the gradient degradation and disappearance issue that frequently occurs in the training process (Du et al., 2023). The added residual connections enable the network to learn from the functions to associate inputs with desired outputs. ResNet50 contains 50 layers that are distributed over four types of layers which are convolutional, pooling, fully connected, and shortcut (or skip connection) (Li et al., 2023). The shortcut means that the model can skip some layers. It can perform efficiently even with a large number of layers. Figure (5.9) shows a network consisting of one residual block with a skip connection.

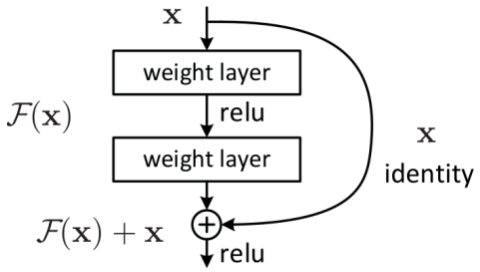


Figure 5.9: Residual block with a skip connection (He et al., 2015)

Residual Networks have different variants of ResNet configurations including ResNet50 as shown in Figure (5.10). In addition, the architecture of ResNet50 layers is presented in Figure (5.11).

layer name	output size	18-layer	34-layer	50-layer	101-layer	152-layer
conv1	112×112	7×7, 64, stride 2				
conv2_x	56×56	3×3 max pool, stride 2				
		[3×3, 64; 3×3, 64] ×2	[3×3, 64; 3×3, 64] ×3	[1×1, 64; 3×3, 64; 1×1, 256] ×3	[1×1, 64; 3×3, 64; 1×1, 256] ×3	[1×1, 64; 3×3, 64; 1×1, 256] ×3
conv3_x	28×28	[3×3, 128; 3×3, 128] ×2	[3×3, 128; 3×3, 128] ×4	[1×1, 128; 3×3, 128; 1×1, 512] ×4	[1×1, 128; 3×3, 128; 1×1, 512] ×4	[1×1, 128; 3×3, 128; 1×1, 512] ×8
conv4_x	14×14	[3×3, 256; 3×3, 256] ×2	[3×3, 256; 3×3, 256] ×6	[1×1, 256; 3×3, 256; 1×1, 1024] ×6	[1×1, 256; 3×3, 256; 1×1, 1024] ×23	[1×1, 256; 3×3, 256; 1×1, 1024] ×36
conv5_x	7×7	[3×3, 512; 3×3, 512] ×2	[3×3, 512; 3×3, 512] ×3	[1×1, 512; 3×3, 512; 1×1, 2048] ×3	[1×1, 512; 3×3, 512; 1×1, 2048] ×3	[1×1, 512; 3×3, 512; 1×1, 2048] ×3
	1×1	average pool, 1000-d fc, softmax				
FLOPs		1.8×10^9	3.6×10^9	3.8×10^9	7.6×10^9	11.3×10^9

Figure 5.10: ResNet Variants (He et al., 2015)

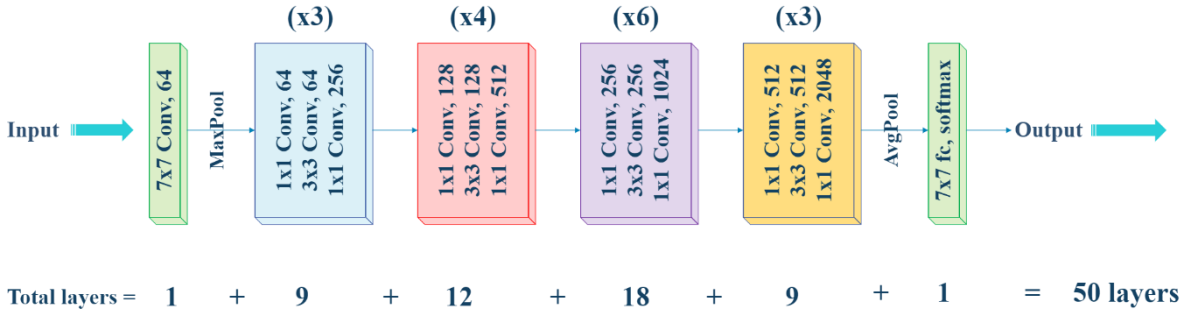


Figure 5.11: ResNet50 Architecture.

ResNet50 in general consists of the following components:

- **Convolutional layers** to extract features from the input object by applying filters. The number of filters varies from one convolutional layer to another (He et al., 2015).
- **Batch Normalization** component used for performance improvement which is usually applied after the conv layer or the fully connected layer and before the activation function (He et al., 2015).
- **ReLU Activation** is a rectified linear unit function. It's responsible for transforming the summed weights into the output. If the input is positive, the output will be equal to the input, otherwise, the output will be zero. The ReLU formula is shown in the following equation (He et al., 2015).

  $$ReLU(x) = max(0, x) \quad (3)$$

- **Max pooling** layer which is used for spatial dimensions reduction of the output mapped features by the convolutional layers. Subsequently, prevent overfitting (He et al., 2015).
- **Fully Connected layers** that are known as dense layers. This type of layer connects all the neurons in the current layer with the all neurons in the previous layer. It's applied usually in the final stages or blocks of the model and is responsible for making the final classification (He et al., 2015).
- **The Average Pooling** layer is used to calculate the feature map patches' average value to down-sample the feature map and is used after the convolutional layers (He et al., 2015).

5.5.5 EfficientNetB7

EfficientNets is a CNN family of models of eight variants namely EfficientNetB0 to EfficientNetB7 that obtained better efficiency and accuracy than the previous CNNs. It was developed by scaling up the base EfficientNet architecture. The scaling method was utilized to scale up depth, width, and resolution dimensions rather than dependency through a compound coefficient to enhance performance with fewer parameters as shown in Figure (5.12). The depth dimension indicates the number of layers, the width represents the number of channels or the feature maps in the layer, and the resolution is the input image (width × height). EfficientNetB7 in particular, achieved top-1 accuracy on ImageNet of all EfficientNet families (Tan & Le, 2019).

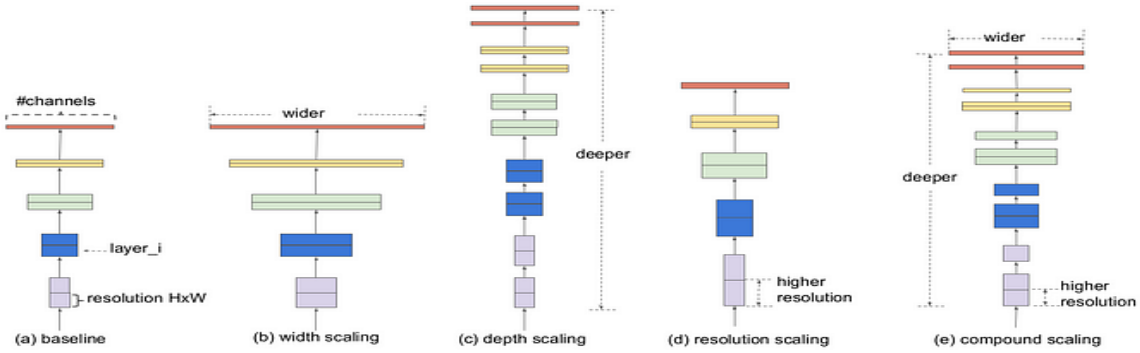


Figure 5.12: Idea behind EfficientNet *(Tan & Le, 2019)*

The base architecture layers consist of the MBConv block which is a combination of squeeze-and-excitation (SE) and inverted residual blocks. Additionally, the overall architecture of EfficientNetB7 has seven blocks as shown in Figure (5.13). Number of filters varies from one layer to another.

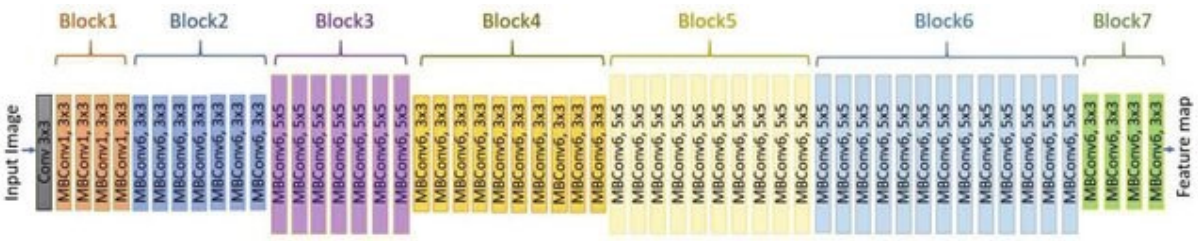


Figure 5.13: EfficientNet architecture (Baheti et al., 2020)

5.5.6 Xception

The Extreme Inception; Xception was proposed by Google that utilizes CNN structure and relies on depthwise-separable convolution (Deb et al., 2023). It was used in a lot of classification and recognition tasks with high accuracy. According to the official documentation, Xception architecture involves 36 convolutional layers that form the feature extraction base of the network (Chollet, 2017). The standard architecture of Xception included three flows as illustrated in Figure (5.14) below. The first flow (entry) consisted of convolutional layers along with separable convolutional and pooling layers. The middle flow consisted of only separable convolutional layers repeated 8 times, while the final flow (exit) contained separable convolutional, pooling, and fully connected layers. All the convolutional and separable convolutional layers are followed with a batch normalization.

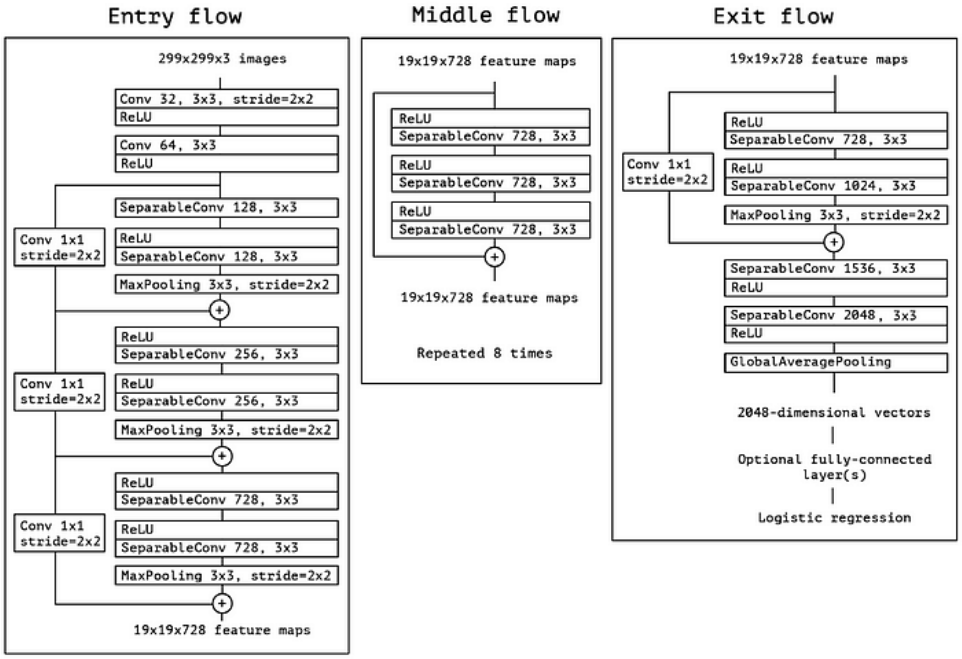


Figure 5.14: The standard architecture of Xception model (Chollet, 2017)

The Xception model consists of the following components (Chollet, 2017):

- **Depthwise Separable Convolutions**: this type of layer separates the spatial and depth dimensions, unlike the traditional convolutions that utilize them simultaneously. By doing this separation, it reduces the computational costs.
- **Separable Convolutions**: the Xception structure is built of many sequential blocks of separable convolutional. Each one of these blocks has a depthwise separable convolution followed by a pointwise convolution to enhance the recognition of complex patterns.
- **Entry and Exit Flow**: these flows are considered the standards of the Xception hierarchy. The first flow is responsible for feature extraction, while the last flow is responsible for tuning them for better prediction or classification.
- **Skip Connections**: the Xception model integrates the skip connections that are used in Residual Networks in its layers to enhance data flow across different layers.

5.6 Regularization

Regularization is a set of techniques to reduce and avoid overfitting or underfitting while training the model to increase its generalizability (Santos & Papa, 2022). Overfitting happens when the model fits exactly or too close to the training data, subsequently making inaccurate predictions or classifications from datasets other than the training dataset. On the other hand, underfitting happens when the model doesn't train enough time to draw the correct relationship between the input and the output. Figure (5.15) shows how the regularization enabled the model to find the best-fit line for the data on a given dataset.

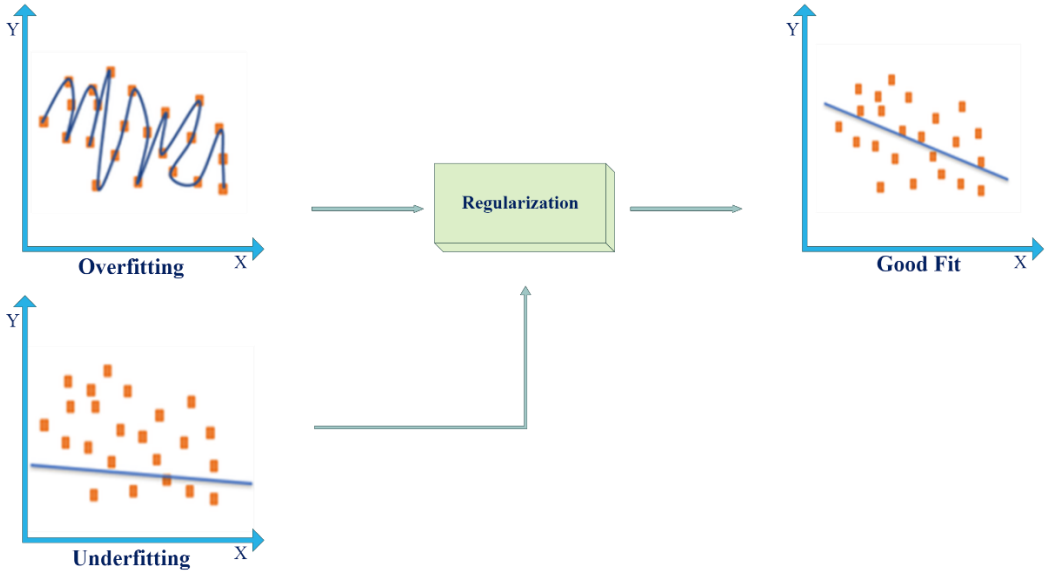


Figure 5.15: Regularization.

There are two main types of regularization namely Lasso Regularization (L1) and Ridge Regularization (L2). The L1 type is a technique that minimizes the high-value, correlated coefficient by penalizing them (Ogutu et al., 2012). It's usually done by adding the absolute value of the sum of the coefficient to the loss function. Comparatively, L2 regularization adds the squared magnitude of the coefficients as the penalty term to the loss function (Ogutu et al., 2012). Moreover, L2 is used in this thesis with the selected models presented in the previous sections based on its effective results presented in the state-of-art. Equations 4 and 5 show the formula of both L1 and L2 methods.

- L1 Regularization:

$$Modified\ Loss\ function = Loss\ function + \lambda \sum_{i=1}^{n} |W_i| \quad (4)$$

- L2 Regularization:

$$Modified\ Loss\ function = Loss\ function + \lambda \sum_{i=1}^{n} W_i^2 \quad (5)$$

Where;

– Loss function is the sum of the squared residuals
– λ is the penalty for the errors
– W is the fit-line slope

In general, Regularization methods are used to prevent overfitting or underfitting, balance bias, and variance, control the model complexity, and generalization.

5.7 Activation Functions

In neural network layers, the neurons are associated with their weight and bias. The activation function is used to control these neurons by deciding which neurons should be activated by adding the bias to their weighted sum (Oostwal et al., 2021). Without an activation function, the network acts as a linear regression model. Subsequently, adding the activation function will introduce the non-linearity to the model which enables it to handle more complex tasks like muti classification.

In addition, there are many types of activation functions depending on the type of task, for example, prediction or classification. In this thesis, two types were used namely, ReLU and Softmax (Oostwal et al., 2021), (Es-sabery et al., 2021).

- **ReLU**: as mentioned in the previous sections, it's the Rectified linear unit. It's commonly used in classification tasks. In this function, the neurons are not activated all at once (Oostwal et al., 2021). That is, the output will be equal to the input if it is positive, otherwise, it converts the output to zero so the neurons will not get activated (Es-sabery et al., 2021). Figure (5.16) and Equation (6) illustrate the ReLU function graphically and mathematically.

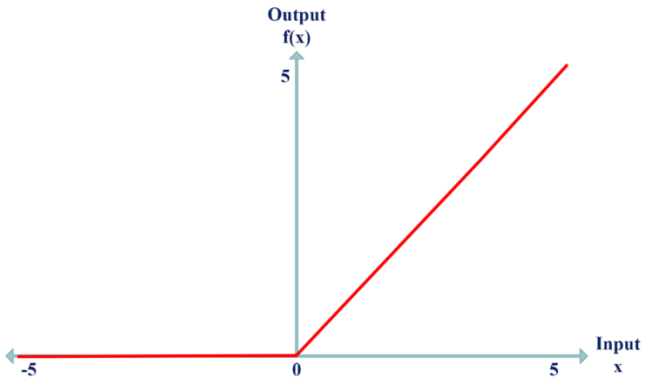


Figure 5.16: ReLU activation function.

$$f(x) = \max(0, x) \tag{6}$$

Where x is the input.

- **Softmax**: it's a type of activation function that is usually used in the final layer of a CNN model in classification tasks to convert outputs into probabilities (Oostwal et al., 2021). The function takes the output vector and exponentiates it using constant *e* and then normalizes it. The normalization is done by dividing by the sum of all exponentials to ensure that the values of the output range between 0 and 1, and the total of summation is 1. Figure (5.17) and Equation (7) present the Softmax function graphically and mathematically.

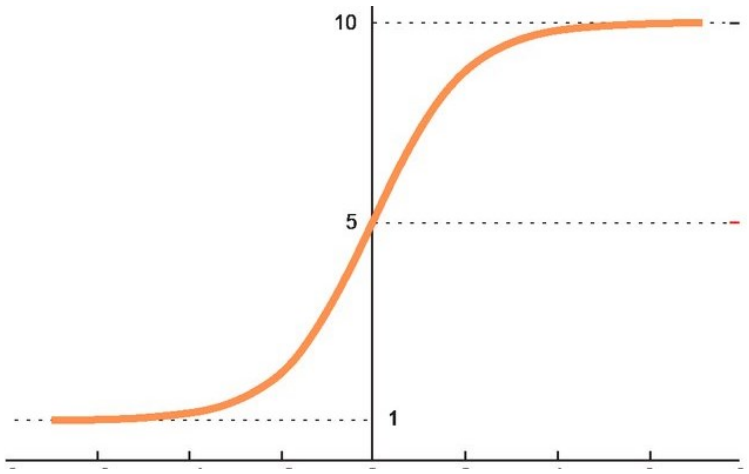


Figure 5.17: Softmax activation function (Es-sabery et al., 2021).

$$Softmax\ function = \frac{e^{z_i}}{\sum_{j=1}^{K} e^{z_j}} \quad (7)$$

Where zi represents the input for the class i

5.8 Optimizer

Optimizers are machine-learning algorithms to minimize the loss function by changing the network attributes such as weights and learning rates (Reyad et al., 2023). The goal of this algorithm is to reach the optimal model parameters to reduce the error function. There are many types of optimizers used in previous studies for different applications. The most popular one is Adam Optimizer which is used in this thesis.

Adam stands for Adaptive Moment Estimation and is used for gradient descent (Reyad et al., 2023). It is a combination of momentum and Root Mean Square Propagation (RMSP) methods (Reyad et al., 2023). Adam doesn't require much tuning, works very fast, and converges rapidly in big datasets while consuming less memory. It outperformed all other optimizers in terms of performance and gradient descent with a significant difference in margin.

5.9 Learning Rate

The learning rate is a hyperparameter that determines the step size to update the weights of the network during the training process. It controls how fast or slow the network moves toward the optimal weights concerning the loss gradient. The learning rate value ranges between 1 and 1×10^{-7} (Apaydin et al., 2020). Figure (5.18) depicts the effect of the learning rate value on the network epochs and loss function.

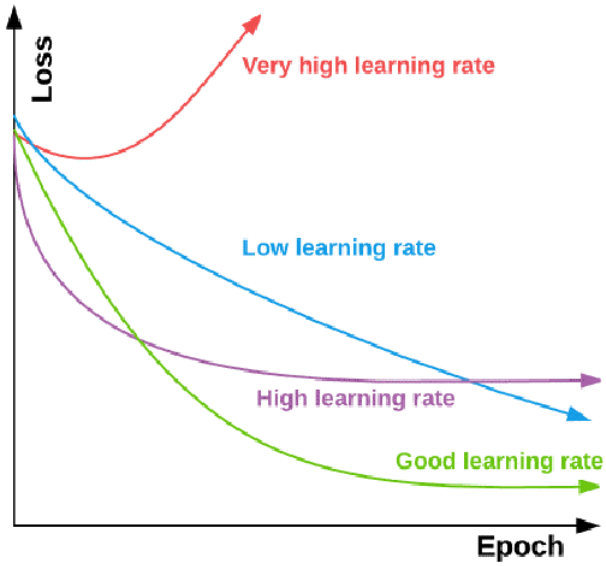


Figure 5.18: Learning rate effects on the loss function (Apaydin et al., 2020)

The effect on loss function and epochs is related to the size of the move or in other words, the value of learning rate. The network can take a long time to reach the optimal weights in case of using low values. On the other hand, the best solution is most probably rejected when using very high values. Moreover, the optimal solution is obtained when the size of the move slows down while the model progresses.

5.10 Evaluation and Performance Measures

Model performance evaluation is the last stage of the model life cycle and is usually evaluated with different measurements on the testing subset to ensure that it achieves the desired goal. The measures were implemented on the testing subset. This section will present the measures used in this thesis to evaluate the effectiveness and performance of the implemented models in detecting breast abnormalities. The metrics are the confusion matrix, accuracy, recall, precision, F1-score, and AUC (Hossin & M.N, 2015).

- **Confusion Matrix**

The confusion matrix is one of the most important measures in the classification models. It's a combination of actual and positive values. The actual is the input, while the predicted is the output classification. Figure (5.19) shows the matrix of these values.

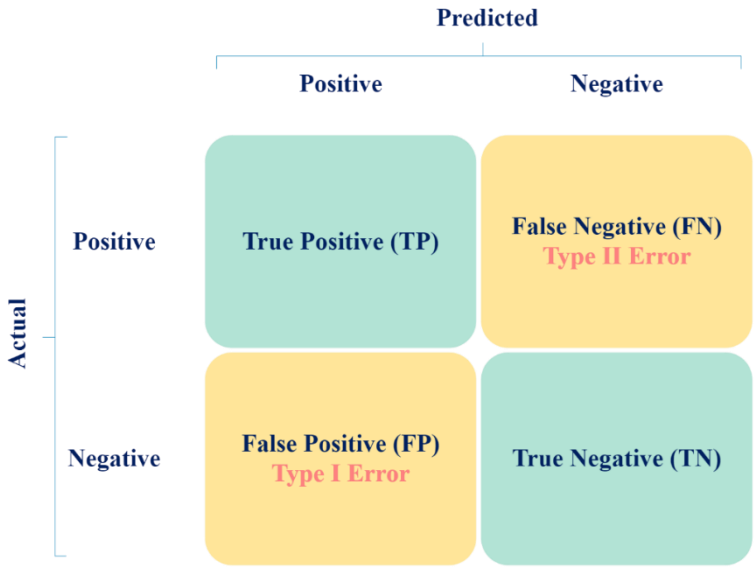


Figure 5.19: Confusion Matrix.

From the figure, it's clear that the matrix consists of four values:

- **TP**: *True Positive*; when the model predicts the positive sample correctly.
- **TN**: *True Negative*; when the model predicts the negative sample correctly.
- **FP**: *False Positive*; when the model predicts the positive sample but it's negative in reality.
- **FN**: *False Negative*; when the model predicts the negative sample but it's positive in reality.

Many performance measures are calculated from these values namely accuracy, recall, precision, F1-score, and ROC. These measures are presented in the following sections along with their formulas.

- **Accuracy**

  Accuracy is the simplest and the most used measure in the classification studies. It measures how often the model correctly predicts the outcome. In general, it represents the ratio of the correct predictions of all predictions. The accuracy formula is shown in Equation (8) below.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} * 100 \quad (8)$$

- **Recall**

  Recall is the metric that measures how often the model identifies "true positives" correctly from all ground truth positives. It's also called the True Positive Rate (TPR). It's used to extract the ratio of true positives to all positives in the ground truth. Moreover, it represents type-II errors which occur when the model accepts the false Null Hypothesis (H_0). The formula is shown in Equation (9).

$$Recall\ (or\ TPR) = \frac{TP}{TP + FN} \quad (9)$$

- **Precision**

  Precision is the ability of the model to predict the positive class correctly. It represents the ratio of "true positive" to all predicted positives. Additionally, it focuses on type-I error which occurs when the model rejects the true Null Hypothesis (H_0) by mistake. The formula is shown in Equation (10).

$$Precision = \frac{TP}{TP + FP} \quad (10)$$

- **F1-score**

The last measure is the F1-score. It's the harmonic mean of precision and recall. It's used in binary and multi-class classifications and ranges between 1 and 0 with 1 being the better score. The formula is shown in Equation (11).

$$F1 - score = 2 * \frac{Precision * Recall}{Precision + Recall} \quad (11)$$

- **ROC - AUC**

ROC stands for Receiver Operating Characteristic, and AUC is the Area Under the Curve. ROC is a graphical plot that plots the true positive rate (TPR) represented in the y-axis versus the false positive rate (FPR) represented in the x-axis as illustrated in Figure (5.20) below. It's an important measure in binary or multi-classification tasks to know if the model can distinguish between classes. In addition, the higher the AUC, the better the model's capability of distinguishing between classes.

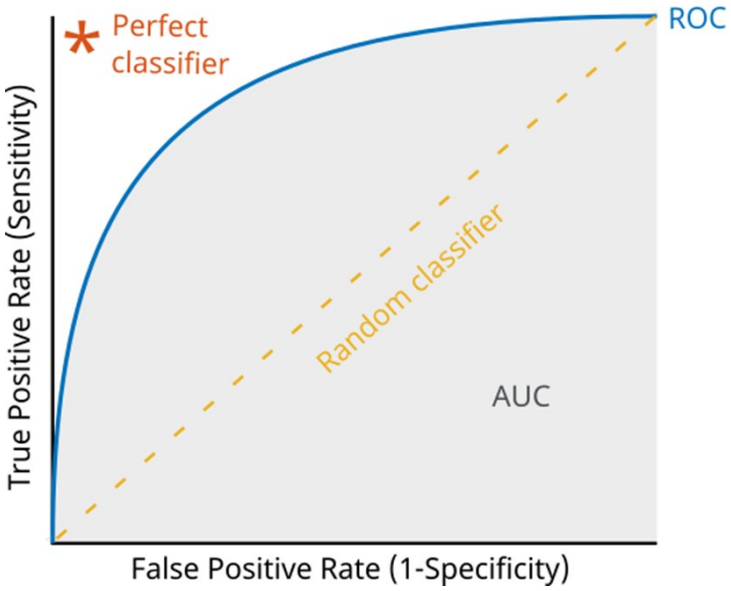


Figure 5.20: ROC Curve.

The True Positive Rate (TPR) is represented by the Recall explained in the previous sections, while the False Positive Rate (FPR) means the probability of falsely rejecting the null hypothesis. The latter is calculated using Equation (12) below.

$$FPR = \frac{FP}{TN + FP} \tag{12}$$

5.11 Summary

In summary, this chapter presented the proposed methodology and approaches in this thesis to diagnose breast abnormalities in mammographies. Previous researchers employed similar approaches but with different conditions such as the pre-processing stage. Some researchers used automatic pre-processing techniques while this thesis used manual for many reasons like the nature of the dataset. Moreover, various machine learning parameters were shown. These parameters were important in tuning the models and improving the performance. Finally, this chapter illustrated several techniques and performance measures used to evaluate the models to ensure their effectiveness. The results will be discussed in the coming chapter.

Chapter Six: Results and Discussion

6.1 Introduction

In this thesis, two approaches were presented to check the power and effectiveness of machine learning models in detecting breast abnormalities in mammogram images, which will help in the early detection of the disease. Both approaches used CNN deep learning models, namely DenseNet121, ResNet50, VGG16, VGG19, Xception, and EffificentNetB7. The first approach classified the images into normal and abnormal, while the second approach compared the classification between different BI-RADS scores such as 3 and 4, 3 and 5, 3 and 6, 4 and 5, 4 and 6, and finally 5 and 6. Moreover, these approaches were evaluated using different performance measures such as confusion matric, accuracy, recall, precision, F1-score, and ROC curve.

In this chapter, the results of implementing six CNN models to classify a first-hand mammogram image dataset into normal and abnormal, and BI-RADS will be presented. Both approaches were done using the same parameter values across all models. The model's input shape was 224*224 and the test subset size was 0.2 of the original datasets. In addition, the batch size was 32 with 20 epochs and Adam optimizer was deployed with a 0.001 learning rate. Finally, the activation functions used in the models were ReLU and Softmax.

6.2 First Approach Results

In the first approach and as mentioned earlier, a training subset of size 1600 was used to learn the models, while a testing subset of size 400 was used for testing and evaluation. This section presents the results obtained from the first approach from 6 CNN models in addition to the results from the performance measures used for evaluation.

6.2.1 DenseNet121

This section presents the results obtained from the DenseNet121 model. Figure (6.1) illustrates the loss function and accuracy of the model versus epochs. As seen in the figure, the training loss function started at around 4.4 in the first epoch and then it started to decrease gradually while the epochs progressed until it reached 0.5 at epoch 20. Comparatively, the validation loss curve started with a value of 2.3 and decreased smoothly until it reached 0.5 at the final epoch. On the other hand, the training accuracy started with the value of 0.6 in the first epoch and then started to increase gradually with small fluctuations until it reached the final accuracy of 0.81. Comparatively, the validation accuracy started with a higher value of 0.68 and increased until it reached a close point of the training accuracy with the value of 0.82 at the final epoch. Unlike the training accuracy curve, the validation accuracy curve was significantly fluctuating as the epochs progressed.

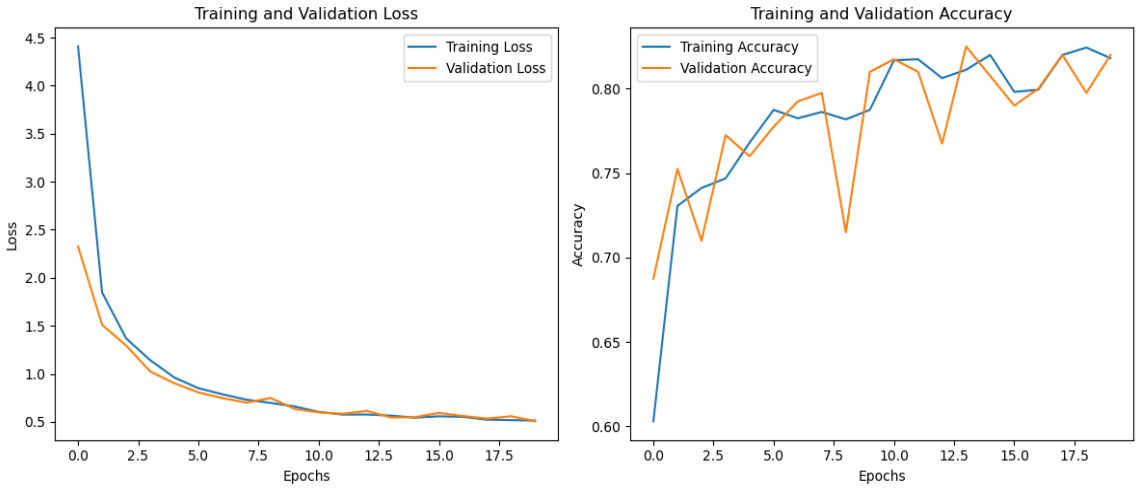


Figure 6.1: Loss and accuracy vs. epoch for Densenet121 in the first approach.

In addition, a confusion matrix resulted from testing the trained model on the testing set. Figure (6.2) depicts the values of the matrix. The values were: TP = 165, FN = 35, FP = 37, TN = 163. That is, the model succeeded in classifying 165 normal images out of

200, called **True Positive**, while misclassified 35 and classified them as abnormal, called **False Negative**. Moreover, the model successfully classified 163 abnormal images which are referred to as **True Negative**, while misclassified 37 and recognized them as normal, which is labeled as **False Positive**.

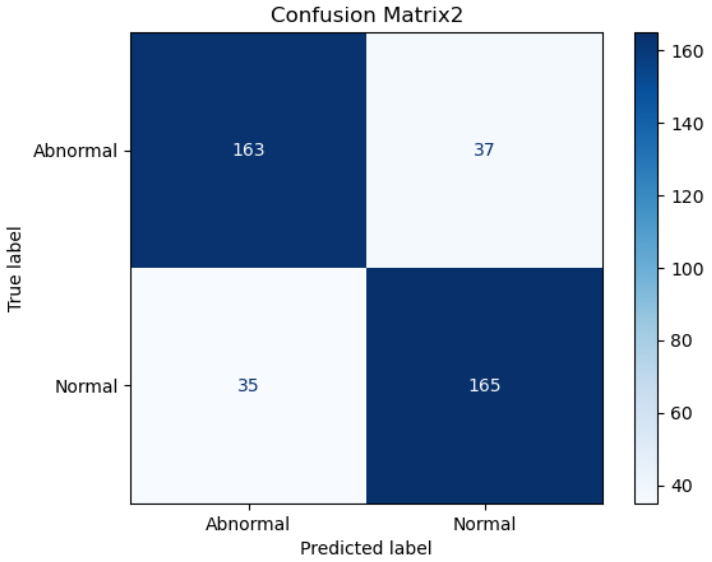


Figure 6.2: Confusion matrix for Densenet121 in the first approach.

From the previous confusion matrix, many measures were calculated. Firstly, the accuracy was equal to 0.82. That is, the probability that the model will successfully classify the output is 80%, and that is a good accuracy in classification tasks. Moreover, the recall or in other words, the true positive rate (TPR) value was equal to the accuracy value of 0.82. That is, the proportion of normal images that were correctly predicted as normal by the model is 82%. On the other hand, the false positive rate (FPR) or the fall-out rate was 0.18, which means that there is an 18% percentage that the model will incorrectly identify the abnormal class as a normal class. In addition, precision was also equal to 0.82. This is the ratio of correctly predicted normal class to all predicted normal. It also means the ability of the model to correctly predict the normal class. Furthermore, the F1-score is calculated from the harmonic mean of recall and precision,

and in this case, is equal to 0.82 which indicates a good classifier. Finally, the ROC curve was plotted from the previous measures with TPR versus the FPR. The x-axis represents the FPR, while the y-axis represents the TPR as shown in Figure (6.3). The area under the curve (AUC) was equal to 0.82. The higher the AUC, the better the classifier, and this can be achieved when the TPR is high and the FPR is low.

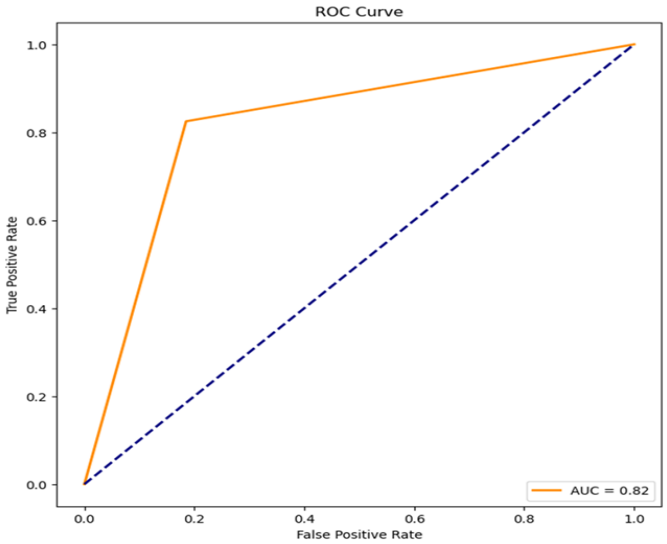


Figure 6.3: AUC for Densenet121 in the first approach.

6.2.2 VGG16

The second used model is VGG16. The results of the loss vs. epochs and the accuracy vs. epochs are illustrated in the following figure. Figure (6.4) shows the curves of the loss and accuracy, where the training loss function started at 3.01 in the first epoch and then started to decrease gradually while the epochs progressed until it reached 0.6 in the final epoch. Comparatively, the validation loss value started at 1.3 and decreased smoothly until it reached the value of 0.59 at epoch 20. On the other hand, the training accuracy started with the value of 0.54 in the first epoch and then started to increase gradually with small fluctuations after the 5^{th} epoch until it reached the final accuracy of 0.68. Comparatively, the validation accuracy started with a higher value of 0.68 and

increased until it reached a close point of the training accuracy with a value of 0.73 at the final epoch. The validation accuracy curve was increasing in general but with unstable steps and significant fluctuation as the epochs progressed. This model showed good results in terms of error function. However, DenseNet121 showed better results in terms of accuracy.

Figure 6.4: Loss and accuracy vs. epoch for VGG16 in the first approach.

Additionally, the confusion matrix in Figure (6.5) reflects the predicted outputs against the actual inputs. In this matrix, the values of TP, FN, FP, and TN were 127, 73, 32, and 168 respectively. That is, the model succeeded in classifying 127 normal images out of 200, called **TP**, while misclassified 73 and classified them as abnormal, called **FN**. Furthermore, the model successfully classified 168 abnormal images which are referred to as **TN**, while misclassified 32 and identified them as normal, which is called **FP**.

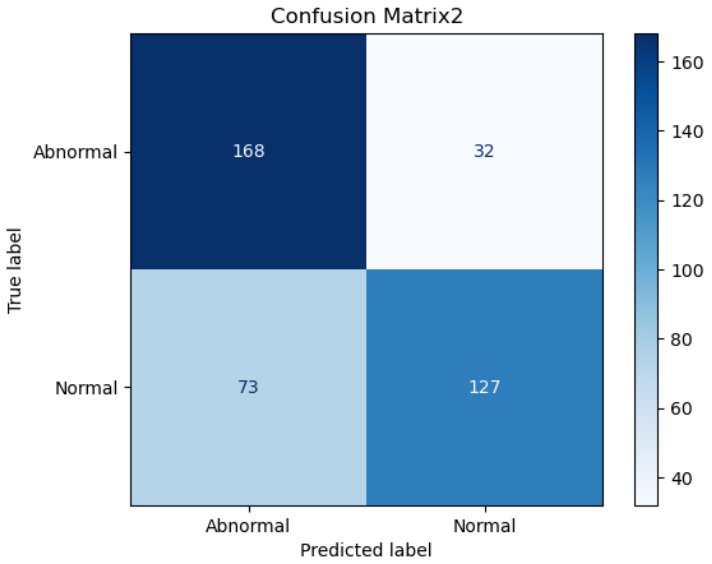


Figure 6.5: Confusion matrix for VGG16 in the first approach.

Moreover, the performance measures were also calculated from the previous values in the matrix. First of all, the accuracy was 0.74 which means that 74% of the outputs were classified correctly. In addition, the recall (TPR) was less than the accuracy with a value of 0.64. That is, the proportion of normal images that were correctly predicted as normal by the model is 64%. On the other hand, the FPR or the fall-out rate was 0.16, which means that there is a 16% percentage that the model will incorrectly identify the abnormal class as a normal class. In the same way, precision resulted in 0.8 which is the rate of correctly predicted normal class to all predicted normal. By the same token, the F1-score was equal to 0.71 which is less than the F1-score result obtained in the DenseNet121. Lastly, the curve of TPR and FPR was drawn in Figure (6.6) where the x-axis represents the FPR and the y-axis represents the TPR. The area under the curve (AUC) was equal to accuracy with a value of 0.74. Here, the TPR was not as high as it was in the DenseNet121, which led to AUC = 0.74.

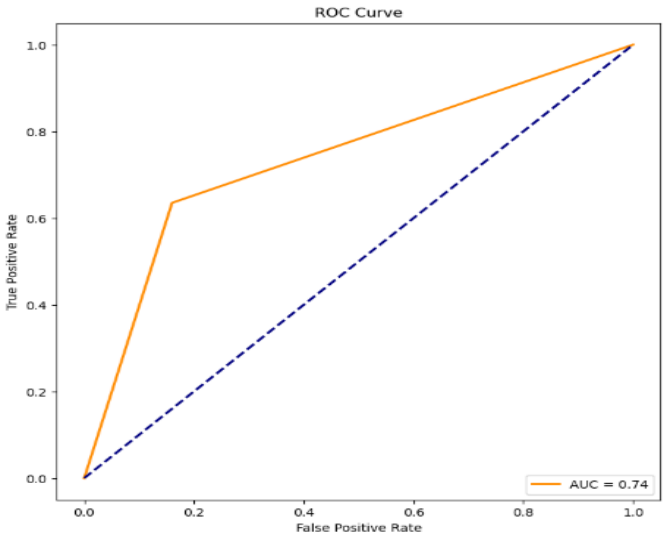


Figure 6.6: AUC for VGG16 in the first approach.

Generally speaking, VGG16 acted better in identifying the abnormal class with only 32 misclassified while there were 73 instances of misclassified in the normal class.

6.2.3 VGG19

The third model is the VGG19. Figure (6.7) shows the loss curve versus epochs and the accuracy curve versus epochs as well. In the same way as the previous models, the training and validation loss functions decreased gradually and smoothly as epochs progressed, where the training loss started at 3.1 and ended at the value of 0.6 and the validation loss started at 1.4 and ended at the value of 0.67. in addition, the training accuracy started at 0.51 and reached 0.59 at the last epoch, while the validation accuracy was slightly higher with the starting value of 0.64 in the first epoch and reaching 0.68 in the last epoch. As seen in the figure, the validation accuracy curve was extremely unstable where no trend or pattern can be found. This might be attributed to many reasons such as the data itself and the learning rate. Note that the learning rate in this approach is 0.001 which outperformed other rates like 0.1 and 0.001. Overall, the accuracy was not improved enough throughout the iterations.

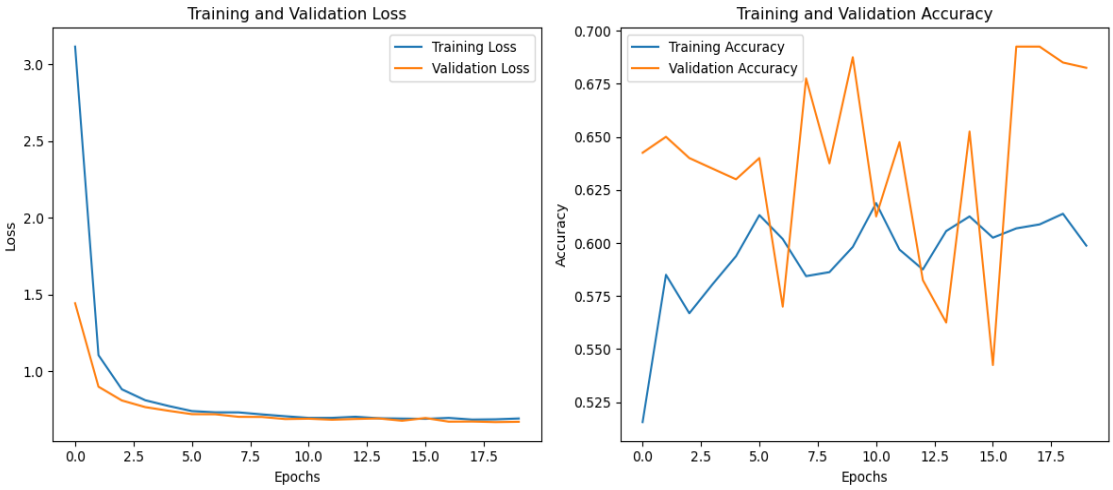


Figure 6.7: Loss and accuracy vs. epoch for VGG19 in the first approach.

Similar to the previous model, the confusion matrix was populated from the testing set predictions as shown in Figure (6.8) below. The four main values in the matrix were as follows: TP = 121, FN = 79, FP = 48, and TN = 152. That is, the model succeeded in classifying 121 normal images out of 200, which is the **True Positive**, while misclassified 79 and classified them as abnormal, called **False Negative**. Moreover, the model successfully recognized 152 abnormal images which are called **True Negative**, while misclassified 48 and recognized them as normal, which is called **False Positive**.

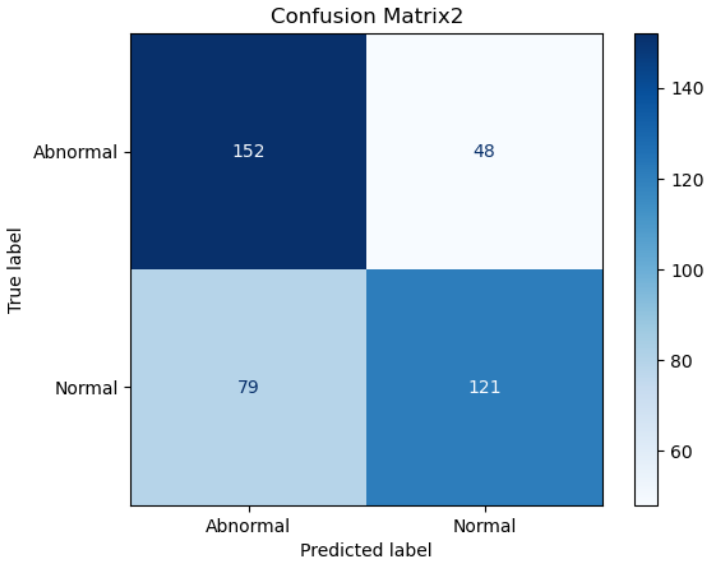


Figure 6.8: Confusion matrix for VGG16 in the first approach.

In the same manner, the performance measures were also produced from the previous values in the matrix. First thing, the accuracy was 0.68 which means that 68% of the outputs were identified correctly. Additionally, the recall (TPR) was less than the accuracy with a value of 0.60. That is, only 60% of normal images were correctly classified as normal by the model. On the other hand, the FPR or the fall-out rate was 0.24, which means that there is a 24% percentage that the model will incorrectly classify the abnormal class as a normal class. By the same token, precision resulted in 0.72 which is the rate of correctly predicted normal class to all predicted normal. In the same way, the F1-score was equal to 0.66 which is less than the F1-score result obtained in the DenseNet121 and VGG16 models. Lastly, the curve of TPR against FPR was drawn in Figure (6.9) where the x-axis is the FPR and the y-axis is the TPR. The area under the curve (AUC) was equal to accuracy with a value of 0.68. Here, the TPR was not as high as it was in the DenseNet121 and VGG16, which led to AUC = 0.68. Generally, VGG16 fits the data more accurately than VGG19. It might be attributed to the number of layers in both of them, and with more complex models there is always a space for inaccurate classifications.

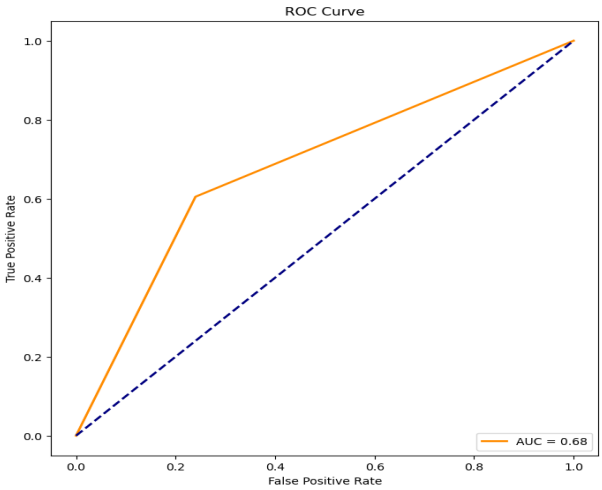


Figure 6.9: AUC for VGG16 in the first approach.

6.2.4 Xception

Xception was the fourth implemented model in this approach. Figure (6.10) depicts the curves of the error vs. epochs and the accuracy vs. epochs. In the figure down below, the blue line which is the training loss starts at 4.5 and decreases rapidly after the first epoch then stabilizes until it reaches the final value of 0.6. Comparatively, the testing loss starts at 2.1 and decreases rapidly just like the training loss curve until stops at the value of 0.7. Moreover, there was a convergence in some areas of the two curves which might indicate overfitting where the model cannot act well on the unseen data. On the other hand, the training accuracy started with a value of 0.61 in the first epoch and then rapidly increased by reaching the second epoch. After that, it had many ups and downs until it reached the final accuracy of 0.72. Comparatively, the validation accuracy started with a higher value of 0.64 and increased significantly. The increment progress significantly fluctuated and dropped significantly after the 17th epoch until it reached the value of 0.61, which is less than the initial value. This also indicates overfitting in the model as the model cannot perform well on the new data.

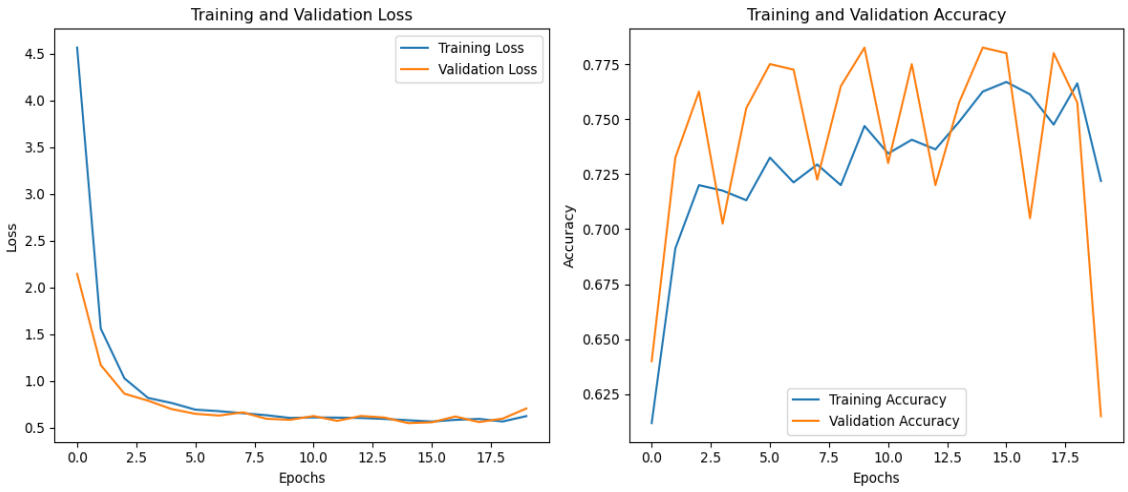


Figure 6.10: Loss and accuracy vs. epoch for Xception in the first approach.

Similarly, the confusion matrix was populated from the testing set predictions as shown in Figure (6.11) below. The four main values in the matrix were as follows: TP = 147, FN = 53, FP = 34, and TN = 166. That is, the model succeeded in classifying 147 normal images out of 200, which is the **True Positive**, while misidentified 53 and classified them as abnormal, called **False Negative**. Moreover, the model successfully recognized 166 abnormal images which are called **True Negative**, while misclassified 34 and recognized them as normal, which is called **False Positive**.

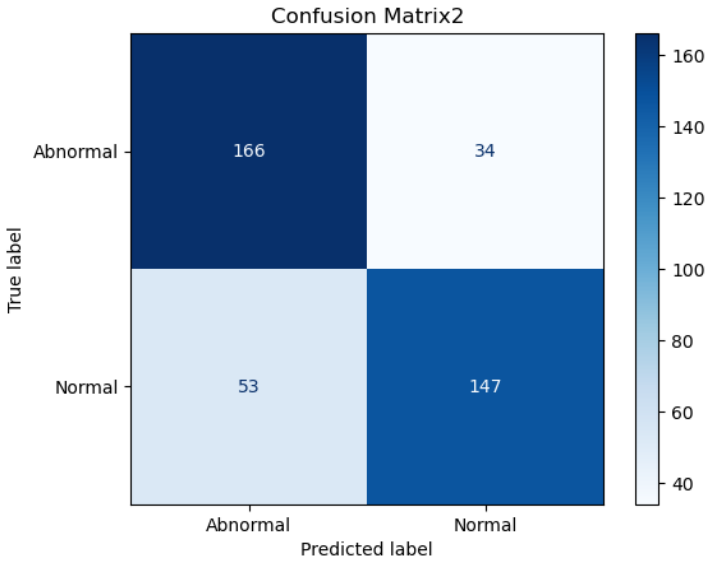


Figure 6.11: Confusion matrix for Xception in the first approach.

By the same token, the performance measures were also populated. First of all, the accuracy was 0.78 which means that 78% of the outputs were classified correctly. Furthermore, the recall (TPR) was less than the accuracy with a value of 0.73. That is, only 73% of normal images were correctly recognized as normal by the model. On the other hand, the FPR or the fall-out rate was 0.17, which means that there is a 17% percentage that the model will incorrectly classify the abnormal class as a normal class. In the same manner, precision resulted in 0.81 which is the rate of correctly predicted

normal class to all predicted normal. In the same way, the F1-score was equal to 0.77 which is greater than the F1-score result obtained in the VGG16 and VGG19 models. Lastly, the curve of TPR against FPR was drawn in Figure (6.12) where the x-axis is the FPR and the y-axis is the TPR. The area under the curve (AUC) was equal to accuracy with a value of 0.78. Here, the TPR was also higher than it was in the VGG16 and VGG19, which led to AUC = 0.78. Generally, Xception fits the data better than VGG16 and VGG19.

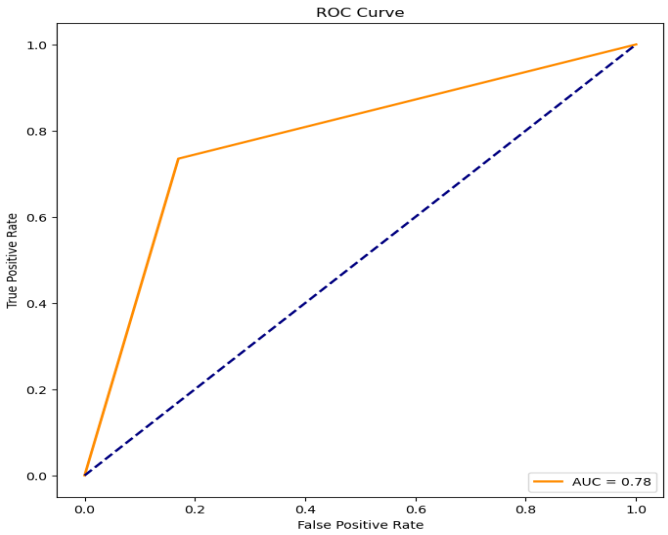


Figure 6.12: AUC for Xception in the first approach.

6.2.5 EfficientNetB7

EffificentNetB7 is one of the Effificent Networks family that was the fifth implemented model. In Figure (6.13), the error curve (loss) and the accuracy curve for the training and validation sets are presented. The loss curve is not very different from the curves resulting from previous models. The training loss curve started at 4.7 then dropped rapidly after the first epoch, and then stabilized until it reached the final value at 0.6. Besides, the validation curve started at 2.08 and gradually decreased as the epochs progressed and then increased slightly between epochs 7 and 12 until it reached the final value of 0.55 in the last epoch. On the other hand, as the epochs progressed, the training accuracy steadily increased until it had the final value of 0.7.

Furthermore, the validation accuracy showed a significant drop between epochs 7 and 10 then it increased again until the final value of 0.78.

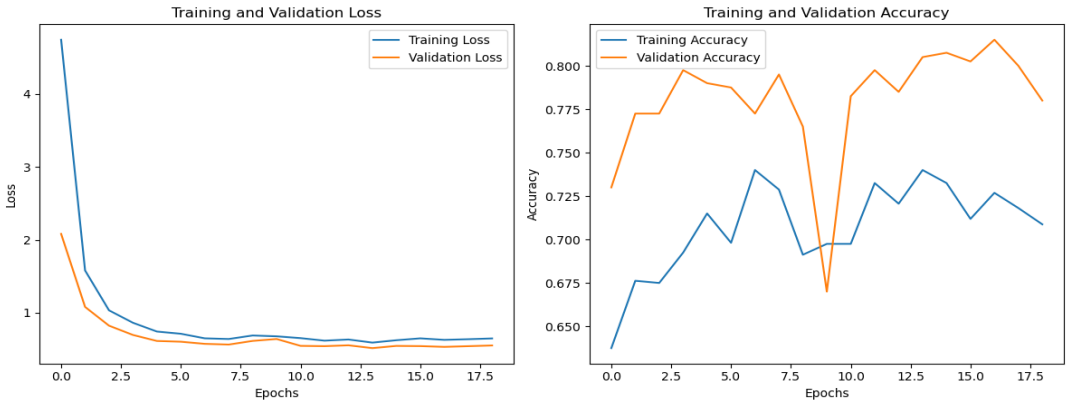


Figure 6.13: Loss and accuracy vs. epoch for EfficientNetB7 in the first approach.

In the same way, the confusion matrix was created from the testing set predictions as shown in Figure (6.14) below. The four main values in the matrix were as follows: TP = 155, FN = 45, FP = 33, and TN = 167. That is, the model succeeded in classifying 155 normal images out of 200, which is a **True Positive**, while misidentified 45 and identified them as abnormal, which is a **False Negative**. Moreover, the model successfully recognized 167 abnormal images which are **True Negatives**, while misclassified 33 and recognized them as normal, which is called **False Positive**.

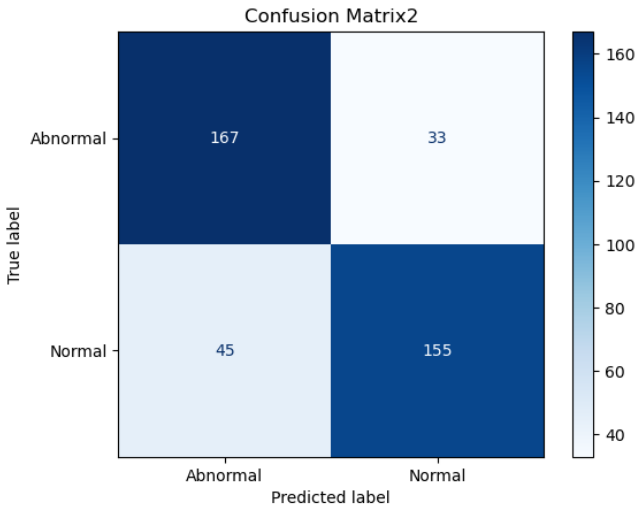


Figure 6.14: Confusion matrix for EfficientNetB7 in the first approach.

Similarly, the confusion matrix produced an accuracy of 0.81 which means that 81% of the outputs were recognized correctly. Moreover, the recall (TPR) was less than the accuracy with a value of 0.78. That is, only 78% of normal images were correctly classified as normal by the model. Also, the FPR or the fall-out rate was 0.165, which means that there is a 16.5% percentage that the model will incorrectly identify the abnormal class as a normal class. Likewise, precision resulted in 0.82 which is the rate of correctly predicted normal class to all predicted normal. The precision here was equal to the precision value resulting from the DenseNet121 model. Additionally, the F1-score was equal to 0.80 which outperformed all previous F1-score results obtained from the implemented models except for DenseNe121. Lastly, the curve of TPR against FPR was illustrated in Figure (6.15) where the x-axis is the FPR and the y-axis is the TPR. The area under the curve (AUC) was close to accuracy with a value of 0.80. Here, the TPR was also higher than it was in the previous models except for DenseNet121. Lastly, the results obtained from EfficientNetB7 indicated to good fit model.

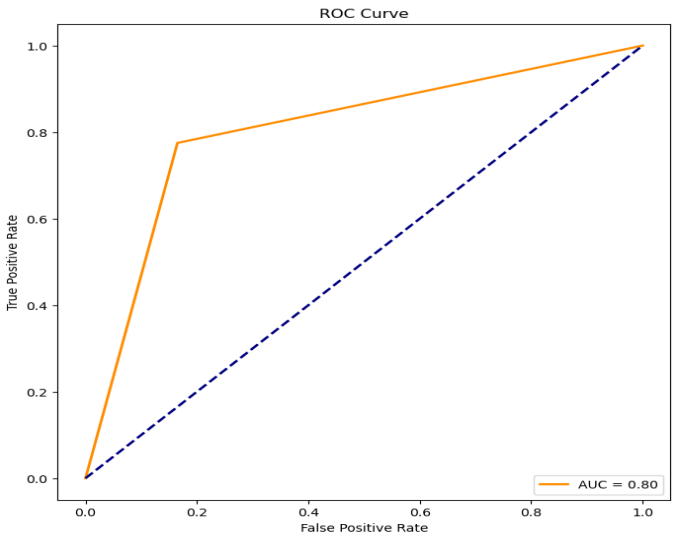


Figure 6.15: AUC for EfficientNetB7 in the first approach.

6.2.6 ResNet50

Residual Network 50 is the last implemented model in this approach. In this model, the results obtained were significantly different from all previous models. Figure (6.16) shows the error curve (loss) and the accuracy curve for the training and validation sets. The training loss curve started at 3.24 then dropped immediately after the first iteration of learning, and then stabilized until it reached the final value at 0.69. Besides, the validation curve started at 1.04 and gradually decreased as the iterations progressed and then converged completely until it reached the final value of 0.69 in the last iteration. This behavior indicates overfitting in the model where the model cannot predict unseen data correctly, and thus cannot be generalized. On the other hand, the training accuracy started to increase after the first iteration and then decreased with ups and downs during the learning process until the final value of 0.47, while the validation curve was constant with the value of 0.5. that is, the model acted like a regular random classifier on unseen data, which also suggests overfitting.

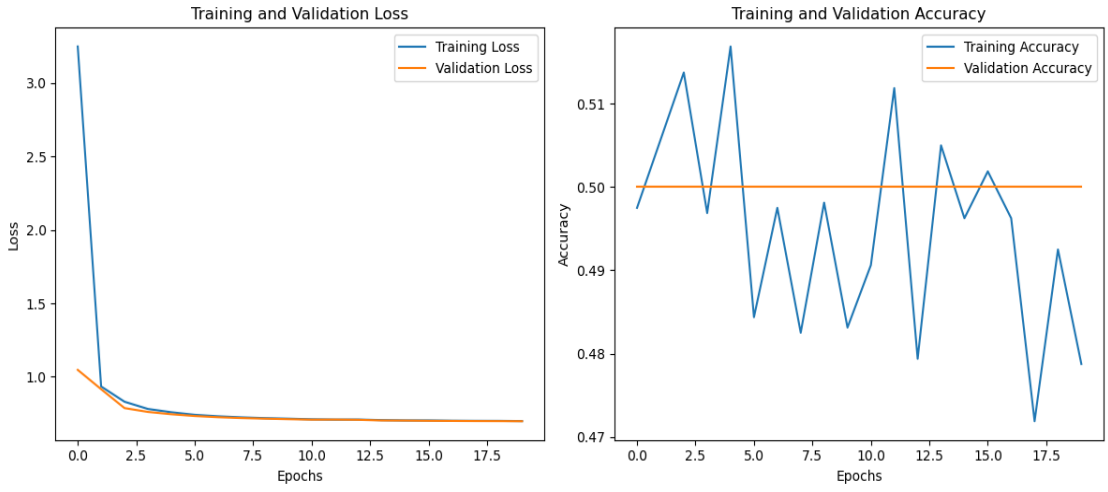


Figure 6.16: Loss and accuracy vs. epoch for ResNet50 in the first approach.

Likewise, the confusion matrix was created from the testing set predictions as shown in Figure (6.17) below. The four main values in the matrix were as follows: TP = 200, FN

= 0, FP = 200, and TN = 0. That is, the model succeeded in classifying 200 normal images out of 200, which is a **True Positive**, while didn't misidentify any of the normal class, which is a **False Negative**. Moreover, the model didn't recognize any image in the abnormal class which is the **True Negative**. Also, it misclassified 200 images out of 200 in the abnormal class and recognized them as normal, which is called **False Positive**.

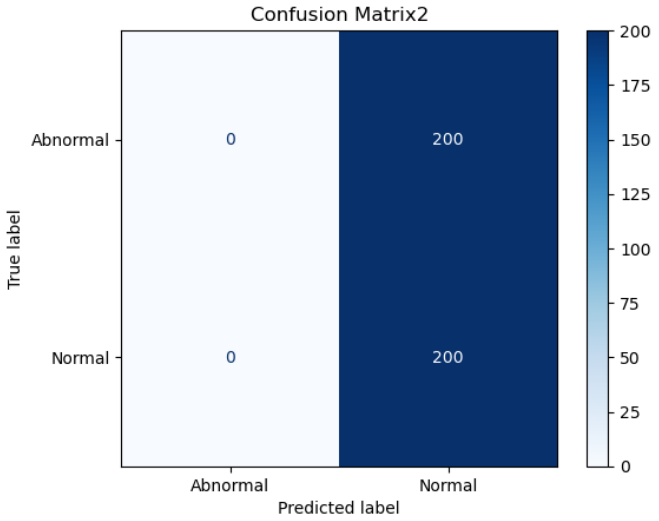


Figure 6.17: confusion matrix for ResNet50 in the first approach.

In addition to the previous, the confusion matrix produced an accuracy of 0.5 which means that 50% of the outputs were classified correctly. Additionally, the recall (TPR) was greater than the accuracy with a value of 1. That is, 100% of normal images were correctly classified as normal by the model. Also, the FPR or the fall-out rate was 1, which means that there is a 100% percentage that the model will incorrectly identify the abnormal class as a normal class. Likewise, precision resulted in 0.52 which is the rate of correctly predicted normal class to all predicted normal. This precision value was the least of all previous models. Moreover, the F1-score was equal to 0.67. Furthermore, the curve of TPR against FPR was drawn in Figure (6.18) where the x-axis is the FPR and

the y-axis is the TPR. The area under the curve (AUC) was equal to accuracy with a value of 0.50. That is, the model's performance is no better than random chance, and is making predictions without any meaningful discrimination between normal and abnormal instances. Lastly, the results obtained from ResNet50 indicated to poor fit model. The AUC's value

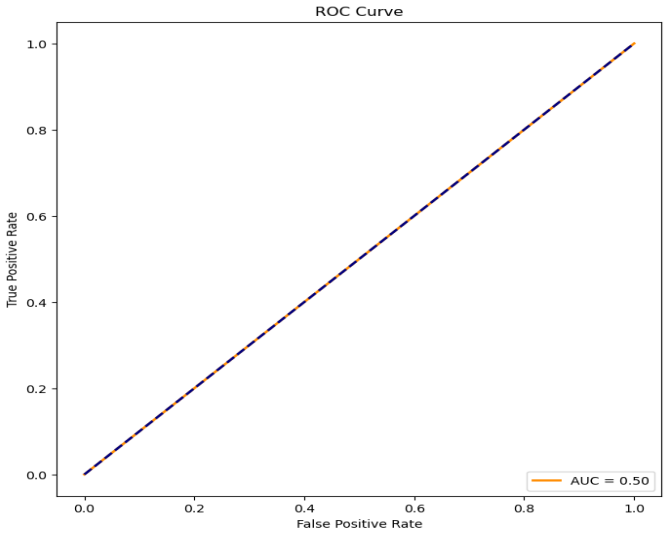


Figure 6.18: AUC for ResNet50 in the first approach.

6.2.7 First Approach Results Summary

In summary and based on the results obtained from the first approach summarized in Table (6.1), DenseNet121 outperformed all other models with the highest accuracy, AUC, Recall, Precision, and F1-score. The highest accuracy for detecting breast abnormalities obtained was 0.82 from DenseNet121, while EfficientNetB7 showed promising and close results of DenseNet121. EfficientNetB7 accuracy was 0.81 and got 0.80, 0.78, 0.82, and 0.80 as AUC, Recall, Precision, and F1-score respectively. Moreover, VGG16, VGG19, and Xception got accuracies of 0.74, 0.68, and 0.78 respectively. The performance in these models was better in detecting the abnormal

class than the normal class. Furthermore, the lowest accuracy was 0.50 from ResNet50 where the model predicted all images as normal class.

Table 6.1: Performance results for the models in the first approach.

Model	Test Accuracy	Recall (TPR)	FPR	Precision	F1-score	ROC-AUC
DenseNet121	**0.82**	**0.82**	**0.18**	**0.82**	**0.82**	**0.82**
VGG16	0.74	0.64	0.16	0.80	0.71	0.74
VGG19	0.68	0.60	**0.24**	0.72	0.66	0.68
Xception	0.78	0.73	0.17	0.81	0.77	0.78
ResNet50	0.50	1.0	1	0.50	0.67	0.50
EfficientNetB7	0.81	0.78	0.16	**0.82**	0.80	0.80

6.3 Second Approach Results

In the Second approach, there were 6 datasets used to feed the same 6 CNN models presented earlier. Each dataset consists of two BI-RADS classes to evaluate the model's performance in recognizing these classes. The datasets are (BI-RAD 3 and BI-RAD 4), (BI-RAD 4 and BI-RAD 5), (BI-RAD 5 and BI-RAD 6), (BI-RAD 3 and BI-RAD 5), (BI-RAD 3 and BI-RAD 6), and (BI-RAD 4 and BI-RAD 6). Furthermore, these six datasets can be grouped into two groups: consecutive categories and non-consecutive categories. Consecutive like BI-RADS (3,4), BI-RADS (4,5), and BI-RADS (5,6), which are consecutive scores in the BI-RADS scale, and non-consecutive categories such as BI-RADS (3,5), BI-RADS (4,6), and BI-RADS (4,6), which are not consecutive in the scale as scores are separating them. Moreover, each dataset is balanced and of size 200 after augmentation and pre-processing. Additionally, this approach used the same parameters presented previously such as input size, learning rate, batch size, optimizer, epochs, and activation functions. Also, the dataset was split into 80% for training (160 images), and 20% for testing (40 images). Finally, this section will discuss the obtained results from this approach.

6.3.1 BI-RAD 3 vs. BI-RAD 4

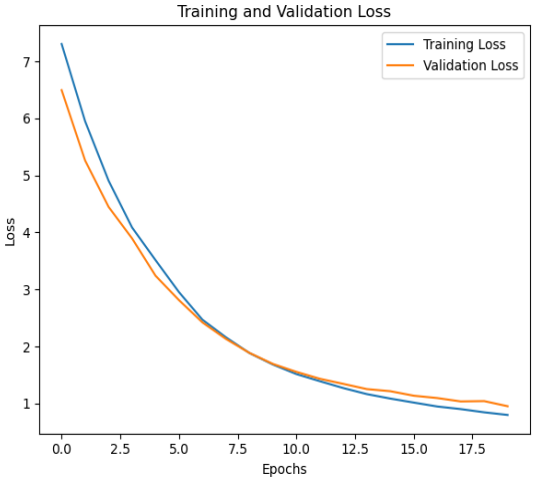
In the first use case, a dataset of 2 classes namely BI-RAD 3 & BI-RAD 4 was fed into 6 models. Figure (6.19) shows the loss function vs. epochs and accuracy vs epochs across 6 models. In DenseNet121, the training and validation loss curves have almost the same behavior. The training curve started at the value of 7.3 and decreased gradually until it reached the final value of 0.7 in the final epoch, while the validation curve started at 5.49 and decreased as the iteration progressed until it reached the final value of 0.9 in the 20^{th} epoch. The two curves intersect between epochs 5 to 12. On the other hand, the training accuracy had an initial value of 0.58 and then started to increase rapidly as expected until it reached the final value of 1 which indicates overfitting. Also, the validation accuracy curve started at 0.5 and rapidly increased after the first epoch and kept fluctuating until it reached the final value of 0.90 in the last epoch. Moreover, VGG16 loss curves showed close behavior as the previous model, but with lower results. The training loss curve started with a value of 5.5 and then began to gradually decrease due to a reduction in the error rate as the iteration progressed until it reached the final value of 0.8. In contrast, the validation curve started at 4.9 and showed the same decreasing rate until the final value of 0.6. Both curves intersected after epoch 10. Comparatively, the training accuracy curve started at the value of 0.51 and then increased gradually as epochs progressed until the final value of 0.86. In the same way, the validation accuracy started at the value of 0.52 and then decreased in the first epoch and increased after that with a significant fluctuation until the final value of 0.80 which is less than the training accuracy.

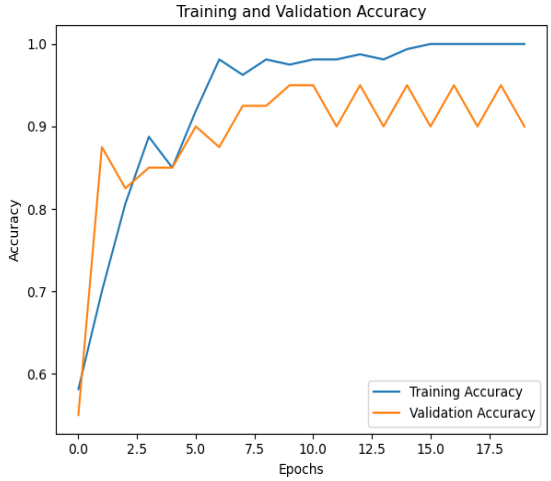
In VGG19, the loss function values in training and validation are almost the same as in VGG16. On the other hand, the training accuracy started at 0.56 and then decreased

after the first iteration and increased after that with significant fluctuations until the final value of 0.77, while the validation accuracy showed similar behavior but with a lower result of 0.65 as the final value in the last epoch. Similarly, the loss curve versus epochs in Xception showed close behavior and results as in DenseNet121 where the initial values of training and validation loss were 8.2 and 7.6, while the final values were 0.54 and 0.65 respectively. Likewise, the accuracy curve had also similar results as in the first model where the initial values of training and validation were 0.56 and 0.57, and the final values were 0.99 and 0.97.

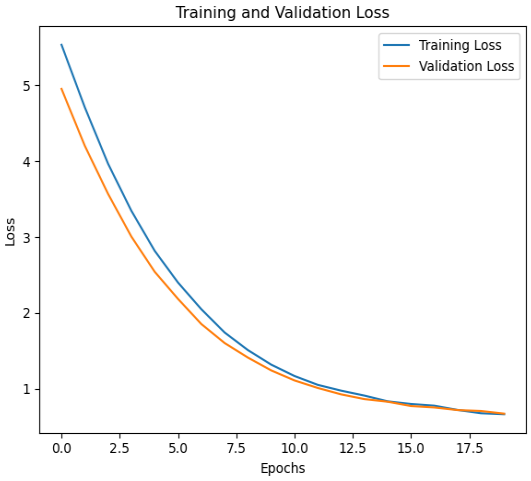
In EfficientNetB7, the initial values of training and validation loss were 8.6 and 7.5, and the final values were 0.86 and 0.77 in order. On the other hand, the initial value of the training accuracy was 0.62 then increased gradually until the final value of 0.92, while the validation had a value of 0.87 as the initial accuracy and 0.92 as the final result. Both training and validation had the same final value in epoch 20, but the first value had a significant difference with the validation value being the highest. Finally, the results in ResNet50 were the lowest. The loss curves were not different from the previous curves but differentiated in accuracy. The training accuracy values ranged between 0.46 and 0.60 throughout the epochs, while the validation accuracy was 0.5 and it was fixed until epoch 13 and then increased significantly and decreased immediately after epoch 16 until it reached the final value of 0.50 which is equal to the initial value.

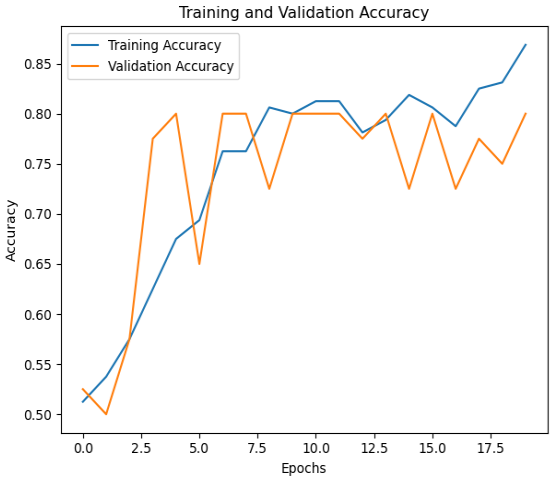
DenseNet121



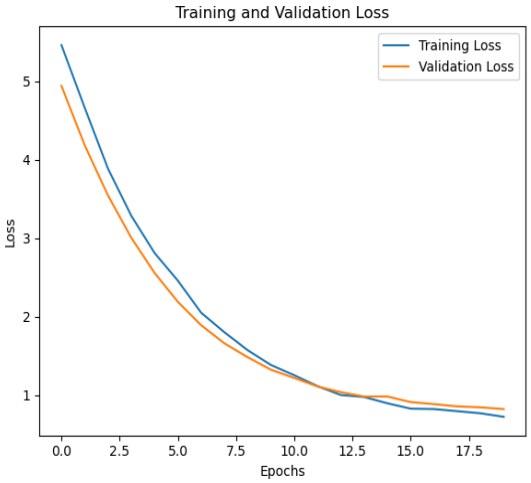


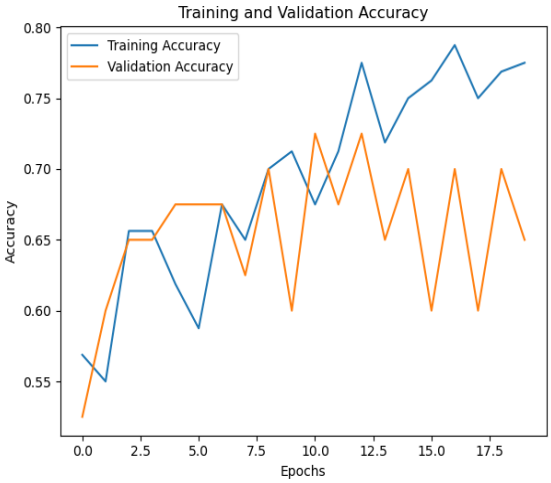
VGG16





VGG19





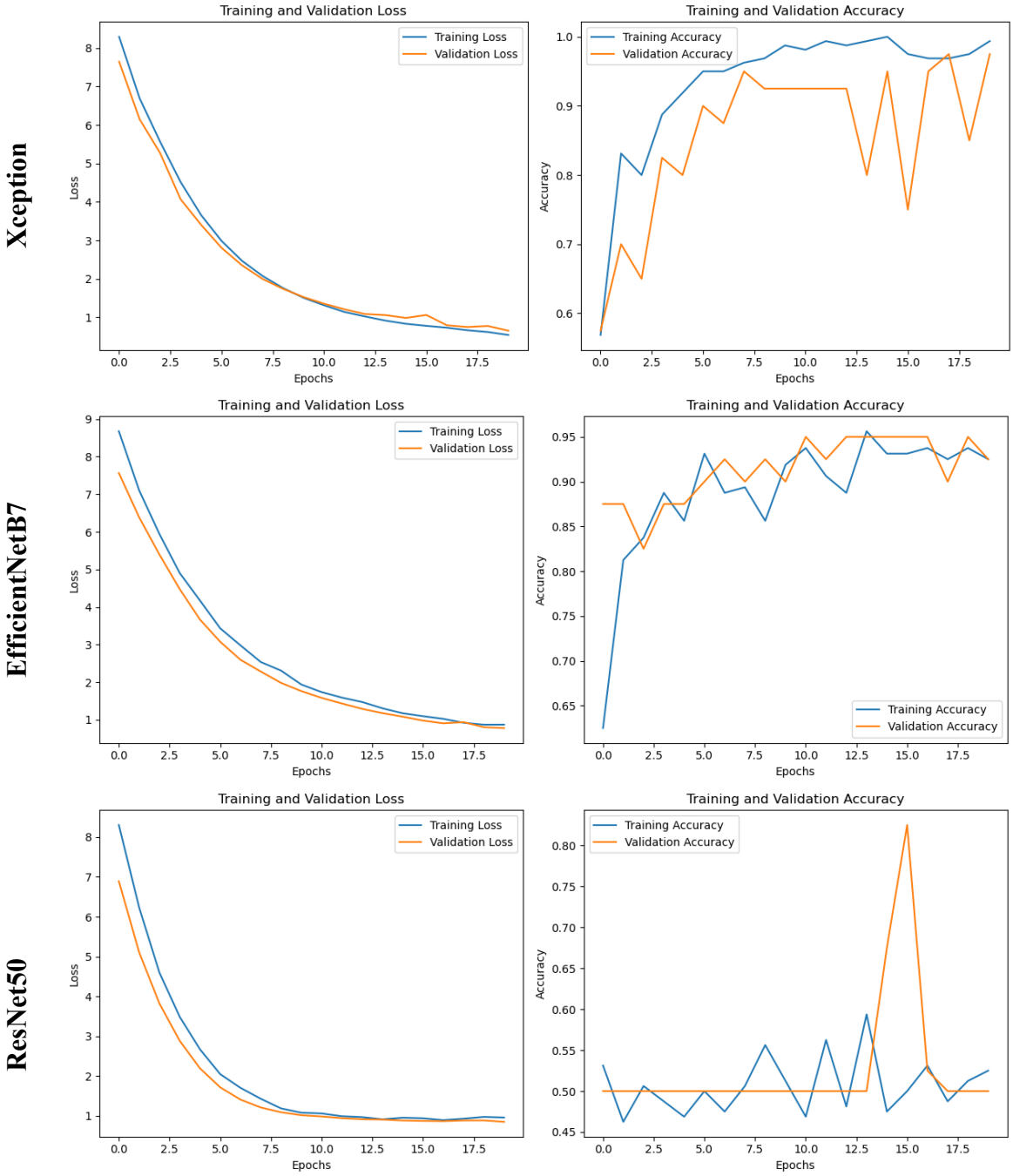


Figure 6.19: Loss and accuracy vs. epoch in BI-RADS 3 and 4.

Additionally, a confusion matrix was calculated from each model of the 6 models. Moreover, the performance measures were also calculated from the confusion matrix. Figure (6.20) below illustrates the confusion matrices in addition to the ROC curves.

In DenseNet121:

The four main values in the matrix were as follows: TP = 18, FN = 2, FP = 2, and TN = 18. That is, the model succeeded in classifying 18 BI-RAD 4 images out of 40, which is the True Positive, while misclassified 2 and classified them as BI-RAD 3, called False Negative. Furthermore, the model successfully recognized 18 BI-RAD 3 images which are called True Negative, while misclassified 2 and recognized them as BI-RAD 4, which is called False Positive. In addition, the performance measures were also produced from the previous values in the matrix. First thing, the accuracy was 0.90 which means that 90% of the outputs were identified correctly. Additionally, the recall (TPR) was equal to the accuracy. That is, 90% of BI-RAD 4 images were correctly classified as BI-RAD 4 by the model. On the other hand, the FPR was 0.1, which means that there is a 10% percentage that the model will incorrectly classify the BI-RAD 3 class as a BI-RAD 4 class. By the same token, precision resulted in 0.90 which is the rate of correctly predicted normal class to all predicted normal. In the same way, the F1-score was equal to 0.90. Lastly, the area under the curve (AUC) was equal to accuracy with a value of 0.90.

In VGG16:

The four main values in the matrix were as follows: TP = 17, FN = 3, FP = 5, and TN = 15. That is, the model succeeded in classifying 17 BI-RAD 4 images out of 40, which is the True Positive, while misclassified 3 and classified them as BI-RAD 3, called False Negative. Moreover, the model successfully identified 15 BI-RAD 3 images which are called True Negative, while misclassified 5 and recognized them as BI-RAD 4, which is called False Positive. Additionally, the accuracy was 0.80 which means that 80% of the outputs were identified correctly. Furthermore, the recall (TPR) was higher than the accuracy with a value of 0.85. That is, 85% of BI-RAD 4 images were correctly

classified as BI-RAD 4 by the model. Comparatively, the FPR was 0.25, which means that there is a 25% percentage that the model will incorrectly classify BI-RAD 3 as BI-RAD 4. In the same way, precision resulted in 0.77, and F1-score was equal to 0.81. Lastly, the area under the curve (AUC) was equal to accuracy with a value of 0.80.

In VGG19:

The four main values in the matrix were as follows: TP = 14, FN = 6, FP = 8, and TN = 12. That is, the model succeeded in identifying 14 BI-RAD 4 images out of 40, which is True Positive, while misclassified 6 and identified them as BI-RAD 3, called False Negative. Likewise, the model successfully identified 12 BI-RAD 3 images which are called True Negative, while misclassified 8 and classified them as BI-RAD 4, which is called False Positive. Similarly, the accuracy was 0.65 which means that 65% of the predictions were classified correctly. Besides, the recall (TPR) was less than the accuracy with a value of 0.7. That is, 70% of BI-RAD 4 images were correctly classified as BI-RAD 4 by the model. On the other side, the FPR was 0.4, which means that there is a 40% percentage that the model will incorrectly recognize BI-RAD 3 as BI-RAD 4. Similarly, precision resulted in 0.64, and F1-score was equal to 0.67. Finally, the area under the curve (AUC) was equal to the accuracy.

In Xception:

The four main values in the matrix were as follows: TP = 19, FN = 1, FP = 0, and TN = 20. That is, the model succeeded in identifying 19 BI-RAD 4, which is True Positive, while misclassified one image and classified it as BI-RAD 3, called False Negative. Similarly, the model successfully recognized 20 BI-RAD 3 images which are all of them and that is called True Negative, while misclassified none. In the same way, the

accuracy was 0.97 which means that 97% of the outputs were recognized correctly. Also, the recall (TPR) was equal to 1. That is, 100% of BI-RAD 4 images were correctly classified as BI-RAD 4 by the model which might indicate overfitting. Comparatively, the FPR was equal to zero. Additionally, precision resulted in 1, and the F1-score was equal to 0.98. Lastly, the area under the curve (AUC) was also equal to the accuracy.

In EfficientNetB7:

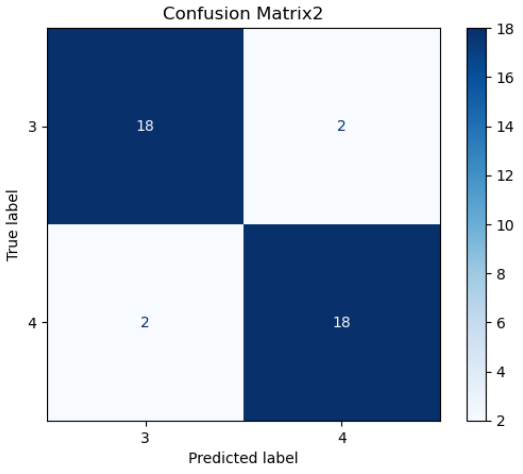
The four main values in the matrix were as follows: TP = 20, FN = 0, FP = 3, and TN = 17. That is, the model succeeded in classifying 20 BI-RAD 4 images out of 40, which is True Positive, while misclassified none and that is called False Negative. Likewise, the model successfully identified 17 BI-RAD 3 images which are called True Negative, while misclassified 3 and classified them as BI-RAD 4, which is called False Positive. Furthermore, the accuracy was 0.93 which means that 93% of the predictions were classified correctly. Besides, the recall (TPR) was higher than the accuracy with a value of 1. That is, 100% of BI-RAD 4 images were correctly classified as BI-RAD 4 by the classifier. Comparatively, the FPR was 0.15, which means that there is a 15% percentage that the model will incorrectly recognize BI-RAD 3 as BI-RAD 4. Likewise, precision resulted in 0.87, and F1-score was 0.93. Finally, the AUC was less than the accuracy with a value of 0.92.

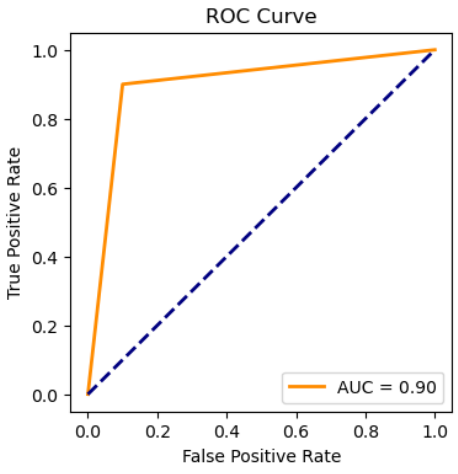
In ResNet50:

The four main values in the matrix were as follows: TP = 20, FN = 0, FP = 20, and TN = 0. That is, the model succeeded in classifying 20 BI-RAD 4 images out of 40, which is True Positive, while misidentified none and that is called False Negative. Similarly, the

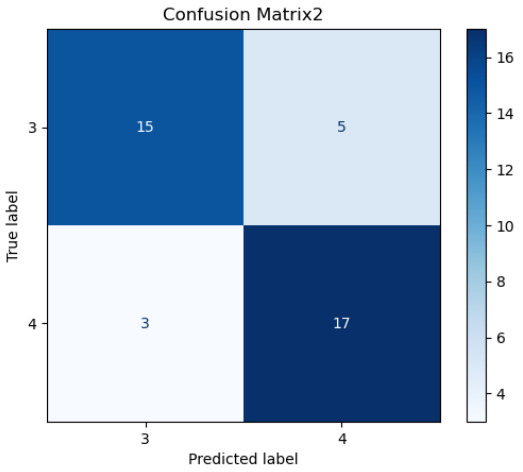
model didn't succeed in classifying any of the BI-RAD 3 images which are called True Negative, while misclassified 20 and classified them as BI-RAD 4, which is called False Positive. In addition, the accuracy was 0.50 which means that only 50% of the predictions were identified correctly. Also, the recall (TPR) was equal to 1. That is, 100% of BI-RAD 4 images were correctly classified as BI-RAD 4 by the classifier. In contrast, the FPR was also equal to 1, which means that there is a 100% percentage that the model will incorrectly recognize BI-RAD 3 as BI-RAD 4. Likewise, precision resulted in 0.50, and F1-score was 0.67. Finally, the AUC was less than the accuracy with a value of 0.50.

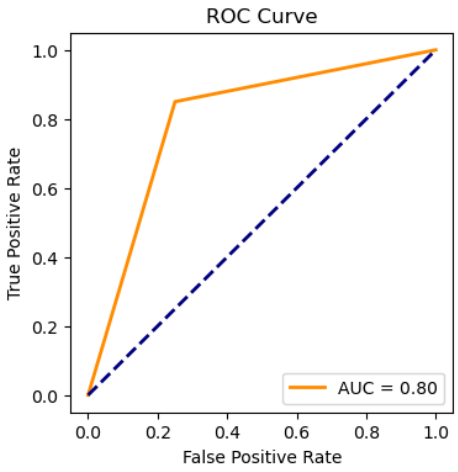
DenseNet121



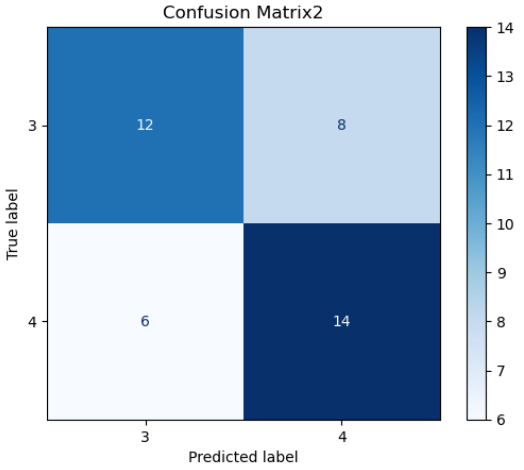


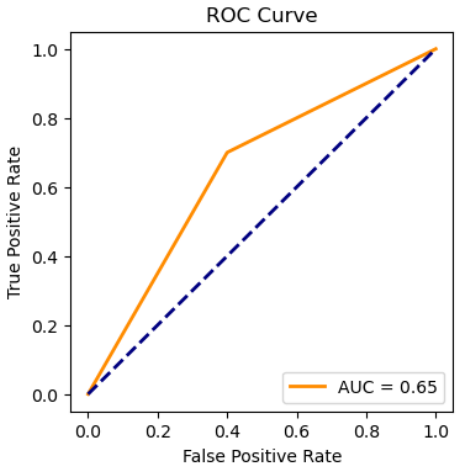
VGG16



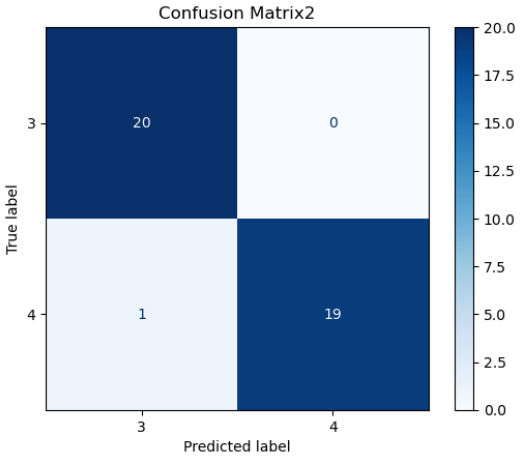


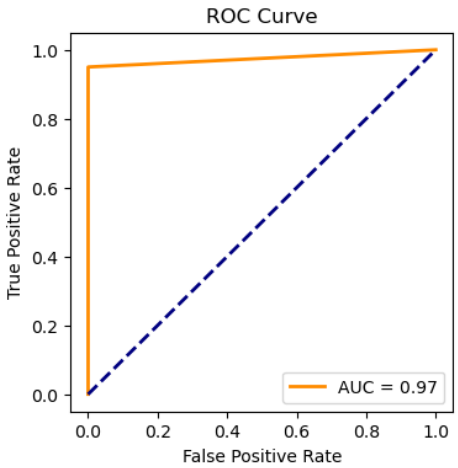
VGG19



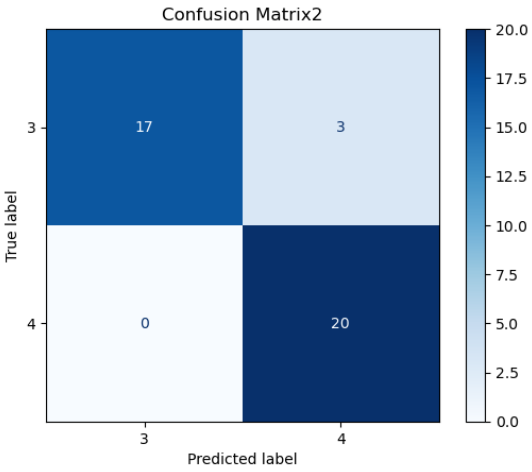


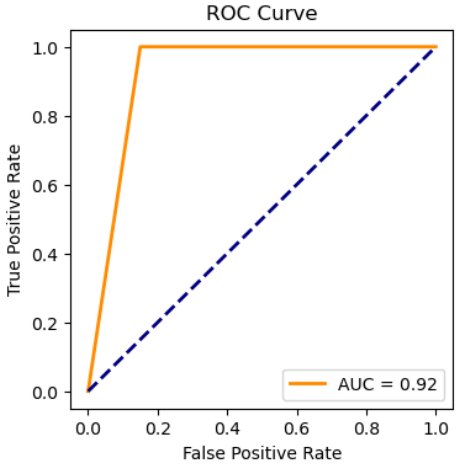
Xception





EfficientNetB7





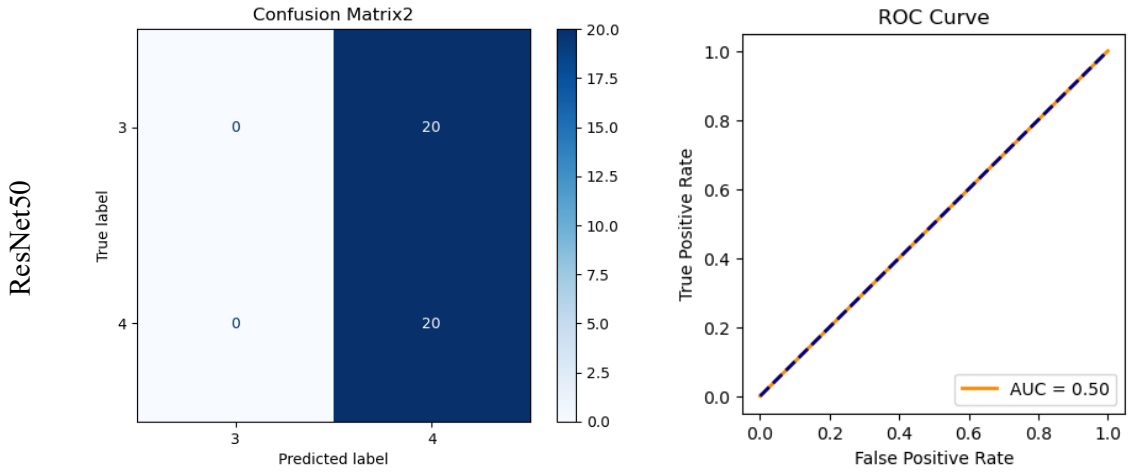


Figure 6.20: Confusion matrix and AUC in BI-RADS 3 and 4.

In the same way, Table (6.2) summarizes the performance measures that were discussed before with the bold values being the best results obtained among all models in this use case.

Table 6.2: Performance measures for all models in BI-RADS 3 and 4.

Model	Test Accuracy	Recall (TPR)	FPR	Precision	F1-score	ROC-AUC
DenseNet121	0.90	0.90	0.10	0.90	0.90	0.90
VGG16	0.80	0.85	0.25	0.77	0.81	0.80
VGG19	0.65	0.70	0.40	0.64	0.67	0.65
Xception	**0.97**	**0.95**	**0.00**	**1.00**	**0.98**	**0.97**
EfficientNetB7	0.93	1.00	0.15	0.87	0.93	0.92
ResNet50	0.50	1.00	1.00	0.50	0.67	0.50

6.3.2 BI-RAD 4 vs. BI-RAD 5

In the second use case, a dataset of BI-RAD 4 & BI-RAD 5 categories was fed into the models. Figure (6.21) shows the results of loss function vs. epochs and accuracy vs epochs across 6 models. In the first model which is the DenseNet121, the training and validation loss curves have almost the same behavior as previous models but with lower results. The training loss curve started at the value of 7.4 and decreased gradually until

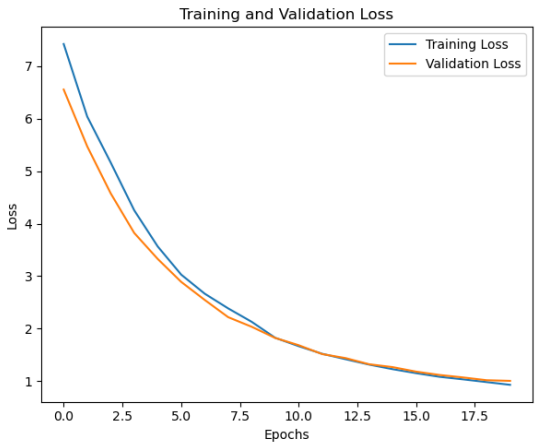
it reached the final value of 0.92 in the final epoch. In contrast, the validation curve started at 6.5 and decreased as the iteration progressed until it reached the final value of 1 in the last epoch. The two curves intersect after epoch 8. Comparatively, the training accuracy had an initial value of 0.52 and then started to increase rapidly as expected until it reached the final value of 0.99. Besides, the validation accuracy curve started at 0.57 and gradually increased after the first epoch and kept fluctuating until it reached the final value of 0.90 in epoch 20. Additionally, VGG16 loss curves showed close behavior as the previous model, but with lower results. The training loss curve started with a value of 5.67 and then began to gradually decrease due to a reduction in the error rate as the iteration progressed until it reached the final value of 0.69. In contrast, the validation curve started at 4.9 and showed the same decreasing rate until the final value of 0.67. Both curves intersected after epoch 15. In addition, the training accuracy curve started at the value of 0.45 and then increased gradually as epochs progressed until the final value of 0.80. In the same way, the validation accuracy started at the value of 0.52 and then increased after that with significant fluctuations until the final value of 0.85 which is higher than the training accuracy.

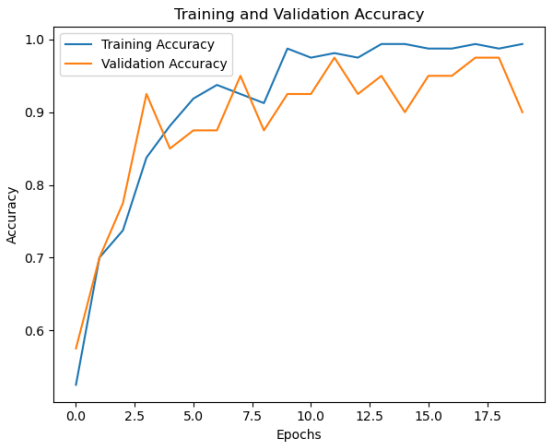
In VGG19, the loss function values in training and validation are almost the same as in VGG16, where the validation loss and accuracy were 0.77 and 0.70 in the last epoch respectively. Moreover, the training accuracy started at 0.44 and then increased after that with significant fluctuations until the final value of 0.72, while the validation accuracy showed similar behavior but with a higher result of 0.70 as the final value in the last iteration. Likewise, the loss curve initial values in Xception of training and validation loss were 8.57 and 7.39, while the final values were 0.72 and 0.81

respectively. Similarly, the initial values of training and validation accuracy were 0.49 and 0.72, and the final values were 0.98 and 0.95.

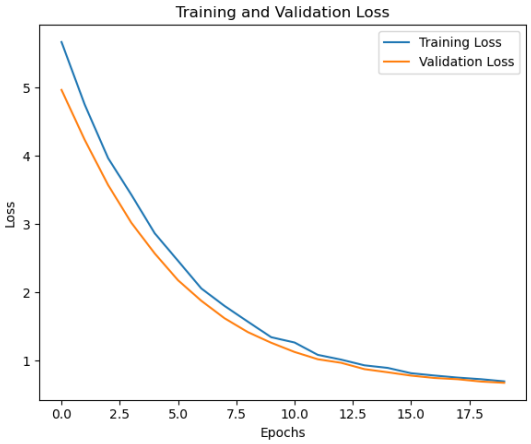
In EfficientNetB7, the initial values of training and validation loss were 9.04 and 7.97, and the final values were 1.09 and 0.8 in order. On the other hand, the initial value of the training accuracy was 0.49 then increased gradually until the final value of 0.86, while the validation had a higher value of 0.67 as the initial accuracy and 1.00 as the final result, which indicates overfitting. Lastly, the results in ResNet50 were the lowest in this section. The loss curves were not different than the previous curves but differentiated in accuracy. The training accuracy values ranged between 0.57 and 0.46 throughout the epochs. In contrast, the validation accuracy was 0.5 and it was fixed until epoch 17 and then increased significantly and decreased immediately after epoch 18 until it reached the final value of 0.55.

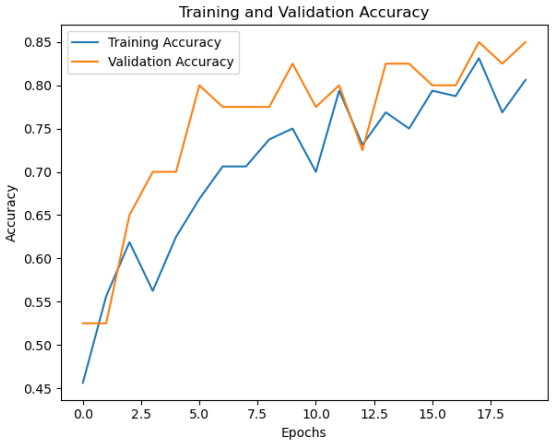
DenseNet121



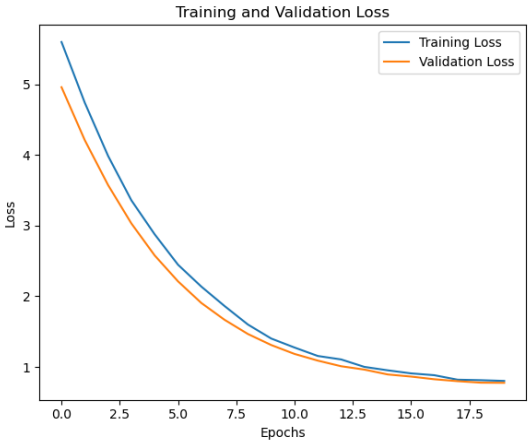


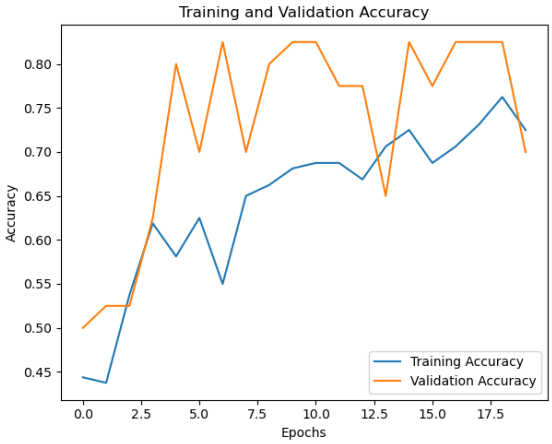
VGG16



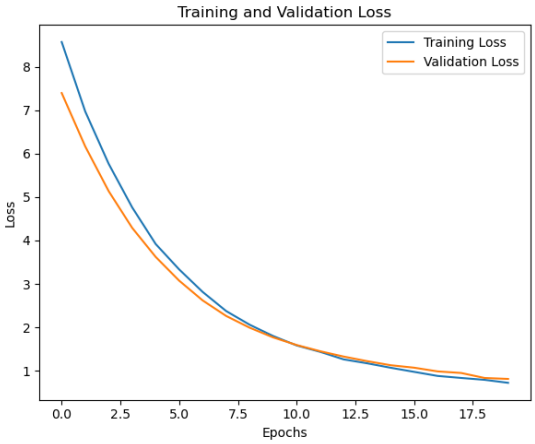


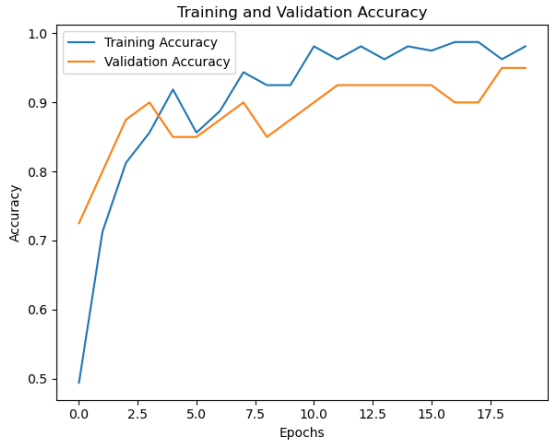
VGG19





Xception





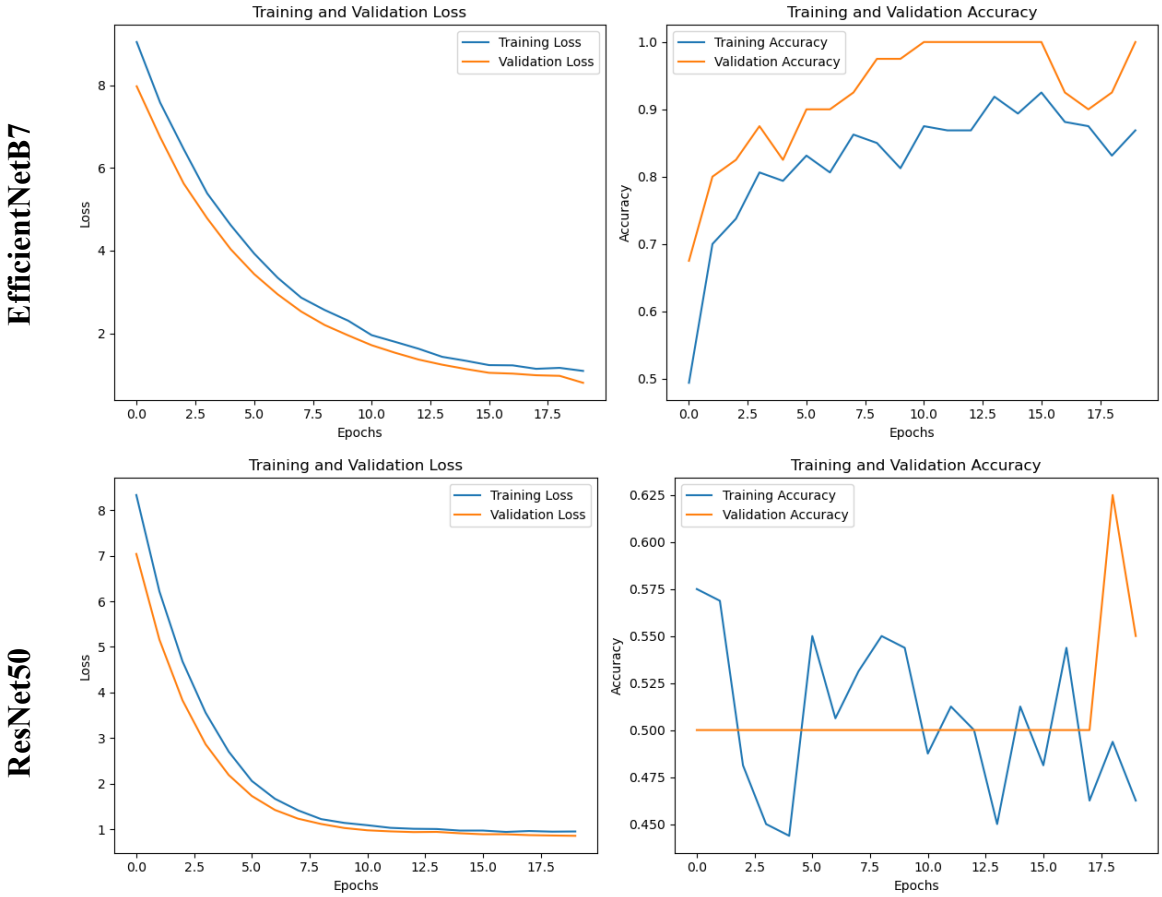


Figure 6.21: Loss and accuracy vs. epoch in BI-RADS 4 and 5.

Moreover, the confusion matrix values, performance measures results, and ROC curve were also calculated and presented in Figure (6.22).

In DenseNet121:

The four main values in the matrix were as follows: TP = 20, FN = 0, FP = 4, and TN = 16. That is, the model succeeded in identifying 20 BI-RAD 5 images out of 40, which is True Positive, while misclassified none, called False Negative. Moreover, the model successfully recognized 16 BI-RAD 4 images which are called True Negative, while misclassified 4 and recognized them as BI-RAD 5, which is called False Positive. Additionally, the accuracy was 0.90 which means that 90% of the predictions were identified correctly. Furthermore, the recall (TPR) was equal to 1. That is, 100% of BI-

RAD 5 images were correctly classified as BI-RAD 5 by the model. On the other hand, the FPR was 0.2, which means that there is a 20% percentage that the model will incorrectly classify the BI-RAD 4 class as a BI-RAD 5 class. By the same token, precision resulted in 0.83 which is the rate of correctly predicted BI-RAD 5 class to all predicted BI-RAD 5. In the same way, the F1-score was equal to 0.91. Lastly, the area under the curve (AUC) was equal to accuracy.

In VGG16:

The four main values in the matrix were as follows: TP = 15, FN = 5, FP = 1, and TN = 19. That is, the model succeeded in classifying 15 BI-RAD 5 images out of 40, which is the True Positive, while misclassified 5 and classified them as BI-RAD 4, called False Negative. Moreover, the model successfully identified 19 BI-RAD 4 images which are called True Negative, while misclassified 1 and recognized them as BI-RAD 5, which is called False Positive. Moreover, the accuracy was 0.85 which means that 85% of the predicted instances were identified correctly. Furthermore, the recall (TPR) was less than the accuracy with a value of 0.75. That is, 75% of BI-RAD 5 images were correctly classified as BI-RAD 5 by the model. In contrast, the FPR was 0.05, which means that there is a 5% percentage that the model will incorrectly classify BI-RAD 4 as BI-RAD 5. Similarly, precision resulted in 0.94, and F1-score was equal to 0.83. Finally, the AUC was equal to accuracy with a value of 0.85.

In VGG19:

The four main values in the matrix were as follows: TP = 11, FN = 9, FP = 3, and TN = 17. That is, the model succeeded in identifying 11 BI-RAD 5 images out of 40, which is True Positive, while misclassified 9 and identified them as BI-RAD 4, called False

Negative. Likewise, the model successfully identified 17 BI-RAD 4 images which are called True Negative, while misclassified only 3 and classified them as BI-RAD 5, which is called False Positive. Moreover, the accuracy was 0.70 which means that 70% of the predictions were classified correctly. Also, the recall (TPR) was less than the accuracy with a value of 0.55. That is, 55% of BI-RAD 5 images were correctly classified as BI-RAD 5 by the model. On the other side, the FPR was 0.15, which means that there is a 15% percentage that the classifier will incorrectly classify BI-RAD 4 as BI-RAD 5. In the same way, the precision was 0.79, and F1-score was equal to 0.65. Finally, the AUC was equal to the accuracy.

In Xception:

The four main values in the matrix were as follows: TP = 18, FN = 2, FP = 0, and TN = 20. That is, the model succeeded in identifying 18 BI-RAD 5, which is True Positive, while misclassified two images and identified them as BI-RAD 4, called False Negative. Similarly, the model successfully recognized 20 BI-RAD 4 images which are all of them and that is called True Negative, while misclassified none. Similarly, the accuracy was 0.95 which means that 95% of the predicted images were classified correctly. Besides, the recall (TPR) was equal to 0.9. That is, 90% of BI-RAD 5 images were correctly classified as BI-RAD 5 by the model. On the other hand, the FPR was zero. In addition, precision resulted in 1.00, and F1-score was equal to 0.95. Lastly, the area under the curve (AUC) was 0.95.

In EfficientNetB7:

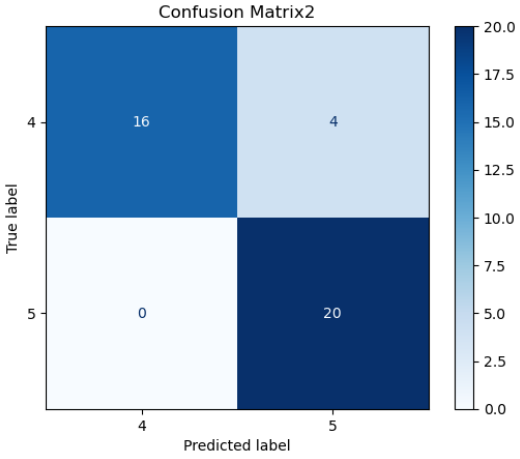
The four main values in the matrix were as follows: TP = 20, FN = 0, FP = 0, and TN = 20. That is, the model succeeded in classifying 20 BI-RAD 5 images out of 40, which is

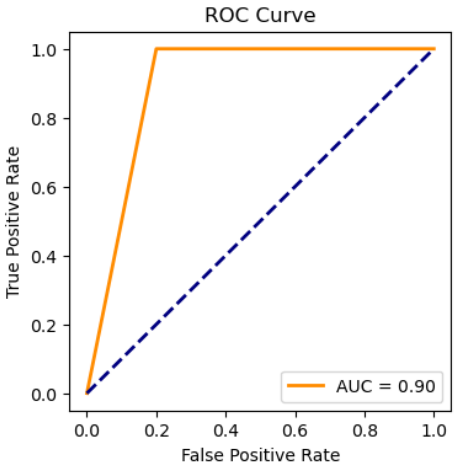
True Positive, while misclassified none and that is called False Negative. Likewise, the model successfully identified 20 BI-RAD 4 images which are called True Negative, while misclassified none, which is called False Positive. Moreover, the accuracy resulted in 1.00 which means that 100% of the outputs were classified correctly, which potentially suggests overfitting in this use case. Also, the recall (TPR) was equal to the accuracy. That is, 100% of BI-RAD 5 images were correctly classified as BI-RAD 5 by the classifier. In contrast, the FPR was zero, meaning that there is no chance that the model will incorrectly recognize BI-RAD 4 as BI-RAD 5. Likewise, precision resulted in 1.00, and F1-score was also equal to 1. Finally, the AUC was equal to 1.

In ResNet50:

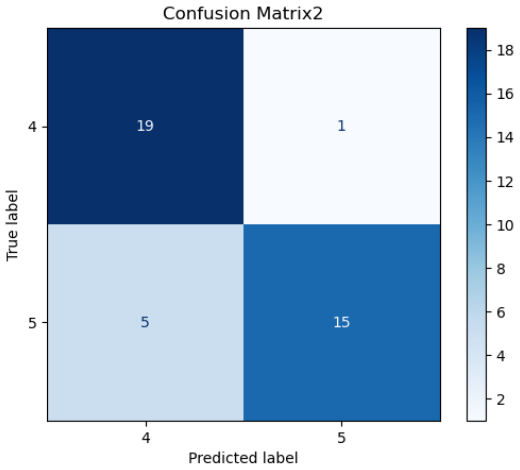
The four main values in the matrix were as follows: TP = 20, FN = 0, FP = 18, and TN = 2. That is, the model succeeded in classifying 20 BI-RAD 5 images out of 40, which is True Positive, while misidentified none and that is called False Negative. Likewise, the model successfully identified only two BI-RAD 4 images which are called True Negative, while misclassified 18 and classified them as BI-RAD 5, which is called False Positive. Additionally, the accuracy was 0.55 meaning that only 55% of the predictions were identified correctly. Also, the recall (TPR) was equal to 1. That is, 100% of BI-RAD 5 images were correctly classified as BI-RAD 5 by the classifier. Comparatively, the FPR was also equal to 0.90, meaning that there is a 90% percentage that the model will incorrectly recognize BI-RAD 4 as BI-RAD 5. Similarly, precision resulted in 0.53, and F1-score was 0.69. Finally, the AUC was 0.55.

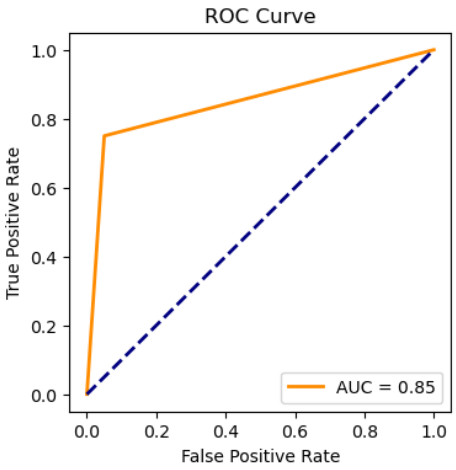
DenseNet121



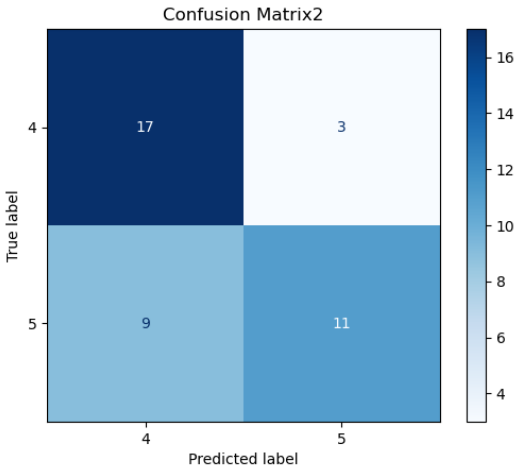


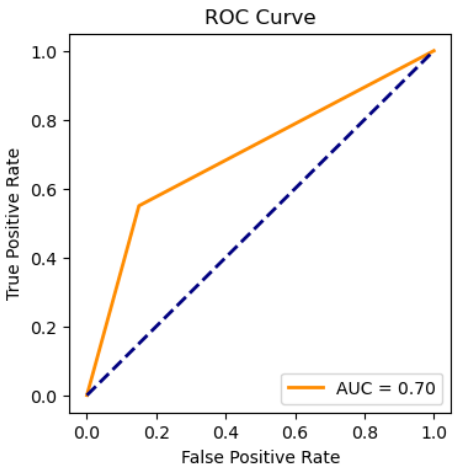
VGG16





VGG19





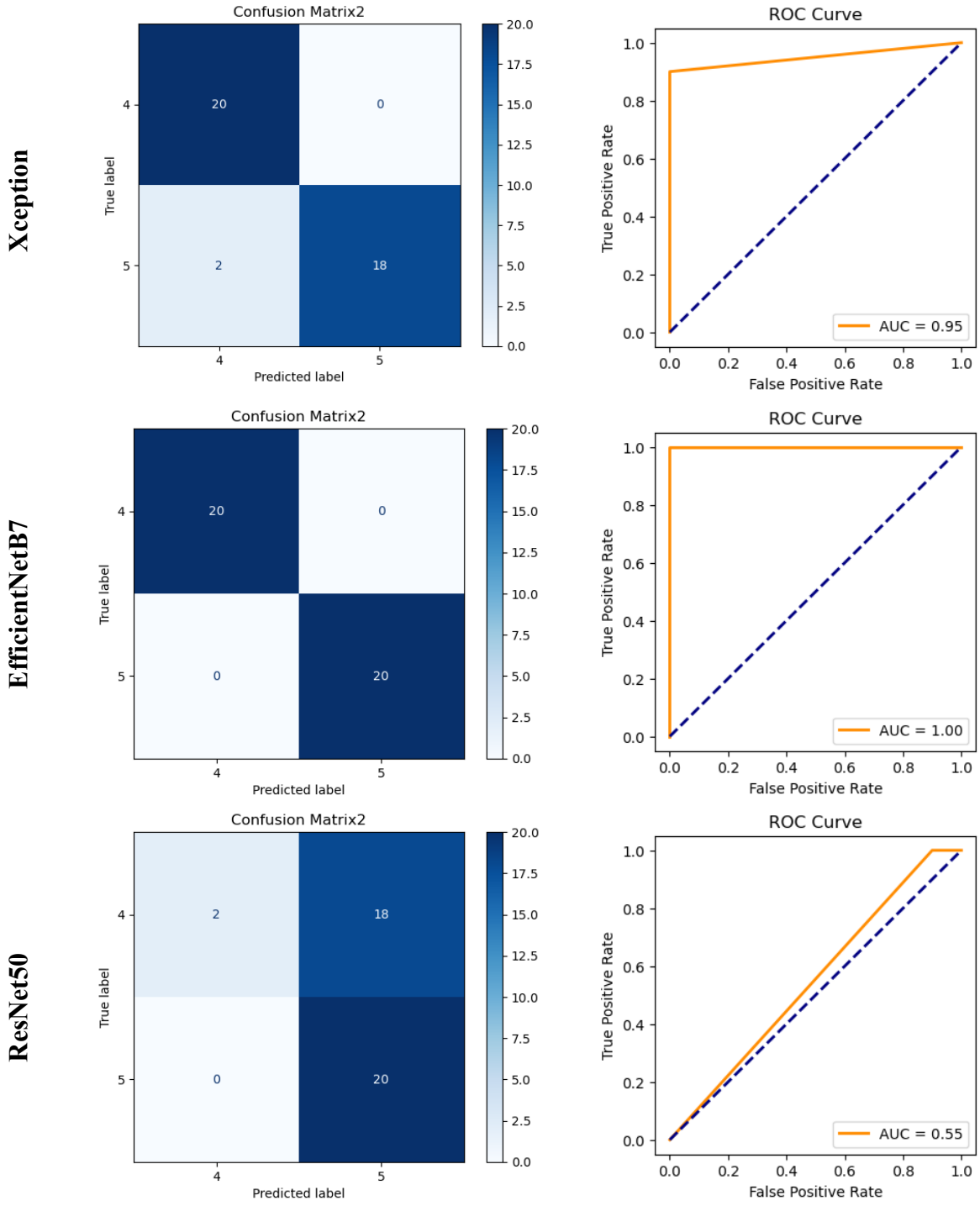


Figure 6.22: Confusion matrix and AUC in BI-RADS 4 and 5.

Moreover, Table (6.3) below summarizes the performance measures resulting in this use case. Overall, Xception had the best performance among all implemented models.

Table 6.3: Performance measures for all models in BI-RADS 4 and 5.

Model	Test Accuracy	Recall (TPR)	FPR	Precision	F1-score	ROC-AUC
DenseNet121	0.90	1.00	0.20	0.83	0.91	0.90
VGG16	0.85	0.75	0.05	0.94	0.83	0.85
VGG19	0.70	0.55	0.15	0.79	0.65	0.70
Xception	**0.95**	**0.90**	**0.00**	**1.00**	**0.95**	**0.95**
EfficientNetB7	1.00	1.00	0.00	1.00	1.00	1.00
ResNet50	0.55	1.00	0.90	0.53	0.69	0.55

6.3.3 BI-RAD 5 vs. BI-RAD 6

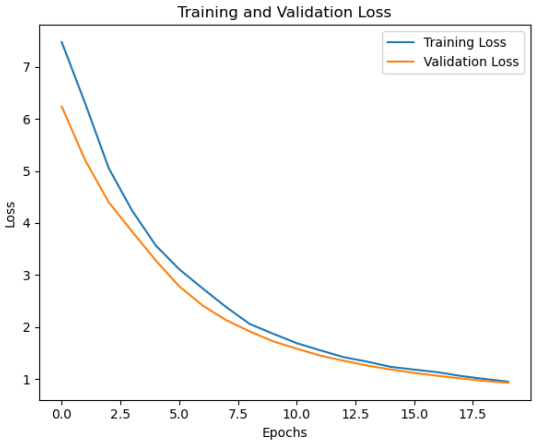
In the third use case, a dataset of BI-RAD 5 & BI-RAD 6 classes was fed into the models. Figure (6.23) shows the results of loss function vs. epochs and accuracy vs epochs across all implemented models. In DenseNet121, the training and validation loss curves have almost the same behavior as previous models. The training loss curve started at the value of 7.4 and decreased gradually until it reached the final value of 0.95 in the final epoch. In contrast, the validation curve started at 6.2 and decreased as the iteration progressed until it reached the final value of 0.92 in the final epoch. The two curves intersect after epoch 13. On the other hand, the training accuracy had an initial value of 0.56 and then started to increase rapidly as expected until it reached the final value of 0.99. Also, the validation accuracy curve started at 0.72 and gradually increased after the first epoch until it reached the final value of 1.00 in the 20^{th} iteration. In addition, VGG16 loss curves showed close behavior as the previous model, but with lower results. The training loss curve started with a value of 5.54 and then began to gradually decrease due to a reduction in the error rate as the epochs progressed until it reached the final value of 0.59. In contrast, the validation curve started at 4.8 and showed the same decreasing rate until the final value of 0.48. Furthermore, the training accuracy curve started at the value of 0.52 and then increased gradually as epochs progressed until the final value of 0.84. By the same token, the validation accuracy

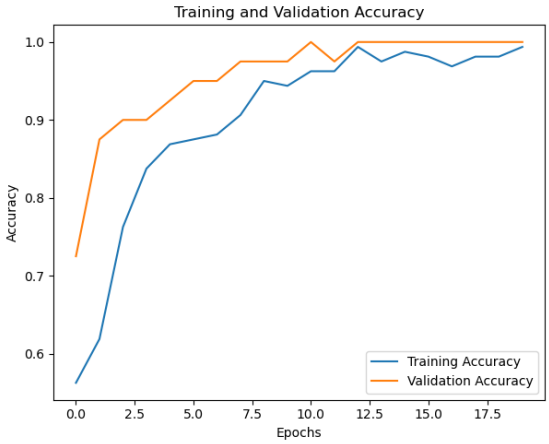
started at a high value of 0.87 and then increased slightly after that until the final value of 0.90 which is higher than the training accuracy.

In VGG19, the loss function values in training and validation are very similar to VGG16, where the validation loss and accuracy were 0.69 and 0.85 in the last epoch respectively. Moreover, the training accuracy started at 0.48 and then increased after that with fluctuations until the final value of 0.79, while the validation accuracy showed similar behavior but with a higher result of 0.85 as the final value in the last iteration. Likewise, the loss curve initial values in Xception of training and validation loss were 8.4 and 7. 2, while the final values were 0.52 and 0.62 respectively. Similarly, the training accuracy started at 0.52 and rapidly increased after that to the final value of 0.99. In contrast, the validation accuracy started at a higher value of 0.8 and slightly increased to 0.95.

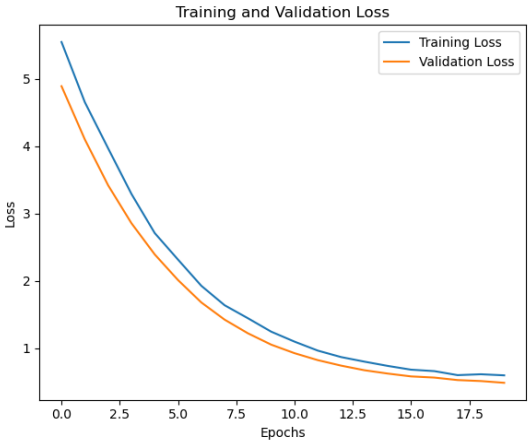
In EfficientNetB7, both loss function and accuracy were similar to the results in Xcepton. The initial values of training and validation loss were 8.58 and 7.43, and the final values were 0.76 and 0.62 in order. On the other hand, the initial value of the training accuracy was 0.66 then increased rapidly with significant fluctuations until the final value of 0.93, while the validation had a higher value of 0.90 as the initial accuracy and 0.97 as the final result. Finally, the results in ResNet50 were the lowest in this section. The loss curves were not different than the previous curves but differentiated in accuracy. The training accuracy initial value was 0.44 then increased until epoch 2.5. After that, it significantly decreased and then gradually increased to 0.56. Besides, the validation accuracy had a value of 0.5 and didn’t improve as epochs progressed and this indicates a poor fit model that acts like the random classifier.

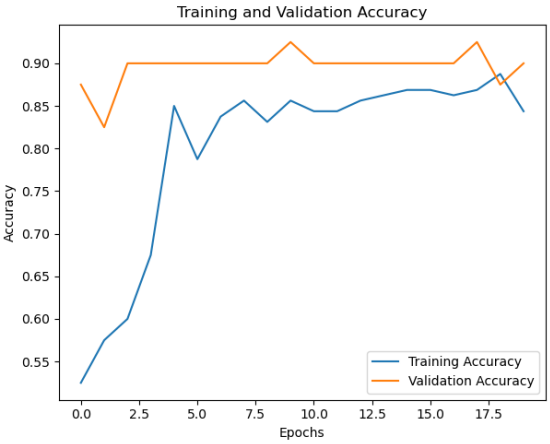
DenseNet121



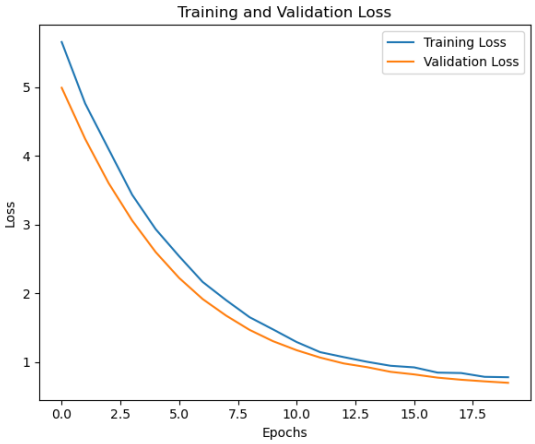


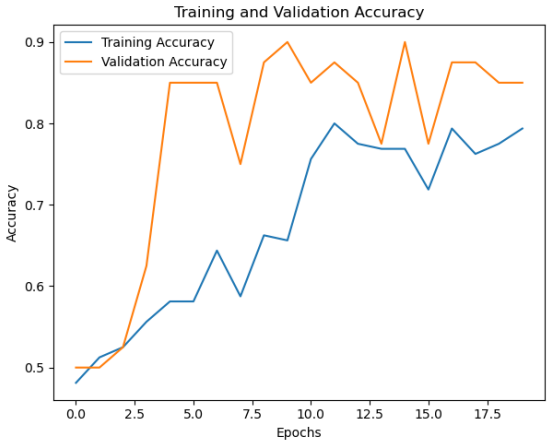
VGG16





VGG19





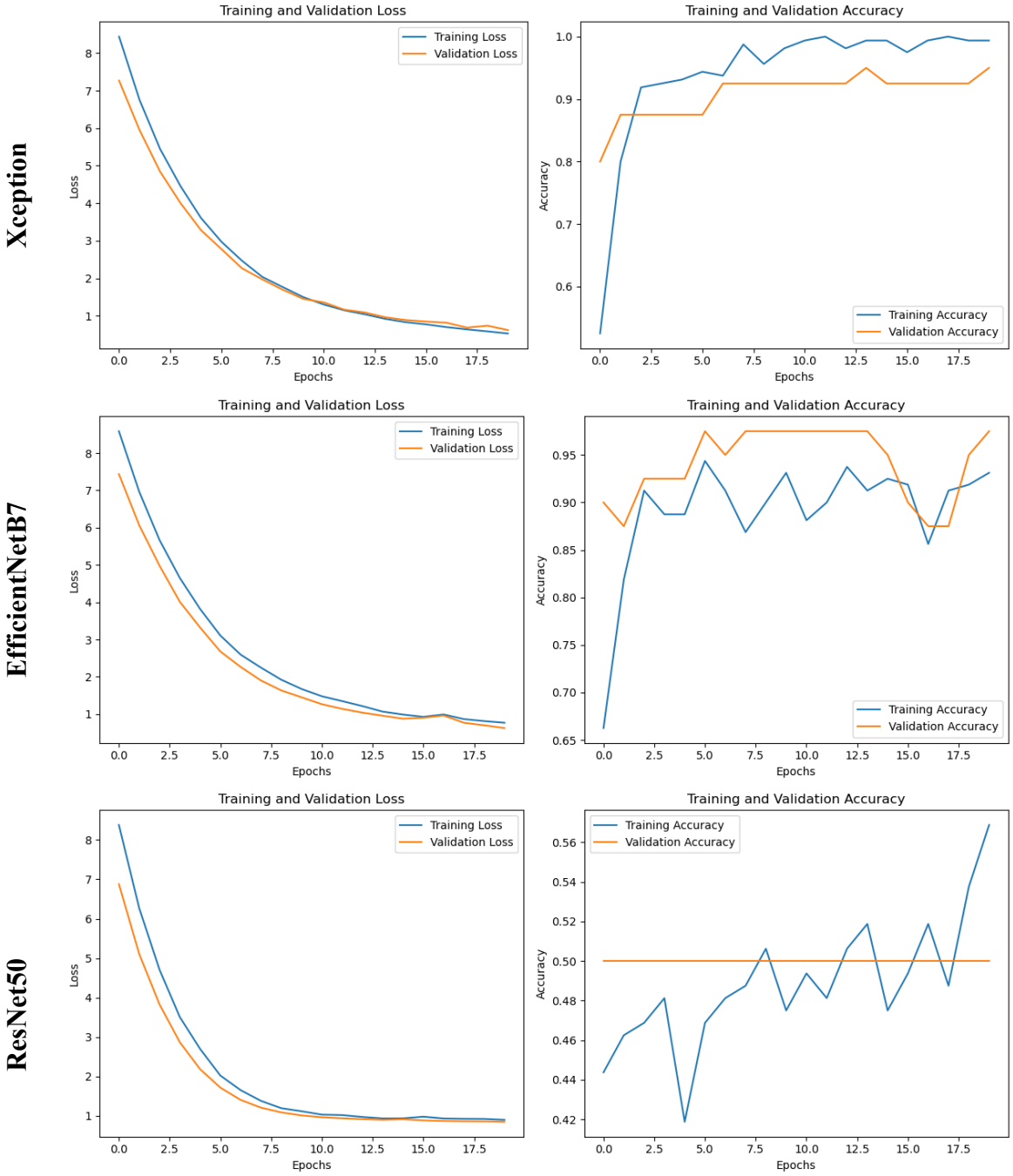


Figure 6.23: Loss and accuracy vs. epoch in BI-RADS 5 and 6.

Additionally, the confusion matrix values, performance measures results, and ROC curve were also calculated and illustrated in Figure (6.24).

In DenseNet121:

The four main values in the matrix were as follows: TP = 20, FN = 0, FP = 0, and TN = 20. That is, the model succeeded in classifying 20 BI-RAD 6 images out of 40, which is True Positive, while misclassified none, called False Negative. Additionally, the classifier successfully identified 20 BI-RAD 5 images which are called True Negative, while misidentified none, which is called False Positive. Moreover, the accuracy was 1 which means that 100% of the predicted images were identified correctly, and that indicates overfitting. Moreover, the recall (TPR) was equal to 1. That is, 100% of BI-RAD 6 images were correctly recognized as BI-RAD 6 by the classifier. In the same way, the FPR was zero, which means that there is a 0% chance that the model will incorrectly identify the BI-RAD 5 class as a BI-RAD 6 class. By the same token, precision resulted in 1, which is the rate of correctly predicted BI-RAD 6 class to all predicted BI-RAD 6. Furthermore, the F1-score was equal to 1. Finally, the area under the curve (AUC) was equal to 1.

In VGG16:

The four main values in the matrix were as follows: TP = 18, FN = 2, FP = 2, and TN = 18. That is, the model succeeded in classifying 18 BI-RAD 6 images out of 40, which is the True Positive, while misclassified only 2 and classified them as BI-RAD 5, called False Negative. Moreover, the model successfully identified 18 BI-RAD 5 images which are called True Negative, while misclassified 2 and recognized them as BI-RAD 6, which is called False Positive. Moreover, the accuracy was 0.90 which means that 90% of the predicted instances were identified correctly. Furthermore, the recall (TPR) was equal to the accuracy with a value of 0.90. That is, 90% of BI-RAD 6 images were correctly classified as BI-RAD 6 by the model. In contrast, the FPR was 0.1, which means that there is a 10% chance that the model will incorrectly classify BI-RAD 5 as

BI-RAD 6. Similarly, precision resulted in 0.90, and F1-score was also equal to 0.90. Finally, the AUC was equal to accuracy with a value of 0.90.

In VGG19:

The four main values in the matrix were as follows: TP = 16, FN = 4, FP = 2, and TN = 18. That is, the model succeeded in identifying 16 BI-RAD 6 images out of 40, which is True Positive, while misclassified 4 and identified them as BI-RAD 5, called False Negative. Likewise, the model successfully identified 18 BI-RAD 5 images which are called True Negative, while misclassified only 2 and classified them as BI-RAD 6, which is called False Positive. In addition, the accuracy was 0.85 which means that 85% of the predictions were classified correctly. Also, the recall (TPR) was less than the accuracy with a value of 0.80. That is, 80% of BI-RAD 6 images were correctly classified as BI-RAD 6 by the model. In contrast, the FPR was 0.1, which means that there is a 10% percentage that the model will incorrectly classify BI-RAD 5 as BI-RAD 6. Also, the precision was 0.89, and F1-score was equal to 0.84. Finally, the AUC was equal to the accuracy.

In Xception:

The four main values in the matrix were as follows: TP = 18, FN = 2, FP = 0, and TN = 20. That is, the model succeeded in classifying 18 BI-RAD 6, which is True Positive, while misclassified two instances and identified them as BI-RAD 5, called False Negative. Likewise, the classifier successfully recognized 20 BI-RAD 5 images which are all of them and that is called True Negative, while misidentified none. By the same token, the accuracy was 0.95 which means that 95% of the outcome images were classified correctly. In addition, the recall (TPR) was equal to 0.9. That is, 90% of BI-

RAD 6 images were correctly classified as BI-RAD 6 by the model. On the other hand, the FPR was zero. In addition, precision resulted in 1.00, and F1-score was equal to 0.95. Lastly, the area under the curve (AUC) was 0.95.

In EfficientNetB7:

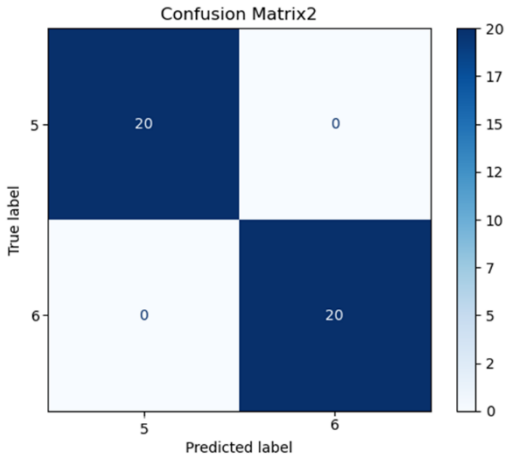
The four main values in the matrix were as follows: TP = 20, FN = 0, FP = 1, and TN = 19. That is, the model succeeded in classifying 20 BI-RAD 6 images out of 40, which is True Positive, while misclassified none and that is called False Negative. Likewise, the model successfully identified 19 BI-RAD 5 images which are called True Negative, while misidentified only one image, which is called False Positive. Furthermore, the accuracy resulted in 0.97 which means that 97% of the outputs were identified correctly. Also, the recall (TPR) was equal to 1. That is, 100% of BI-RAD 6 images were correctly classified as BI-RAD 6 by the model. In contrast, the FPR was 0.05, meaning that there is a 5% chance that the model will incorrectly recognize BI-RAD 5 as BI-RAD 6. Similarly, precision resulted in 0.95, and F1-score was also equal to 0.98. Finally, the AUC was equal to 0.97.

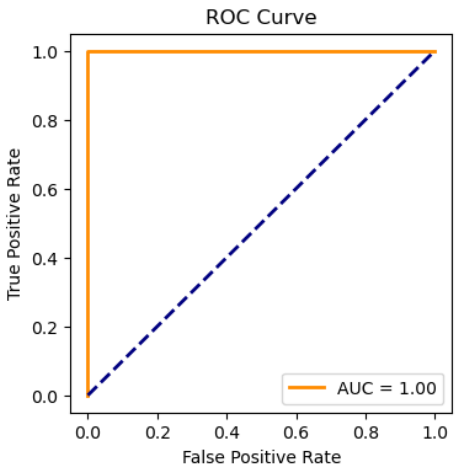
In ResNet50:

The four main values in the matrix were as follows: TP = 0, FN = 20, FP = 0, and TN = 20. That is, the model didn’t succeed in classifying any of the BI-RAD 6 images, which is True Positive, while misidentified 20, which is called False Negative. Likewise, the model successfully identified 20 BI-RAD 5 images which are called True Negative, while misclassified none, which is called False Positive. Additionally, the accuracy was 0.5 meaning that only 50% of the predictions were identified correctly. Also, the recall (TPR) was equal to zero. That is, 0% of BI-RAD 6 images were correctly classified as

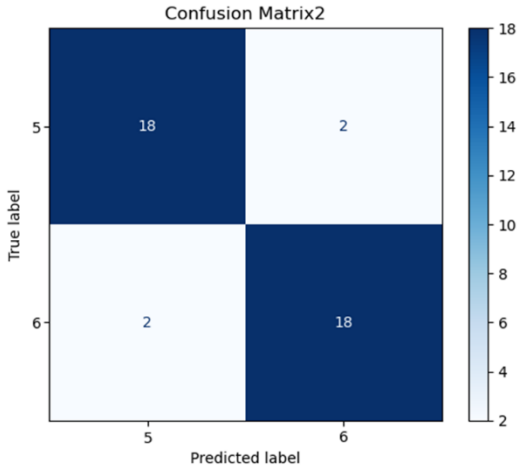
BI-RAD 6 by the classifier. Comparatively, the FPR was also equal to zero, meaning that there is a 0% percentage that the model will incorrectly recognize BI-RAD 5 as BI-RAD 6. In contrast, precision resulted in zero, and the F1-score was also equal to zero. Finally, the AUC was 0.5.

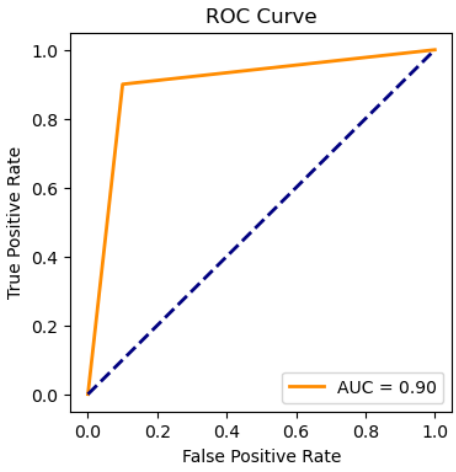
DenseNet121



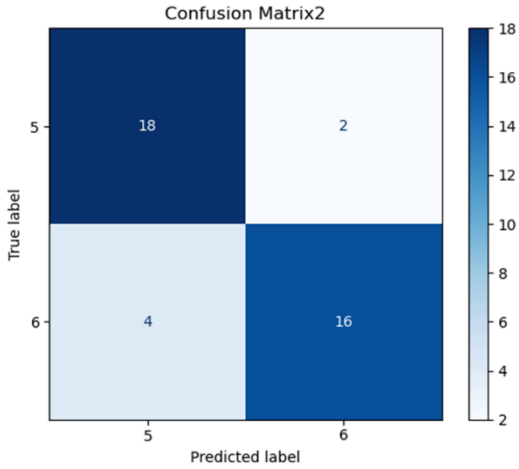


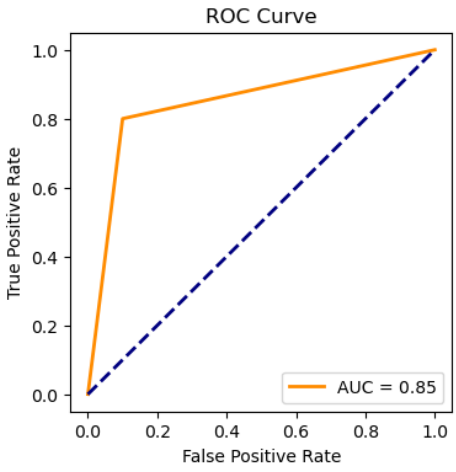
VGG16





VGG19





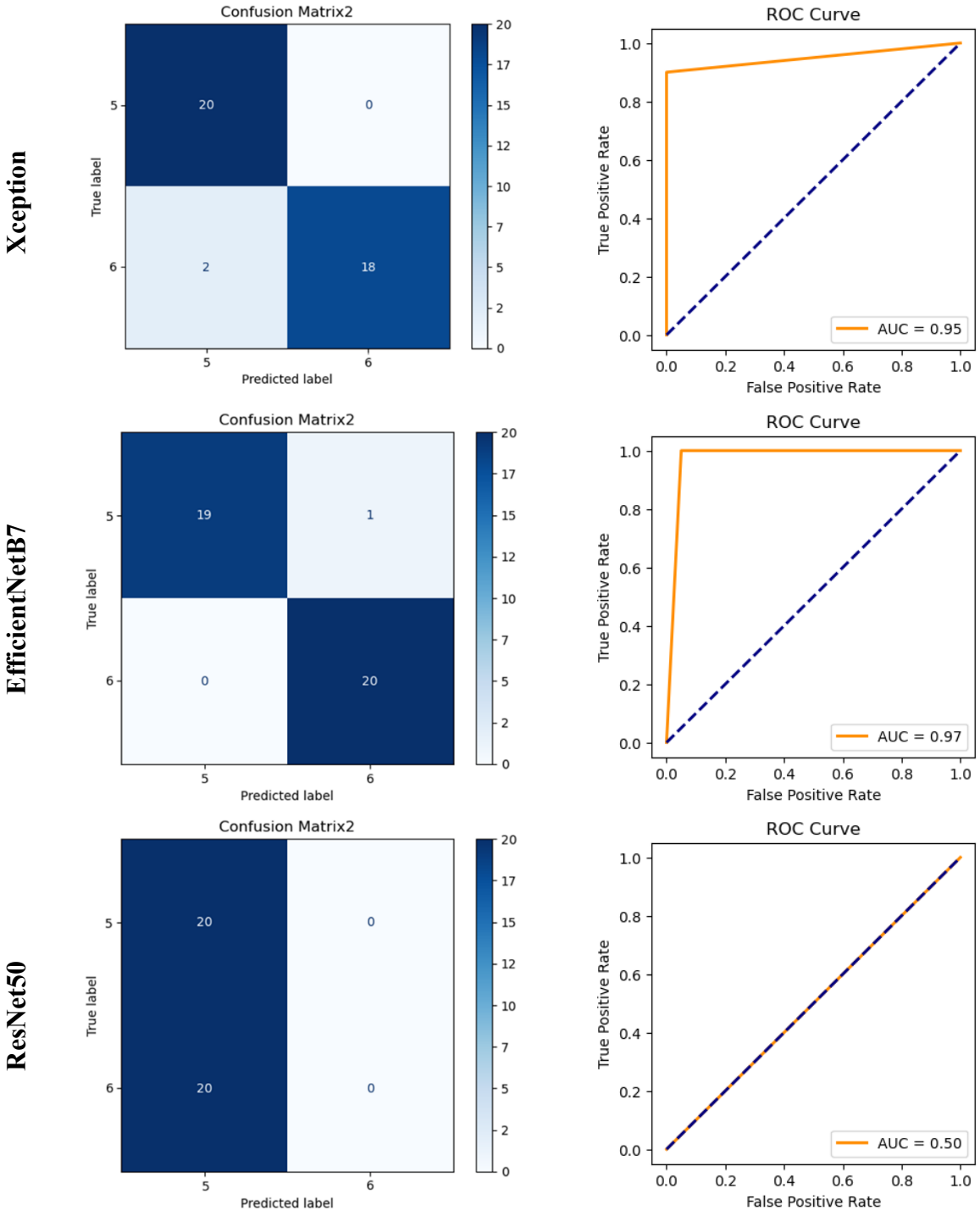


Figure 6.24: Confusion matrix and AUC in BI-RADS 5 and 6.

Furthermore, Table (6.4) below summarizes all obtained results in this use case namely, accuracy, recall, false positive rate, precision, f1-score, and ROC curve.

Table 6.4: Performance measures for all models in BI-RADS 5 and 6.

Model	Test Accuracy	Recall (TPR)	FPR	Precision	F1-score	ROC-AUC
DenseNet121	1.00	1.00	0.00	1.00	1.00	1.00
VGG16	0.90	0.90	0.10	0.90	0.90	0.90
VGG19	0.85	0.80	0.10	0.89	0.84	0.85
Xception	**0.95**	**0.90**	**0.00**	**1.00**	**0.95**	**0.95**
EfficientNetB7	**0.97**	**1.00**	**0.05**	**0.95**	**0.98**	**0.97**
ResNet50	0.50	0.00	0.00	0.00	0.00	0.50

6.3.4 BI-RAD 3 vs. BI-RAD 5

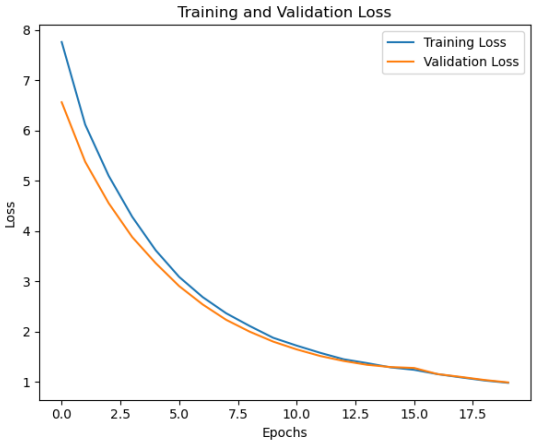
In the fourth use case, a dataset of BI-RAD 3 & BI-RAD 5 classes were fed into the models, which are non-consecutive scores in the BI-RADS scale. Figure (6.25) illustrates the results of loss function vs. epochs and accuracy vs epochs across all used models. In DenseNet121, the training and validation loss curves have similar behavior as previous models. The training loss curve started at the value of 7.7 and decreased gradually until it reached the final value of 0.98 in epoch 20. Comparatively, the validation curve started at 6.5 and decreased as epochs progressed until it reached the final value of 0.99 in the final epoch. The two curves intersect after epoch 11. On the other hand, the training accuracy had an initial value of 0.48 and then started to increase rapidly until it reached the final value of 1.00 which indicates overfitting. Besides, the validation accuracy curve started at 0.55 and gradually increased after the first epoch until it reached the same value obtained in the training which is 1 in the last iteration. Moreover, the VGG16 training loss curve started with a value of 5.7 and then began to gradually decrease as the epochs progressed until it reached 0.7. In contrast, the validation curve started at 5.1 and showed the same decreasing rate until the final value of 0.77. In addition, the training accuracy curve started at the value of 0.48 and then increased gradually as epochs progressed with some fluctuations after epoch 6 until the final value of 0.84. In the same way, the validation accuracy started at a high value of

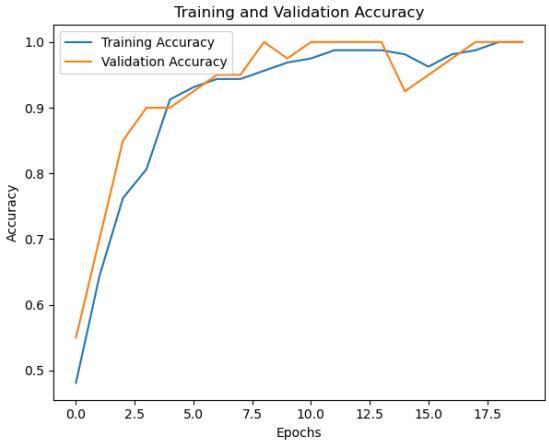
0.50 and then increased slightly after that until the final value of 0.75 which is less than the training accuracy.

In VGG19, the loss function values in training and validation are very similar to VGG16, where the validation's loss and accuracy were 0.79 and 0.72 in the last epoch respectively. Moreover, the training accuracy started at 0.53 and then increased after that until the final value of 0.73. In contrast, the validation accuracy showed similar behavior but with a lower result of 0.72 as the final value in the last iteration. Both training and validation accuracy curves had significant fluctuations as epochs progressed. Similarly, the loss curve initial values in the Xception of training and validation loss were 8.5 and 7.3, while the final values were 0.64 and 0.73 in order. By the same token, the training accuracy started at 0.52 and rapidly increased after that 0.99. In contrast, the validation accuracy started at a higher value of 0.82 and slightly increased to 0.92. Xception obtained similar results as in BI-RADS 5 and 6.

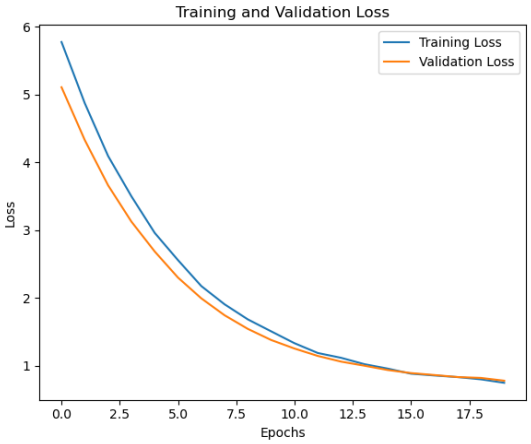
In EfficientNetB7, the initial values of training and validation loss were 8.7 and 7.6, and the final values were 0.73 and 0.59 in order. On the other hand, the initial value of the training accuracy was 0.59 then increased rapidly to the final value of 0.91, while the validation had a higher value of 0.85 as the initial accuracy and 1.00 as the final result. Lastly, the results in ResNet50 were again the lowest in this section. The loss curves were similar to the previous curves but differentiated in accuracy. The training accuracy initial value was 0.49 then had many ups and downs until it had the final value equal to the initial. Besides, the validation accuracy had a value of 0.5 and didn't improve at all as iterations progressed and this indicates a poor data classifier.

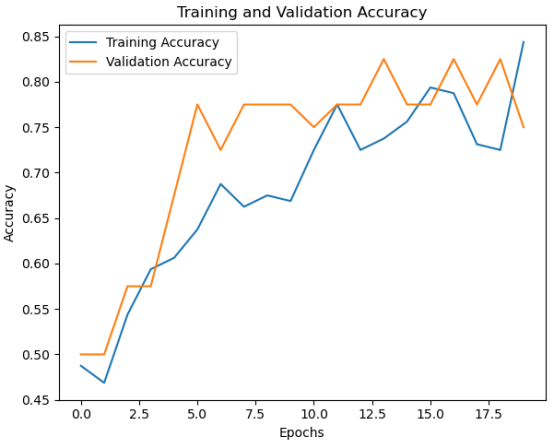
DenseNet121



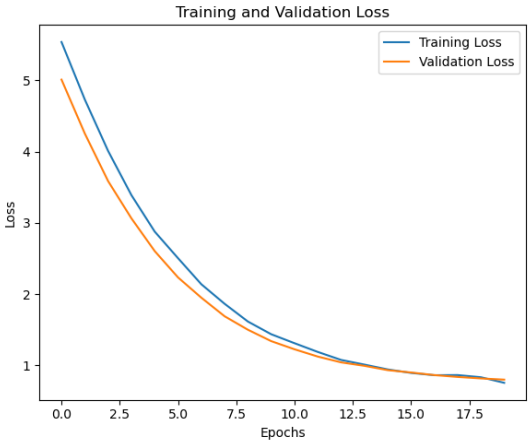


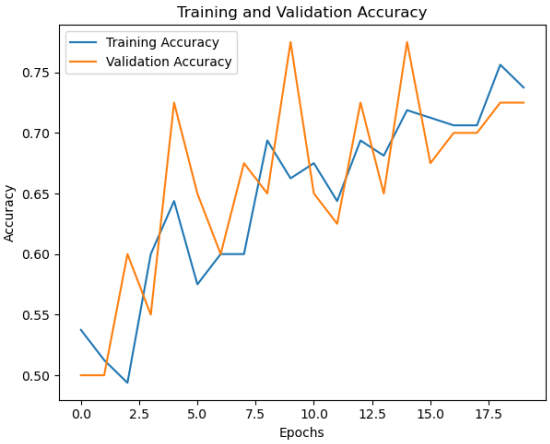
VGG16





VGG19





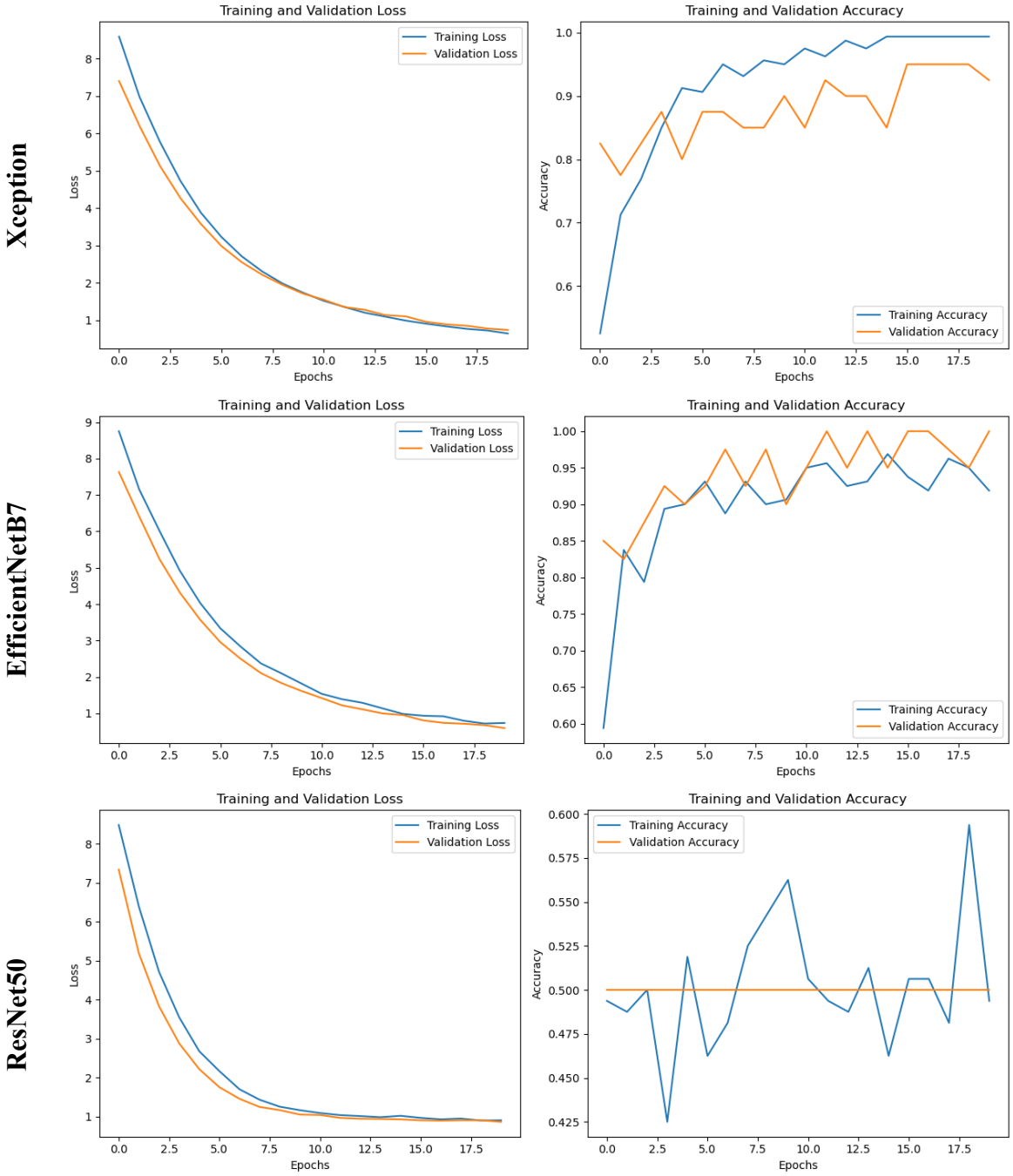


Figure 6.25: Loss and accuracy vs. epoch in BI-RADS 3 and 5.

By the same token, the confusion matrix values, performance measures results, and ROC curve were also calculated and provided in Figure (6.26).

In DenseNet121:

The four main values in the matrix were as follows: TP = 20, FN = 0, FP = 0, and TN = 20. That is, the classifier succeeded in detecting 20 BI-RAD 5 images out of 40, which is True Positive, while misidentified none, called False Negative. Moreover, the classifier successfully identified 20 BI-RAD 3 images which are called True Negative, while misidentified none, which is called False Positive. Furthermore, the accuracy was equal to 1 which means that 100% of the output images were identified correctly, and that means a potential overfitting. Additionally, the recall (TPR) was equal to 1. That is, 100% of BI-RAD 5 images were correctly classified as BI-RAD 5 by the classifier. Similarly, the FPR was zero, which means that there is a 0% chance that the classifier will incorrectly identify the BI-RAD 3 class as a BI-RAD 5 class. Furthermore, precision resulted in 1, which is the rate of correctly predicted BI-RAD 5 class to all predicted BI-RAD 5. Moreover, the F1-score was equal to 1. Finally, the area under the curve (AUC) was also equal to 1.

In VGG16:

The four main values in the matrix were as follows: TP = 15, FN = 5, FP = 5, and TN = 15. That is, the model succeeded in classifying 15 BI-RAD 5 images out of 40, which is the True Positive, while misclassified 5 and classified them as BI-RAD 3, called False Negative. Moreover, the model successfully identified 15 BI-RAD 3 images which are called True Negative, while misclassified 5 and recognized them as BI-RAD 5, which is called False Positive. In addition, the accuracy was 0.75 which means that 75% of the predicted outputs were classified correctly. Similarly, the recall (TPR) was equal to the accuracy with a value of 0.75. That is, 75% of BI-RAD 5 images were correctly identified as BI-RAD 5 by the classifier. In contrast, the FPR was 0.25, which means that there is a 25% chance that the model will incorrectly classify BI-RAD 3 as BI-RAD

5. Likewise, precision resulted in 0.75, and F1-score was also equal to 0.75. Lastly, the AUC was equal to accuracy with a value of 0.75.

In VGG19:

The four main values in the matrix were as follows: TP = 15, FN = 5, FP = 6, and TN = 14. That is, the model succeeded in classifying 15 BI-RAD 5 images out of 40, which is True Positive, while misclassified 5 and classified them as BI-RAD 3, called False Negative. Likewise, the classifier successfully identified 14 BI-RAD 3 images which are called True Negative, while misclassified 6 and identified them as BI-RAD 5, which is called False Positive. Additionally, the accuracy was 0.73 which means that 73% of the predictions were recognized correctly. Besides, the recall (TPR) was higher than the accuracy with a value of 0.75. That is, 75% of BI-RAD 5 images were correctly diagnosed as BI-RAD 5 by the model. Comparatively, the FPR was 0.3, which means that there is a 30% chance that the model will incorrectly diagnose BI-RAD 3 as BI-RAD 5. Also, the precision was 0.71, the F1-score was equal to 0.73, and AUC was equal to 0.72.

In Xception:

The four main values in the matrix were as follows: TP = 18, FN = 2, FP = 1, and TN = 19. That is, the model succeeded in classifying 18 BI-RAD 5, which is True Positive, while misidentified two instances and identified them as BI-RAD 3, called False Negative. Moreover, the classifier successfully diagnosed 19 BI-RAD 3 images which is called True Negative, while misidentified none. Similarly, the accuracy was 0.93 which means that 93% of the outcome images were diagnosed correctly. In addition, the recall (TPR) was equal to 0.9. That is, 90% of BI-RAD 5 images were correctly diagnosed as

BI-RAD 5 by the classifier. In contrast, the FPR was 0.05, the precision resulted in 0.95, and the F1-score was equal to 0.92. Finally, the area under the curve (AUC) was 0.92.

In EfficientNetB7:

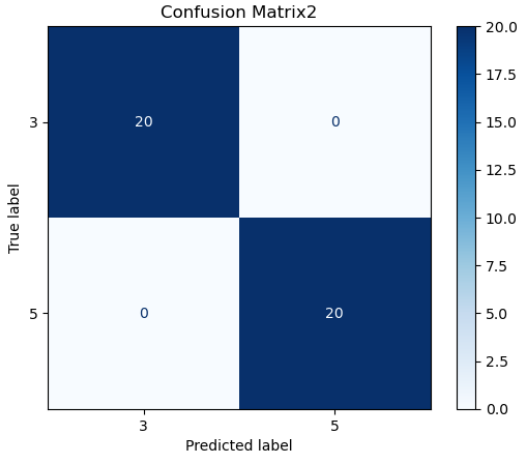
The four main values in the matrix were as follows: TP = 20, FN = 0, FP = 0, and TN = 20. That is, the model succeeded in classifying 20 BI-RAD 5 images out of 40, which is True Positive, while misclassified none and that is called False Negative. Likewise, the model successfully identified 20 BI-RAD 3 images which are called True Negative, while misidentified none, which is called False Positive. Moreover, the accuracy resulted in 1 which means that 100% of the predictions were identified correctly. Besides, the recall (TPR) was equal to 1. That is, 100% of BI-RAD 5 images were correctly classified as BI-RAD 5 by the model. In contrast, the FPR was zero, meaning that there is a 0% chance that the model will incorrectly recognize BI-RAD 3 as BI-RAD 5. Likewise, precision resulted in 1, the F1-score was also equal to 1, and the AUC was equal to 1. This model indicates overfitting in this user case.

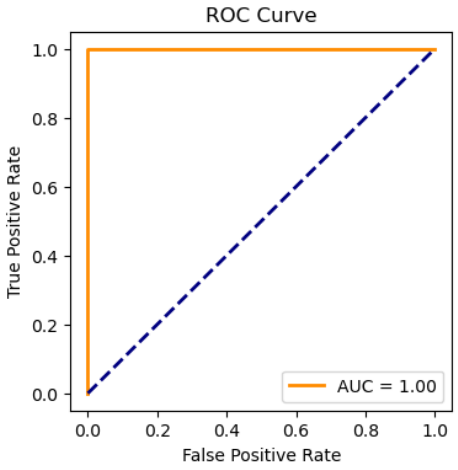
In ResNet50:

The four main values in the matrix were as follows: TP = 20, FN = 0, FP = 20, and TN = 0. That is, the model succeeded in classifying all of the BI-RAD 5 images, which is True Positive, while misidentified none, which is called False Negative. In addition, the classifier didn’t succeed in identifying any of the BI-RAD 3 images which are called True Negative, while misclassified all of them, which is called False Positive. Furthermore, the accuracy was 0.5 meaning that only 50% of the predictions were identified correctly. Also, the recall (TPR) was equal to 1. That is, 100% of BI-RAD 5 images were correctly classified as BI-RAD 5 by the classifier. Comparatively, the FPR

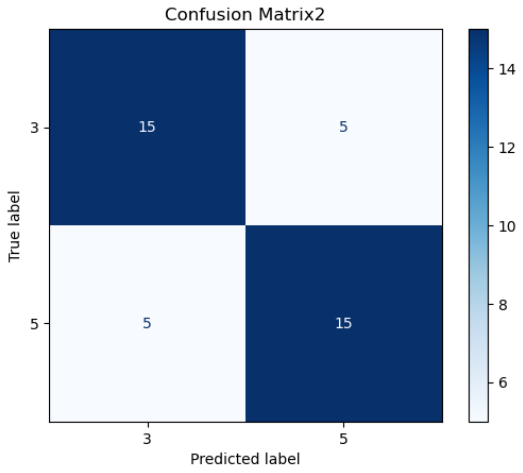
was also equal to 1, meaning that there is a 100% percentage that the model will incorrectly recognize BI-RAD 3 as BI-RAD 5. In contrast, precision resulted in 0.50, the F1-score was also equal to 0.67, and the AUC was 0.5.

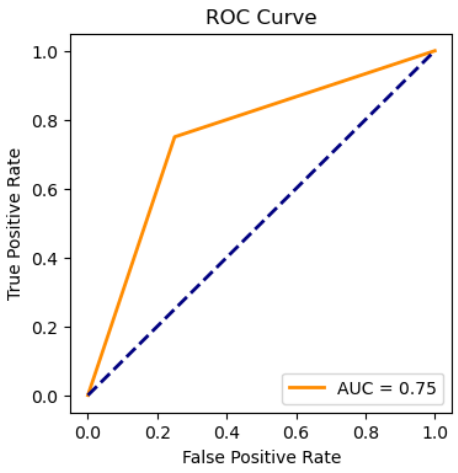
DenseNet121



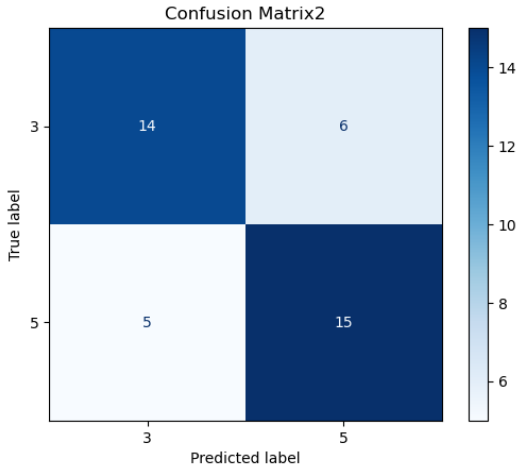


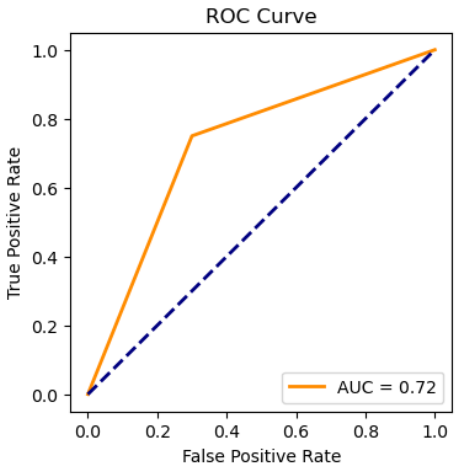
VGG16





VGG19





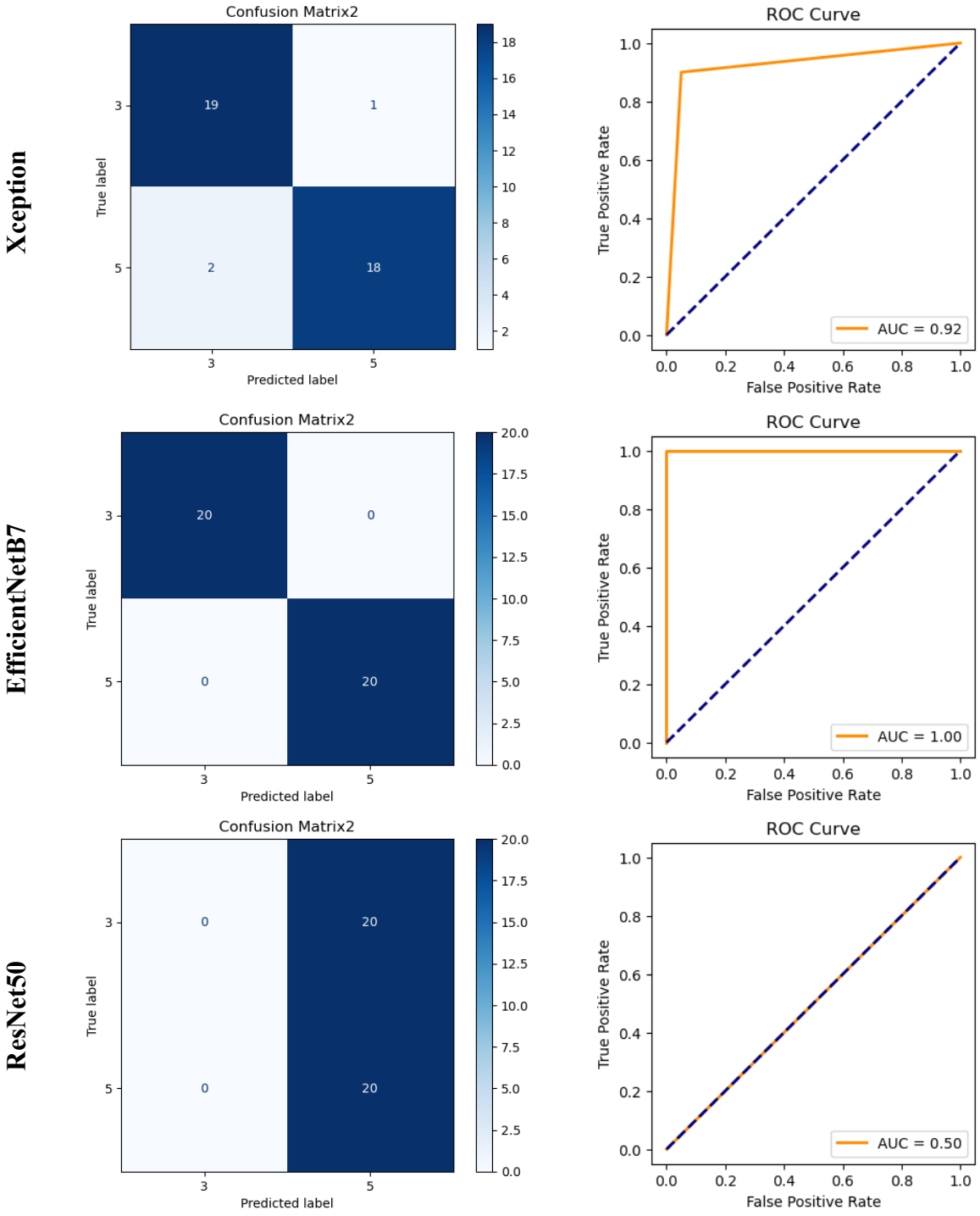


Figure 6.26: Confusion matrix and AUC in BI-RADS 3 and 5.

Moreover, Table (6.5) depicts all results obtained in this use case for BI-RADS 3 and 5. The results were calculated from the confusion matrices above.

Table 6.5: Performance measures for all models in BI-RADS 3 and 5.

Model	Test Accuracy	Recall (TPR)	FPR	Precision	F1-score	ROC-AUC
DenseNet121	1.00	1.00	0.00	1.00	1.00	1.00
VGG16	0.75	0.75	0.25	0.75	0.75	0.75
VGG19	0.73	0.75	0.30	0.71	0.73	0.72
Xception	**0.93**	**0.90**	**0.05**	**0.95**	**0.92**	**0.92**
EfficientNetB7	1.00	1.00	0.00	1.00	1.00	1.00
ResNet50	0.50	1.00	1.00	0.50	0.67	0.50

6.3.5 BI-RAD 3 vs. BI-RAD 6

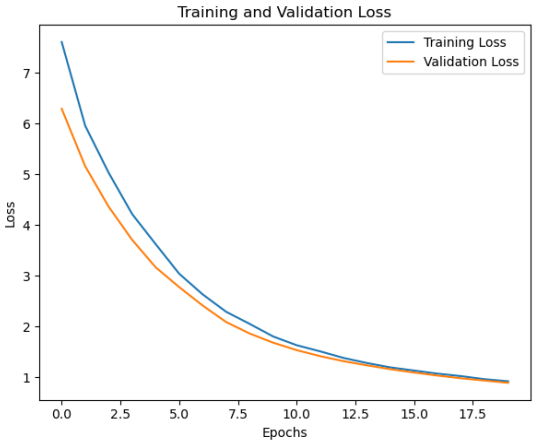
In the fifth use case, a dataset of BI-RAD 3 & BI-RAD 6 classes was fed into the models, which are non-consecutive scores in the BI-RADS scale. Figure (6.27) presents the obtained results of loss and accuracy curves across all used models. In DenseNet121, the training loss curve started at the value of 7.6 and decreased gradually until it reached the final value of 0.92 in epoch 20. Comparatively, the validation curve started at 6.2 and decreased as iterations progressed until it reached the final value of 0.89 in the final epoch. The two curves intersect after epoch 12. In contrast, the training accuracy had an initial value of 0.54 and then started to increase rapidly until it reached the final value of 1.00 which indicates overfitting. Similarly, the validation accuracy curve started at 0.65 and gradually increased after the first epoch until it reached the same value obtained in the training which is 1 in the last iteration. In addition, the VGG16 training loss curve started with a value of 5.6 and then began to gradually decrease as the epochs progressed until it reached 0.7. Likewise, the validation curve started at 5.02 and showed the same decreasing rate until the final value of 0.6. Moreover, the training accuracy curve started at the value of 0.48 and then increased gradually as epochs progressed with some fluctuations after the second epoch until the final value of 0.80. Comparatively, the validation accuracy started at a high value of

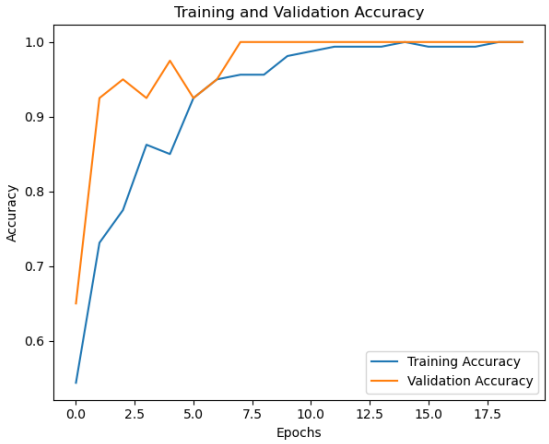
0.50 and then increased gradually after that until the final value of 0.90 which is higher than the training accuracy.

In VGG19, the loss function values in training and validation are very similar to VGG16. Moreover, the training accuracy started at 0.46 and then increased after that until the final value of 0.75. On the other hand, the validation accuracy showed similar behavior but with a higher result of 0.87 as the final value in epoch 20. Both training and validation accuracy curves had significant fluctuations as epochs progressed. Likewise, the loss curve initial values in the Xception of training and validation loss were 8.2 and 7.1, while the final values were 0.38 and 0.43 respectively. In the same way, the training accuracy started at 0.60 and increased after that to 1.00 which suggests overfitting. In contrast, the validation accuracy started at a higher value of 0.90 and slightly increased to 0.95.

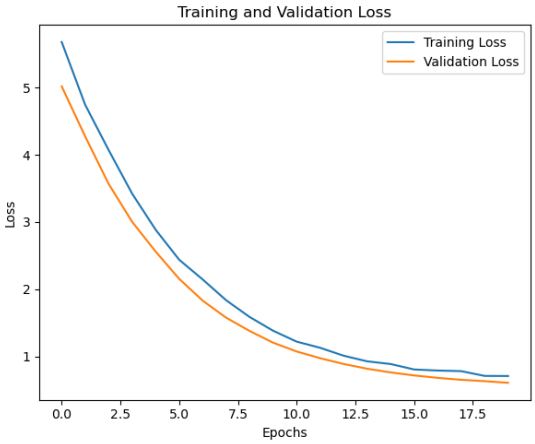
In EfficientNetB7, the initial values of training and validation loss were 9.3 and 8.3, and the final values were 0.76 and 0.70 in order. Furthermore, the initial value of the training accuracy was 0.47 then increased rapidly to the final value of 0.97, while the validation accuracy was 0.50 in the first iteration and increased to 0.97 as the final result. Finally, the results in ResNet50 were low. The loss curves were similar to the previous models. Moreover, the training accuracy initial value was 0.52 then had many ups and downs until it reached 0.53 which it didn’t improve much. Besides, the validation accuracy had a value of 0.5 and didn’t improve at all as epochs progressed and this indicates a poor data classifier.

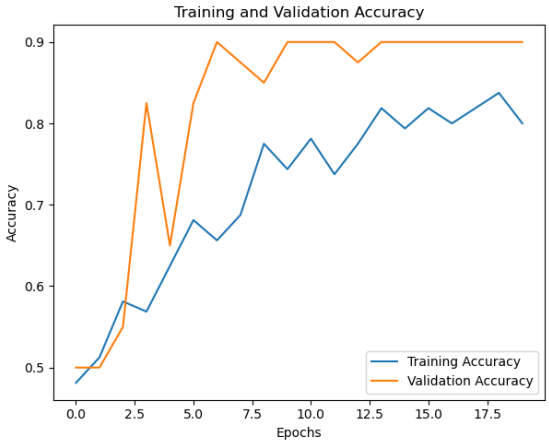
DenseNet121



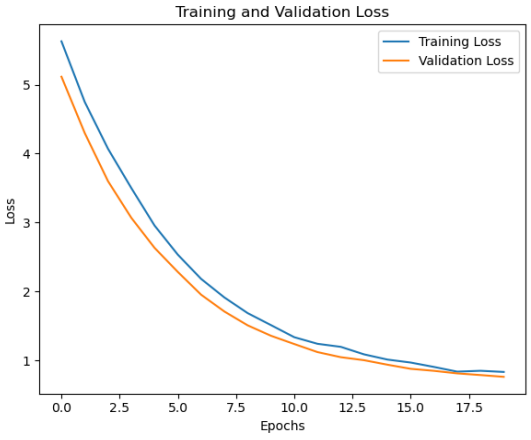


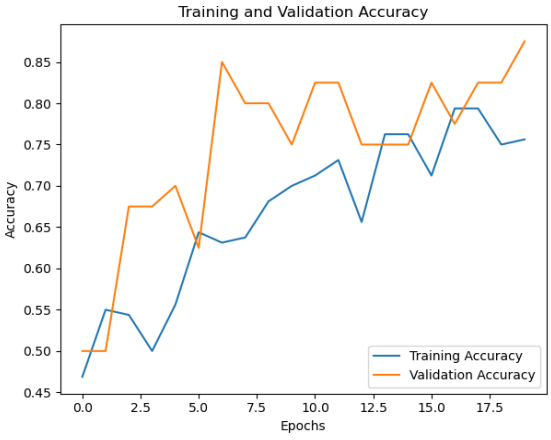
VGG16



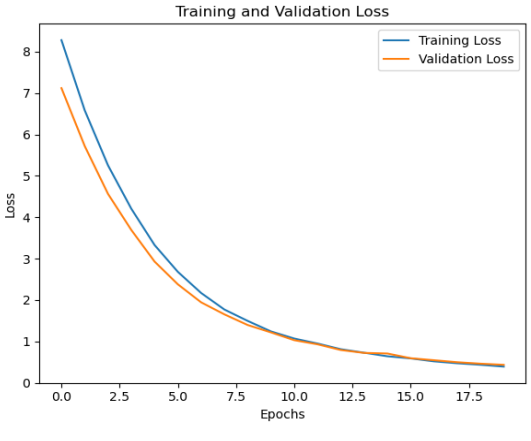


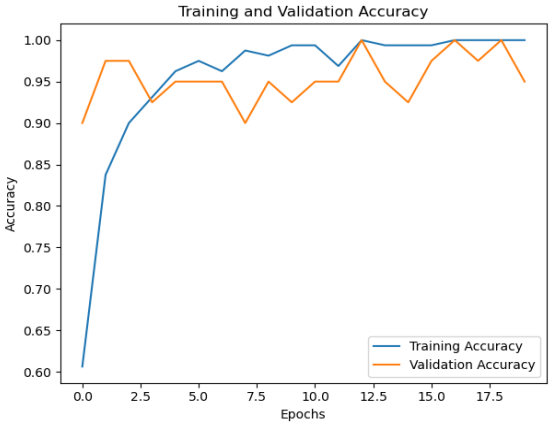
VGG19





Xception





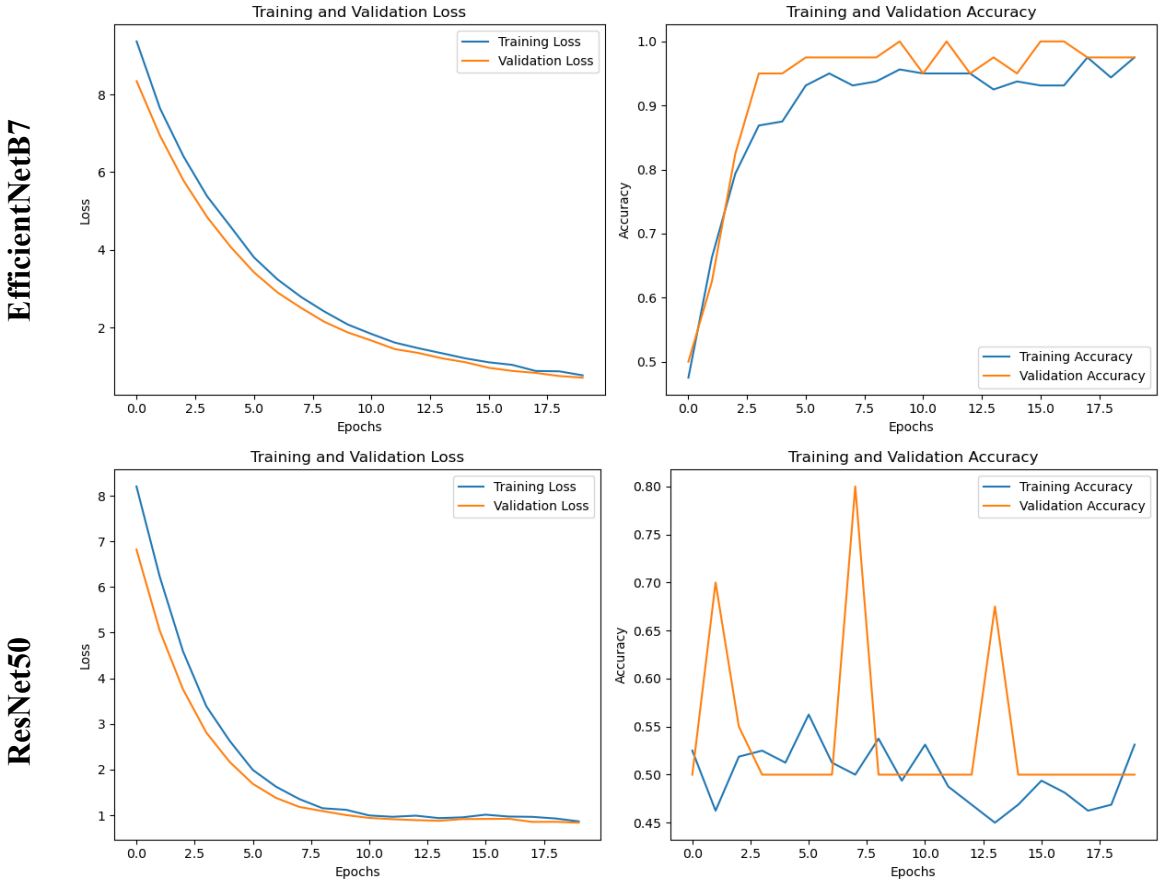


Figure 6.27: Loss and accuracy vs. epoch in BI-RADS 3 and 6.

By the same token, the confusion matrix values, performance measures results, and ROC curve were also calculated and provided in Figure (6.28).

In DenseNet121:

The four main values in the matrix were as follows: TP = 20, FN = 0, FP = 0, and TN = 20. That is, the classifier succeeded in detecting 20 BI-RAD 6 images out of 40, which is True Positive, while misidentified none, called False Negative. Furthermore, the classifier successfully classified 20 BI-RAD 3 images which are called True Negative, while misclassified none, which is called False Positive. Moreover, the accuracy was equal to 1 which means that 100% of the output images were identified correctly, and that means a potential overfitting. Similarly, the recall (TPR) was equal to 1. That is, 100% of BI-RAD 6 images were correctly classified as BI-RAD 6 by the model. In

contrast, the FPR was zero, which means that there is a 0% chance that the classifier will incorrectly identify the BI-RAD 3 class as a BI-RAD 6 class. In addition, precision resulted in 1, which is the rate of correctly predicted BI-RAD 6 class to all predicted BI-RAD 6. Finally, the F1-score was equal to 1, and the area under the curve (AUC) was also equal to 1.

In VGG16:

The four main values in the matrix were as follows: TP = 18, FN = 2, FP = 2, and TN = 18. That is, the model succeeded in classifying 18 BI-RAD 6 images out of 40, which is the True Positive, while misclassified 2 and classified them as BI-RAD 3, called False Negative. Similarly, the model successfully identified 18 BI-RAD 3 images which are called True Negative, while misclassified 2 and classified them as BI-RAD 6, which is called False Positive. Likewise, the accuracy was 0.90 which means that 90% of the predicted images were identified correctly. In addition, the recall (TPR) was equal to the accuracy with a value of 0.90. That is, 90% of BI-RAD 6 images were correctly diagnosed as BI-RAD 6 by the classifier. In contrast, the FPR was 0.01, which means that there is a 10% chance that the model will incorrectly classify BI-RAD 3 as BI-RAD 6. In the same way, precision resulted in 0.90, the F1-score was also equal to 0.90, and the AUC was equal to accuracy with a value of 0.90.

In VGG19:

The four main values in the matrix were as follows: TP = 19, FN = 1, FP = 4, and TN = 16. That is, the model succeeded in classifying 19 BI-RAD 6 images out of 40, which is True Positive, while misclassified only one image and classified it as BI-RAD 3, called False Negative. Similar to that, the classifier successfully identified 16 BI-RAD 3

images which are called True Negative, while misclassified 4 and identified them as BI-RAD 6, which is called False Positive. Additionally, the accuracy was 0.88 which means that 88% of the predictions were recognized correctly. Besides, the recall (TPR) was higher than the accuracy with a value of 0.95. That is, 95% of BI-RAD 6 images were correctly diagnosed as BI-RAD 6 by the model. Comparatively, the FPR was 0.2, which means that there is a 20% chance that the model will incorrectly diagnose BI-RAD 3 as BI-RAD 6. Also, the precision was 0.83, the F1-score was equal to 0.88, and AUC was equal to 0.88.

In Xception:

The four main values in the matrix were as follows: TP = 18, FN = 2, FP = 0, and TN = 20. That is, the model succeeded in classifying 18 BI-RAD 6, which is True Positive, while misidentified two instances and identified them as BI-RAD 3, called False Negative. Moreover, the classifier successfully diagnosed 20 BI-RAD 3 images which is called True Negative, while misidentified none. By the same token, the accuracy was 0.95 which means that 95% of the outcome images were diagnosed correctly. In addition, the recall (TPR) was equal to zero. That is, 0% of BI-RAD 6 images were correctly diagnosed as BI-RAD 6 by the classifier. In contrast, the FPR was zero, the precision resulted in 1, and the F1-score was equal to 0.95. Finally, the area under the curve (AUC) was 0.95.

In EfficientNetB7:

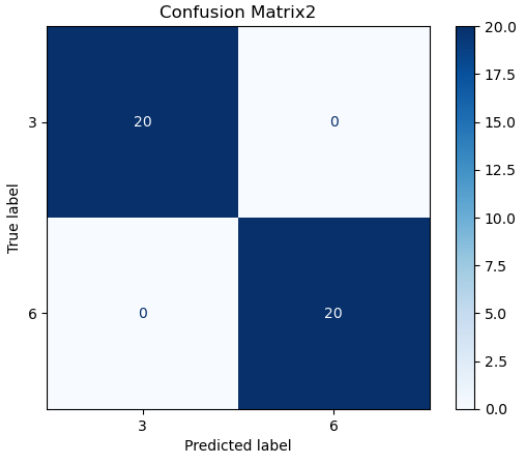
The four main values in the matrix were as follows: TP = 20, FN = 0, FP = 1, and TN = 19. That is, the model succeeded in classifying 20 BI-RAD 6 images out of 40, which is True Positive, while misclassified none and that is called False Negative. Likewise, the

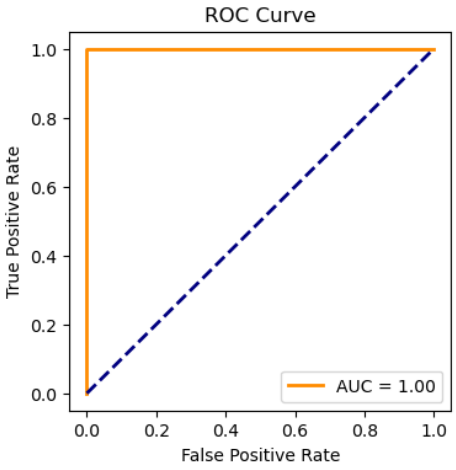
model successfully identified 19 BI-RAD 3 images which are called True Negative, while misidentified only one, which is called False Positive. Moreover, the accuracy resulted in 0.97 which means that 97% of the predictions were identified correctly. Besides, the recall (TPR) was equal to 1. That is, 100% of BI-RAD 6 images were correctly classified as BI-RAD 6 by the model. In contrast, the FPR was 0.05, meaning that there is a 5% chance that the model will incorrectly recognize BI-RAD 3 as BI-RAD 6. Likewise, precision resulted in 0.95, the F1-score was also equal to 0.98, and the AUC was equal to 0.97.

In ResNet50:

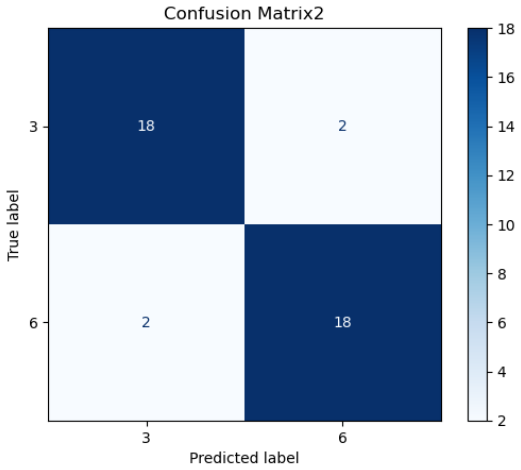
The four main values in the matrix were as follows: TP = 20, FN = 0, FP = 20, and TN = 0. That is, the model succeeded in classifying all of the BI-RAD 6 images, which is True Positive, while misidentified none, which is called False Negative. In addition, the classifier didn't succeed in identifying any of the BI-RAD 3 images which are called True Negative, while misclassified all of them, which is called False Positive. Moreover, the accuracy was 0.5 meaning that only 50% of the predictions were identified correctly. Also, the recall (TPR) was equal to 1. That is, 100% of BI-RAD 6 images were correctly diagnosed as BI-RAD 6 by the classifier. Comparatively, the FPR was also equal to 1, meaning that there is a 100% percentage that the model will incorrectly recognize BI-RAD 3 as BI-RAD 6. In contrast, precision resulted in 0.50, the F1-score was also equal to 0.67, and the AUC was 0.5.

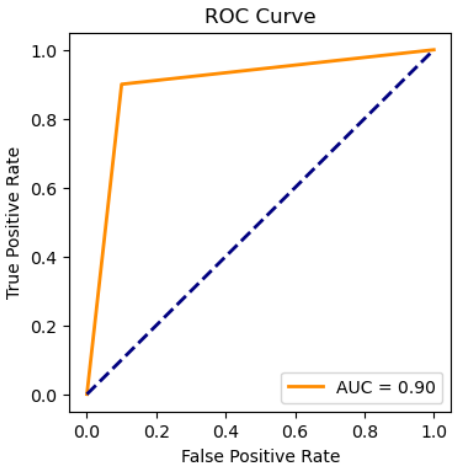
DenseNet121



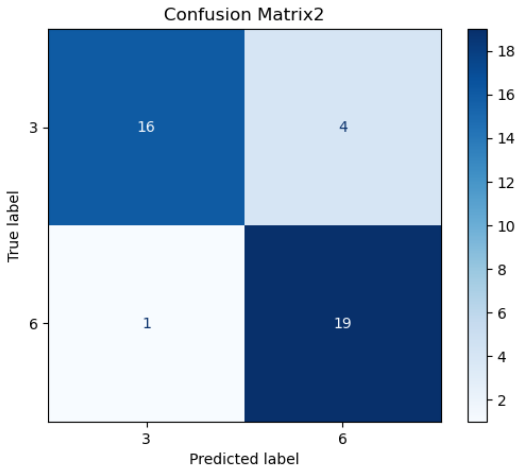


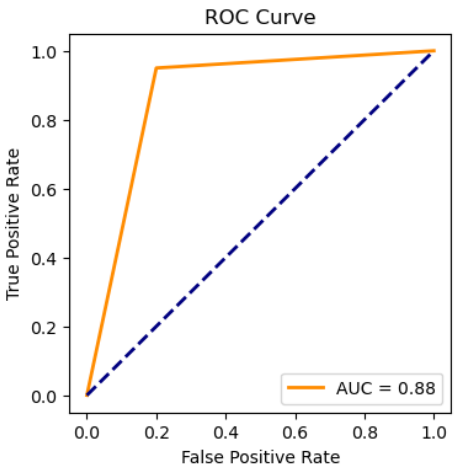
VGG16





VGG19





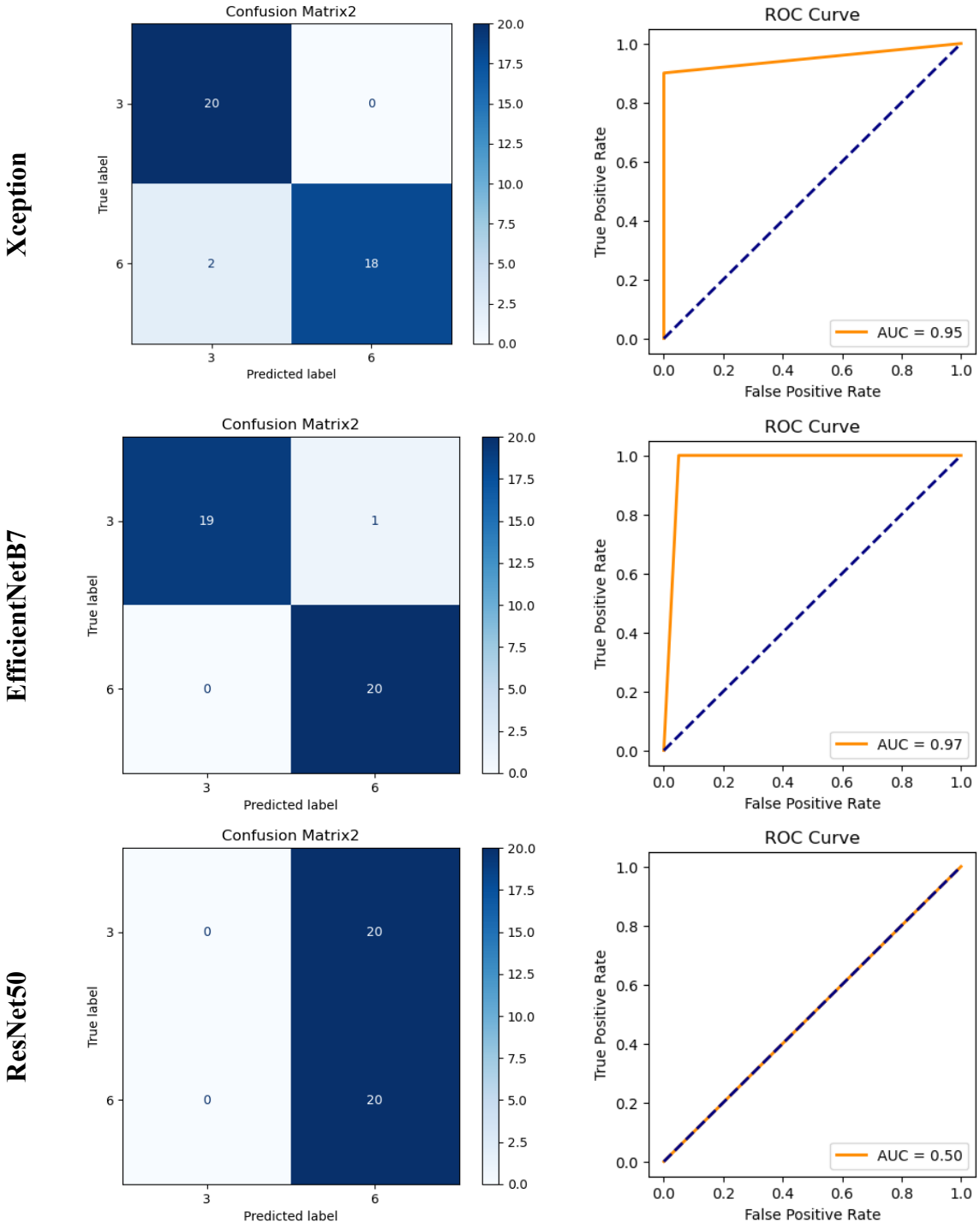


Figure 6.28: Confusion matrix and AUC in BI-RADS 3 and 6.

Table (6.6) below presents the summary results from this use case of classifying BI-RAD 3 and BI-RAD 6 which are non-consecutive scores in the BI-RADS scale.

Table 6.6: Performance measures for all models in BI-RADS 3 and 6.

Model	Test Accuracy	Recall (TPR)	FPR	Precision	F1-score	ROC-AUC
DenseNet121	1.00	1.00	0.00	1.00	1.00	1.00
VGG16	0.90	0.90	0.01	0.90	0.90	0.90
VGG19	0.88	0.95	0.20	0.83	0.88	0.88
Xception	**0.95**	**0.00**	**0.00**	**1.00**	**0.95**	**0.95**
EfficientNetB7	**0.97**	**1.00**	**0.05**	**0.95**	**0.98**	**0.97**
ResNet50	0.50	1.00	1.00	0.50	0.67	0.50

6.3.6 BI-RAD 4 vs. BI-RAD 6

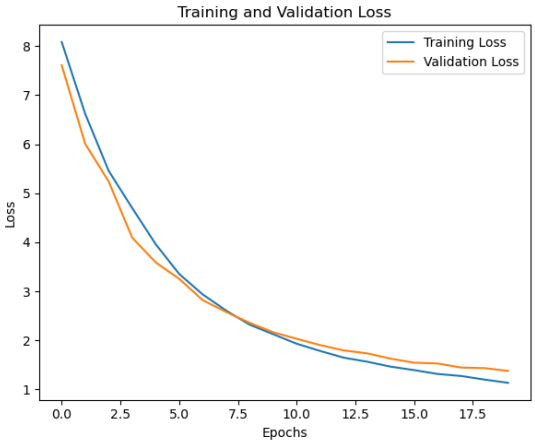
In the sixth and final use case, a dataset of BI-RAD 4 & BI-RAD 6 classes was fed into the models, which are also non-consecutive scores in the BI-RADS scale. Figure (6.29) provides the obtained results of loss and accuracy curves across all implemented models. In DenseNet121, the training loss curve started at the value of 8.08 and decreased gradually until it reached the final value of 1.13 in the last iteration of training. Similarly, the validation curve started at 7.6 and decreased as epochs progressed until it reached the final value of 1.37 in the final epoch. The two curves intersect between epochs 6 and 8. On the other hand, the training accuracy had an initial value of 0.49 and then started to increase gradually until it reached the final value of 0.9. Likewise, the validation accuracy curve started at 0.50 and gradually increased after the first epoch with some fluctuations until it reached the final value of 0.87 which is less than the training accuracy for this model. Additionally, the VGG16 training loss curve started with a value of 5.5 and then began to gradually decrease as the epochs progressed until it reached 0.8. Likewise, the validation curve started at 4.96 and showed the same decreasing rate until the final value of 0.8. Both curves intersect after epoch 11. Furthermore, the training accuracy curve started at the value of 0.50 and then increased gradually as epochs progressed with fluctuations after the second epoch until the final value of 0.65. In contrast, the validation accuracy started at the same value of

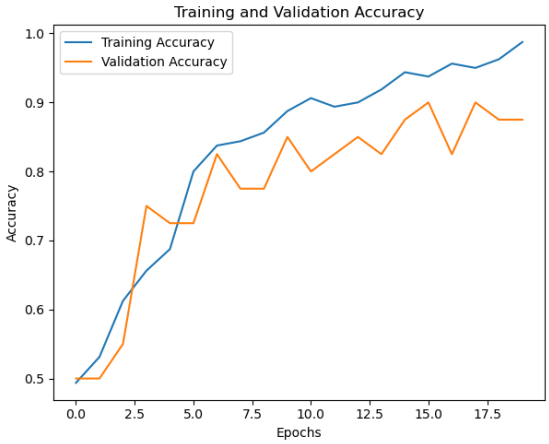
0.50 and then increased gradually after that until the final value of 0.70 which is higher than the training accuracy.

In VGG19, the loss function values in training and validation are almost equal to the VGG16. Moreover, the training accuracy started at 0.50 and reached 0.61 with unstable improvement as epochs progressed. On the other hand, the initial and final values of the validation accuracy were equal to 0.50 which indicates a poor fit model for this use case. Both training and validation accuracy curves had significant fluctuations and were unstable as iterations progressed. Similarly, the loss curve initial values in the Xception of training and validation loss were 8.4 and 7.5, while the final values were 0.83 and 1.5 respectively. By the same token, the training accuracy started at 0.48 and increased after that to 0.95. In contrast, the validation accuracy started at a higher value of 0.57 and slightly increased to 0.60. It didn't significantly improve and this model makes inaccurate classification of unseen data.

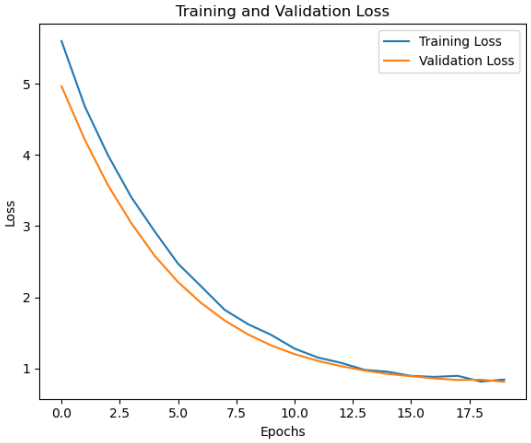
In EfficientNetB7, the initial values of training and validation loss were 8.8 and 7.8, and the final values were 1.02 and 1 in order. Furthermore, the initial value of the training accuracy was 0.53 then increased gradually to the final value of 0.88. In contrast, the validation accuracy was 0.67 in the first epoch and increased to 0.90 as the final result. Finally, the results obtained in ResNet50 were not as good as in the previous models. The loss curves were similar to the previous models. Additionally, the training accuracy initial value was 0.51 then had many ups and downs until it reached 0.46 which is less than the initial value. Also, the validation accuracy curve was not stable as epochs progressed and had a value of 0.5 as the initial and final value which indicates a poor data fit.

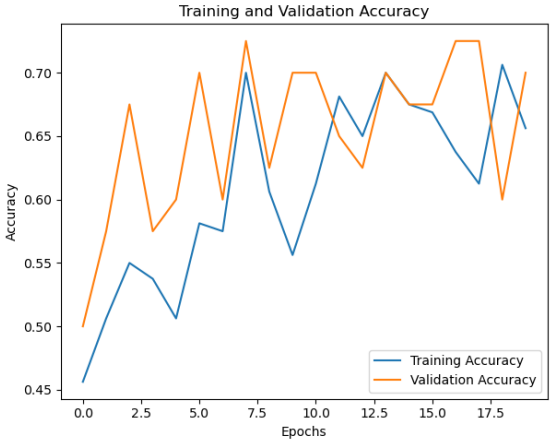
DenseNet121



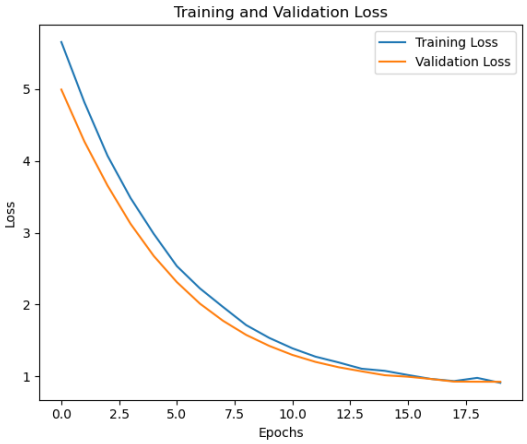


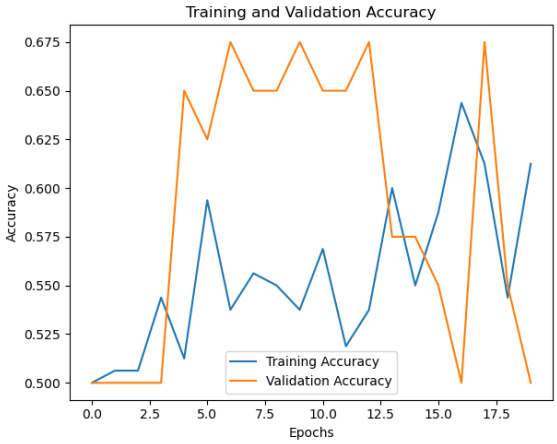
VGG16





VGG19





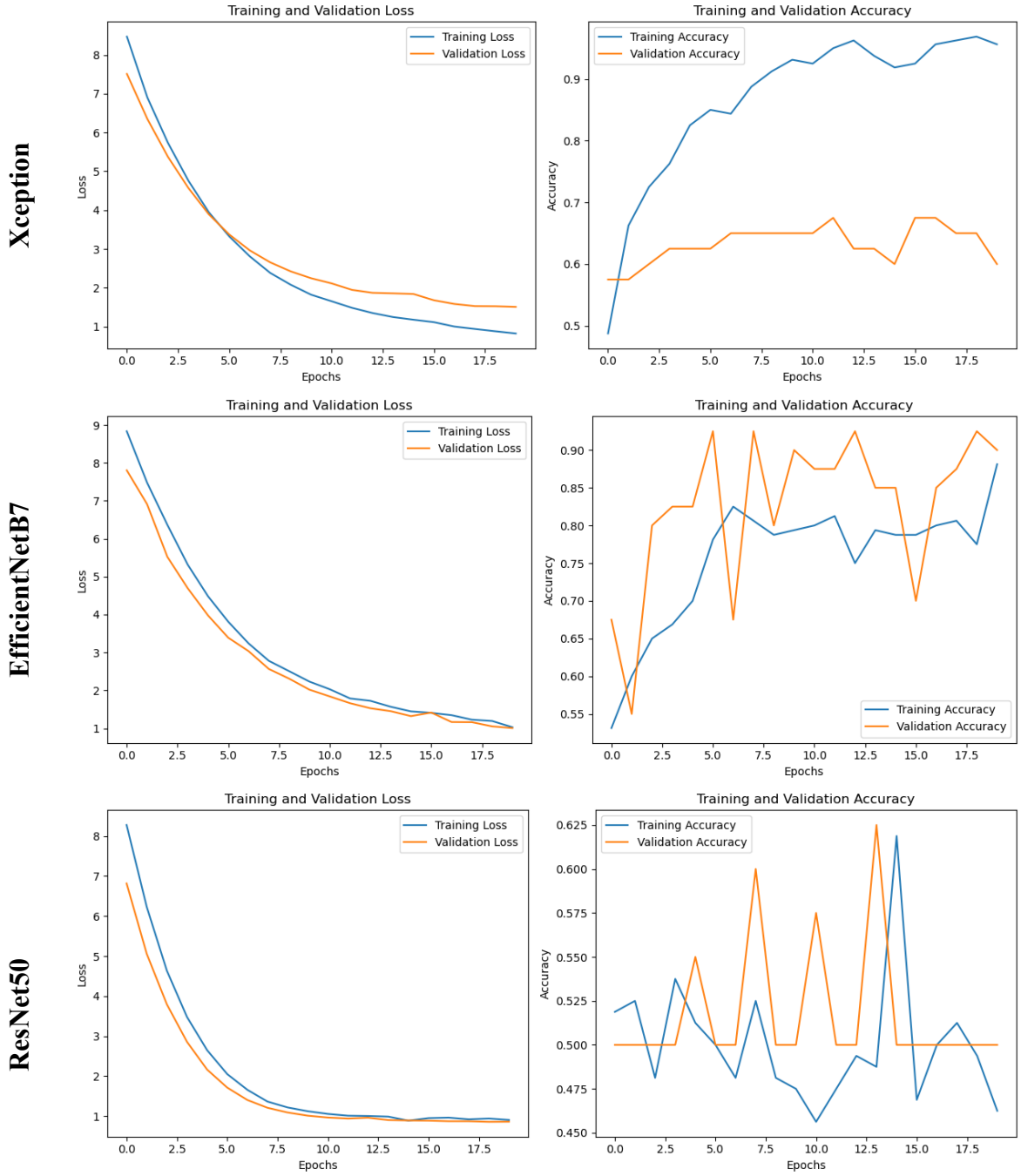


Figure 6.29: Loss and accuracy vs. epoch in BI-RADS 4 and 6.

In addition to the previous, the confusion matrix values, performance measures results, and ROC curve were also calculated and provided in Figure (6.30).

In DenseNet121:

The four main values in the matrix were as follows: TP = 19, FN = 1, FP = 4, and TN = 16. That is, the classifier succeeded in detecting 19 BI-RAD 6 images out of 40, which is True Positive, while misidentified only one image, called False Negative. Moreover, the model successfully classified 16 BI-RAD 4 images which are called True Negative, while misclassified 4, which is called False Positive. Additionally, the accuracy was equal to 0.88 which means that 88% of the output images were classified correctly. Likewise, the recall (TPR) was equal to 0.95. That is, 95% of BI-RAD 6 images were correctly classified as BI-RAD 6 by the model. In contrast, the FPR was 0.2, which means that there is a 20% possibility that the classifier will incorrectly identify the BI-RAD 4 class as a BI-RAD 6 class. In addition, precision resulted in 0.83, which is the rate of correctly predicted BI-RAD 6 class to all predicted BI-RAD 6. Finally, the F1-score was equal to 0.88, and the area under the curve (AUC) was also equal to 0.88.

In VGG16:

The four main values in the matrix were as follows: TP = 19, FN = 1, FP = 9, and TN = 11. That is, the model succeeded in classifying 19 BI-RAD 6 images out of 40, which is the True Positive, while misclassified 1 and classified them as BI-RAD 4, called False Negative. Similarly, the model successfully identified 11 BI-RAD 4 images which are called True Negative, while misclassified 9 and classified them as BI-RAD 6, which is called False Positive. Likewise, the accuracy was 0.70 which means that 70% of the output images were classified correctly. Furthermore, the recall (TPR) was equal to 0.95. That is, 95% of BI-RAD 6 images were correctly diagnosed as BI-RAD 6 by the classifier. In contrast, the FPR was 0.55, which means that there is a 55% chance that the model will incorrectly classify BI-RAD 4 as BI-RAD 6. In the same way, precision

resulted in 0.63, the F1-score was equal to 0.76, and the AUC was equal to accuracy with a value of 0.70.

In VGG19:

The four main values in the matrix were as follows: TP = 0, FN = 20, FP = 0, and TN = 20. That is, the model didn't succeed in classifying any image in the BI-RAD 6 class, which is True Positive, while misclassified all of the images in BI-RAD 6, called False Negative. Similarly, the classifier successfully identified 20 BI-RAD 4 images which are called True Negative, while misclassified none, which is called False Positive. Also, the accuracy was 0.5 which means that 50% of the outputs were classified correctly. Besides, the recall (TPR) was higher equal to zero. That is, 0% of BI-RAD 6 images were correctly diagnosed as BI-RAD 6 by the model. Comparatively, the FPR was also zero, which means that there is a 0% chance that the model will incorrectly diagnose BI-RAD 4 as BI-RAD 6. Also, the precision was zero, the F1-score was equal to zero, and the AUC was equal to 0.5.

In Xception:

The four main values in the matrix were as follows: TP = 16, FN = 4, FP = 12, and TN = 8. That is, the model succeeded in classifying 16 BI-RAD 6, which is True Positive, while misidentified 4 instances and classified them as BI-RAD 4, called False Negative. Moreover, the model successfully diagnosed 8 BI-RAD 4 images which is called True Negative, while misidentified 12. By the same token, the accuracy was 0.6 which means that 60% of the outcome images were diagnosed correctly. In addition, the recall (TPR) was equal to 0.8. That is, 80% of BI-RAD 6 images were correctly diagnosed as BI-

RAD 6 by the classifier. In contrast, the FPR was 0.6, the precision resulted in 0.57, and the F1-score was equal to 0.67. Finally, the area under the curve (AUC) was 0.6.

In EfficientNetB7:

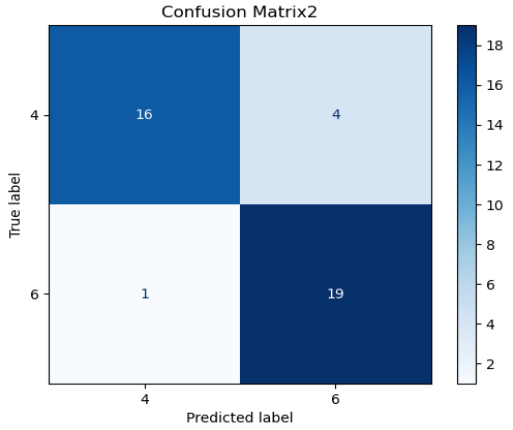
The four main values in the matrix were as follows: TP = 20, FN = 0, FP = 4, and TN = 16. That is, the model succeeded in classifying 20 BI-RAD 6 images out of 40, which is True Positive, while misclassified none and that is called False Negative. Also, the model successfully classified 16 BI-RAD 4 images which are called True Negative, while misidentified 4 images in BI-RAD 4, which is called False Positive. Moreover, the accuracy resulted in 0.9 which means that 90% of the predictions were identified correctly. Besides, the recall (TPR) was equal to 1. That is, 100% of BI-RAD 6 images were correctly classified as BI-RAD 6 by the model. In contrast, the FPR was 0.2, meaning that there is a 20% chance that the model will incorrectly recognize BI-RAD 4 as BI-RAD 6. Likewise, precision resulted in 0.83, the F1-score was equal to 0.91, and the AUC was equal to 0.90.

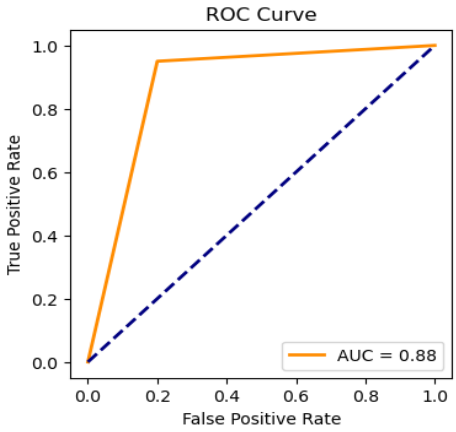
In ResNet50:

The four main values in the matrix were as follows: TP = 20, FN = 0, FP = 20, and TN = 0. That is, the model succeeded in classifying all of the BI-RAD 6 images, which is True Positive, while misidentified none, which is called False Negative. On the other hand, the classifier didn’t succeed in identifying any of the BI-RAD 4 images which are called True Negative, while misclassified all of them, which is called False Positive. Moreover, the accuracy was 0.5 meaning that only 50% of the predictions were identified correctly. Also, the recall (TPR) was equal to 1. That is, 100% of BI-RAD 6 images were correctly diagnosed as BI-RAD 6 by the classifier. Comparatively, the FPR was also equal to 1, meaning that there is a 100% percentage that the model will

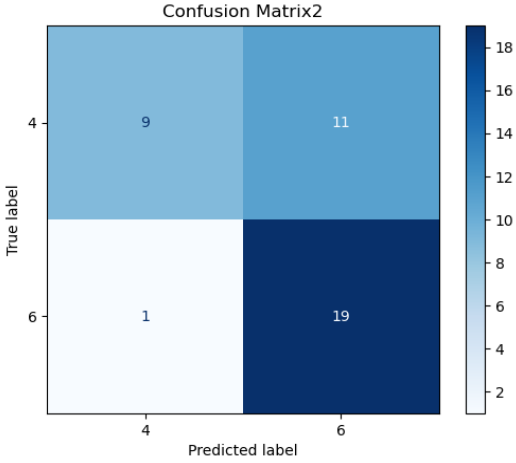
incorrectly recognize BI-RAD 4 as BI-RAD 6. In contrast, precision resulted in 0.5, the F1-score was also equal to 0.67, and the AUC was 0.5.

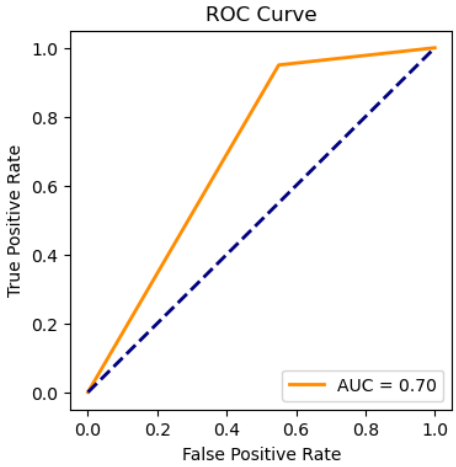
DenseNet121



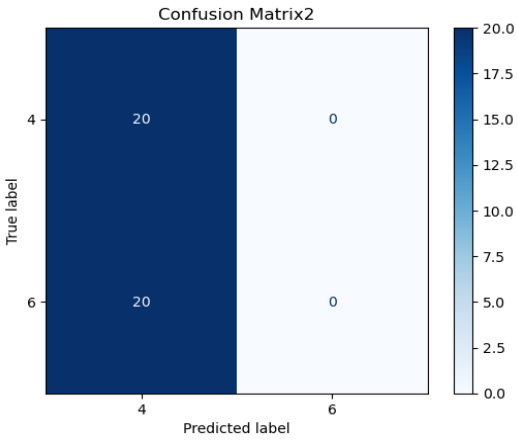


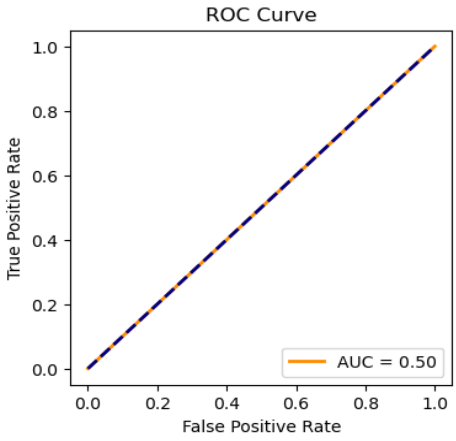
VGG16





VGG19





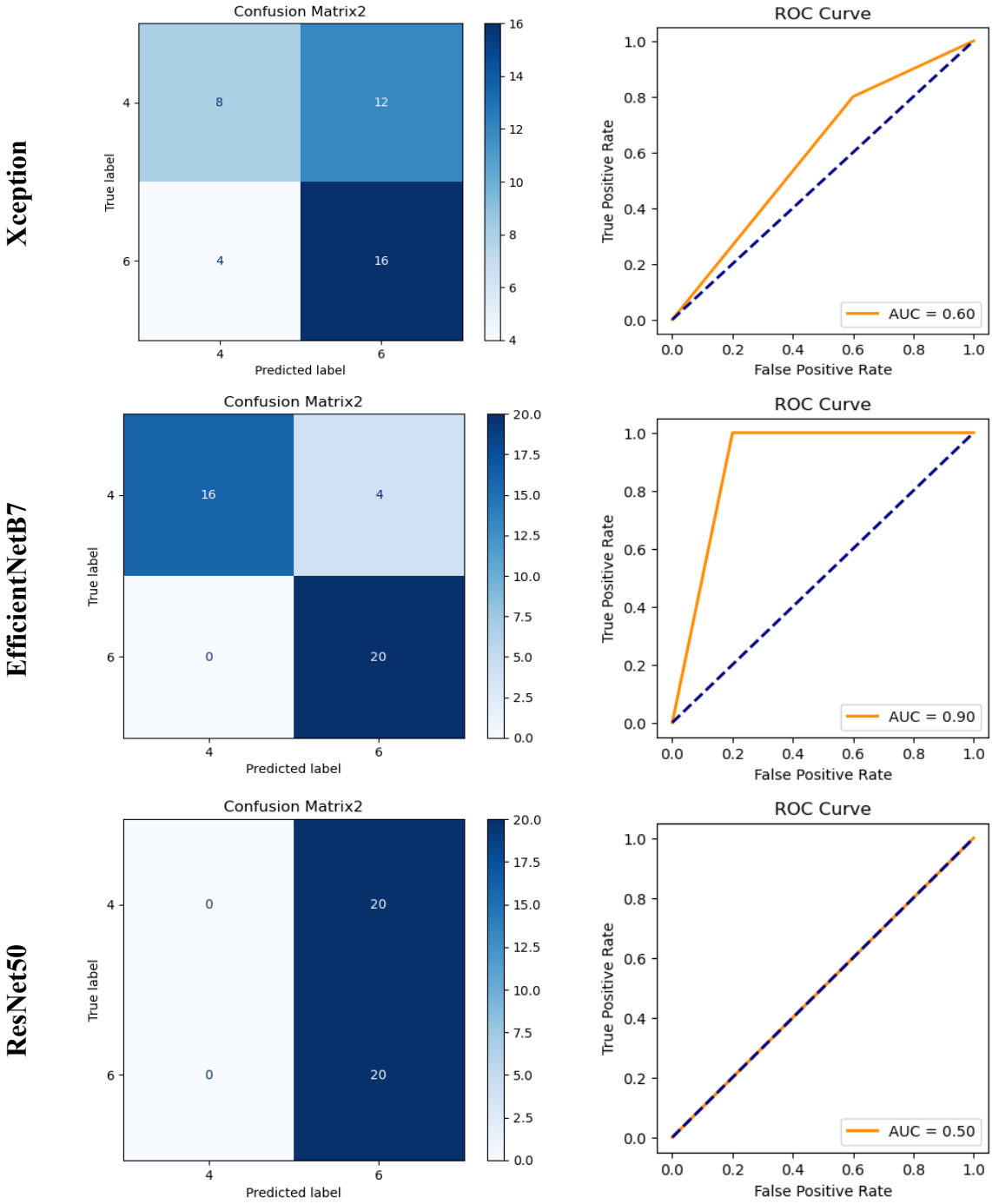


Figure 6.30: Confusion matrix and AUC in BI-RADS 4 and 6.

The summary of the previous results obtained from BI-RADS 4 and BI-RADS can be found in Table (6.7) below.

Table 6.7: Performance measures for all models in BI-RADS 4 and 6.

Model	Test Accuracy	Recall (TPR)	FPR	Precision	F1-score	ROC-AUC

DenseNet121	0.88	0.95	0.20	0.83	0.88	0.88
VGG16	0.70	0.95	0.55	0.63	0.76	0.70
VGG19	0.50	0.00	0.00	0.00	0.00	0.50
Xception	0.60	0.80	0.60	0.57	0.67	0.60
EfficientNetB7	0.90	1.00	0.20	0.83	0.91	0.90
ResNet50	0.50	1.00	1.00	0.50	0.67	0.50

6.3.7 Second Approach Results Summary

In summary, Table (6.8) lists all results obtained from each use case in the second approach regarding model accuracy. Overall, some models resulted in overfitting for some use cases such as EfficientNetB7 in use cases [4, 5], and [3, 5]. Moreover, DenseNcet121 showed overfitting in use cases [5, 6], [3,5], and [3,6]. In addition, results highlighted in yellow indicate overfitting, while green color indicates the best results in each use case. Xception and EfficientNetB7 obtained the best results and can be generalized to be used in other similar tasks.

Table 6.8: Second approach results summary for accuracy.

Model's Accuracy	DenseNet121	VGG16	VGG19	**Xception**	**EfficientNetB7**	ResNet50
[3, 4]	0.90	0.80	0.65	**0.97**	0.92	0.50
[4, 5]	0.90	0.85	0.70	**0.95**	1.00	0.55
[5, 6]	1.00	0.90	0.85	**0.95**	**0.97**	0.50
[3, 5]	1.00	0.75	0.72	**0.92**	1.00	0.50
[3, 6]	1.00	0.90	0.88	**0.95**	**0.97**	0.50
[4, 6]	0.88	0.70	0.50	0.60	**0.90**	0.50

Likewise, Table (6.9) below illustrates the comprehensive results of this approach. This table presents the confusion matrix, accuracy, recall, false positive rate, precision, f1-score, and area under the curve. Note that the bold numbers represent the best results.

Table 6.9: Comprehensive results in the second approach.

BI-RADS	Model	TN	FP	FN	TP	Test Accuracy	Recall (TPR)	FPR	Precision	F1-score	AUC
[3, 4]	DenseNet121	18	2	2	18	0.90	0.90	0.10	0.90	0.90	0.90
	VGG16	15	5	3	17	0.80	0.85	0.25	0.77	0.81	0.80
	VGG19	12	8	6	14	0.65	0.70	0.40	0.64	0.67	0.65

	Xception	20	0	1	19	**0.97**	**0.95**	**0.00**	**1.00**	**0.98**	**0.97**
	EffiicentNetB7	17	3	0	20	0.93	1.00	0.15	0.87	0.93	0.92
	ResNet50	0	20	0	20	0.50	1.00	1.00	0.50	0.67	0.50
[4, 5]	DenseNet121	16	4	0	20	0.90	1.00	0.20	0.83	0.91	0.90
	VGG16	19	1	5	15	0.85	0.75	0.05	0.94	0.83	0.85
	VGG19	17	3	9	11	0.70	0.55	0.15	0.79	0.65	0.70
	Xception	20	0	2	18	**0.95**	**0.90**	**0.00**	**1.00**	**0.95**	**0.95**
	EffiicentNetB7	20	0	0	20	1.00	1.00	0.00	1.00	1.00	1.00
	ResNet50	2	18	0	20	0.55	1.00	0.90	0.53	0.69	0.55
[5, 6]	DenseNet121	20	0	0	20	1.00	1.00	0.00	1.00	1.00	1.00
	VGG16	18	2	2	18	0.90	0.90	0.10	0.90	0.90	0.90
	VGG19	18	2	4	16	0.85	0.80	0.10	0.89	0.84	0.85
	Xception	20	0	2	18	**0.95**	**0.90**	**0.00**	**1.00**	**0.95**	**0.95**
	EffiicentNetB7	19	1	0	20	**0.97**	**1.00**	**0.05**	**0.95**	**0.98**	**0.97**
	ResNet50	20	0	20	0	1.00	1.00	0.00	1.00	1.00	1.00
[3, 5]	DenseNet121	20	0	0	20	1.00	1.00	0.00	1.00	1.00	1.00
	VGG16	15	5	5	15	0.75	0.75	0.25	0.75	0.75	0.75
	VGG19	14	6	5	15	0.73	0.75	0.30	0.71	0.73	0.72
	Xception	19	1	2	18	**0.93**	**0.90**	**0.05**	**0.95**	**0.92**	**0.92**
	EffiicentNetB7	20	0	0	20	1.00	1.00	0.00	1.00	1.00	1.00
	ResNet50	0	20	0	20	0.50	1.00	1.00	0.50	0.67	0.50
[3, 6]	DenseNet121	20	0	0	20	1.00	1.00	0.00	1.00	1.00	1.00
	VGG16	18	2	2	18	0.90	0.90	0.01	0.90	0.90	0.90
	VGG19	16	4	1	19	0.88	0.95	0.20	0.83	0.88	0.88
	Xception	20	0	2	18	**0.95**	**0.00**	**0.00**	**1.00**	**0.95**	**0.95**
	EffiicentNetB7	19	1	0	20	**0.97**	**1.00**	**0.05**	**0.95**	**0.98**	**0.97**
	ResNet50	0	20	0	20	0.50	1.00	1.00	0.50	0.67	0.50
[4, 6]	DenseNet121	16	4	1	19	0.88	0.95	0.20	0.83	0.88	0.88
	VGG16	9	11	1	19	0.70	0.95	0.55	0.63	0.76	0.70
	VGG19	20	0	20	0	0.50	0.00	0.00	0.00	0.00	0.50
	Xception	8	12	4	16	0.60	0.80	0.60	0.57	0.67	0.60
	EffiicentNetB7	16	4	0	20	**0.90**	**1.00**	**0.20**	**0.83**	**0.91**	**0.90**
	ResNet50	0	20	0	20	0.50	1.00	1.00	0.50	0.67	0.50

6.4 Summary

In conclusion, the evaluation process is a very important stage in the model lifecycle. Moreover, a model's evaluation can tell us whether this model is good at fitting the data or not. Poor-fit models should be improved and good-fit models should be generalized. In this chapter, two approaches were implemented. The first model classified the data

into normal and abnormal, while the second approach classified the data into BI-RADS. In addition, six convolutional neural network models were used, namely DenseNet121, ResNet50, VGG16, VGG19, EfficientNetB7, and Xception. Furthermore, DenseNet121 obtained the best result in terms of accuracy with a value of 0.83 in the first approach to classify images into normal and abnormal. On the other hand, Xception and EfficientNetB7 had the best results in classifying images into BI-RADS. Xception accuracy had values of 0.97, 0.95, 0.95, 0.92, and 0.95 for the BI-RAD groups [3, 4], [4, 5], [5, 6], [3, 5], and [3,6] respectively. In contact, EfficientNetB7 resulted in 0.97, 0.97, and 0.90 for the groups [5, 6], [3, 6], and [4, 6] in order. These models can be generalized to be utilized in other tasks specifically, for detecting cancer diseases. Next chapter will present the future work and other enhancements.

Chapter Seven: Conclusion and Future Work

Breast cancer is the top one common cancer in women globally and ranks first as a cause of death in cancer deaths. In addition, many risk factors can help in developing breast cancer such as multiple births, a short breastfeeding period, a family history of cancers, being overweight, having a low socioeconomic level, a low level of education, genetics, and lifestyle. Furthermore, mammogram screening is the most common and efficient method to diagnose abnormalities in the breast. Early detection is very important to control cancer and improve the outcome because at this stage the treatment is cheap and effective. In contrast, machine learning plays a significant role in detecting breast cancer. Many studies proposed different models to diagnose breast cancer with promising results.

This study implements six pre-trained CNN models that obtained the best results in the state-of-the-art on a first-hand dataset collected from a MoH mammogram screening center in Palestine. The size of the dataset was similar to the ones that were used in the literature. The models are: DenseNet121, VGG16, VGG19, ResNet50, Xception, and EfficientNetB7. Additionally, two approaches were presented in this thesis where the first approach classified the images into normal and abnormal and the second classified the images into BI-RADS.

In this thesis, many performance measures like confusion matrix, accuracy, recall, precision, f1-score, and ROC curve were used to check the reliability and efficiency of such models. Based on the results obtained from the first approach, DenseNet121 achieved the highest testing accuracy of 0.83, and an AUC of 0.85, and outperformed all other models. EfficientNetB7 showed close results with an accuracy of 0.81. Overall, the performance in these models was better in detecting the abnormal class than the

normal class. Comparatively, in the second approach, Xception obtained the highest results in most cases. EfficientNetB7 caused overfitting in some cases but got high results in most cases in classifying images into BI-RADS. The Xception model accuracy values were 0.97, 0.95, 0.95, 0.92, and 0.95 in BI-RADs groups [3, 4], [4, 5], [5, 6], [3, 5], and [3,6] respectively. On the other hand, EfficientNetB7 accuracy values were 0.97, 0.97, and 0.90 in BI-RADS [5, 6], [3, 6], and [4, 6] in order. ResNet50 had the lowest results in both approaches and acted like a random classifier in all cases. Lastly, these models can be generalized to other tasks in image processing in the healthcare field.

In future work, a larger dataset of mammograms should be considered and a repository should be created for these kinds of datasets. Additionally, utilizing more automated techniques for image pre-processing such as Morphological transformations provided by the OpenCV library. Furthermore, these models can be used for diseases that are screened and diagnosed using medical images such as X-rays, ultrasound, or CT scans. Similarly, these models should be integrated with other patient data such as genetic information, lifestyle factors, and medical history, to assess an individual's risk of developing certain diseases. This enables proactive measures, such as preventive screenings or lifestyle modifications, to be taken for individuals at higher risk.

References

Aboutalib, S. S., Mohamed, A. A., Berg, W. A., Zuley, M. L., Sumkin, J. H., & Wu, S. (2018). Deep Learning to Distinguish Recalled but Benign Mammography Images in Breast Cancer Screening. *Clinical Cancer Research*, *24*(23), 5902–5909. https://doi.org/10.1158/1078-0432.CCR-18-1115

Abukmeil, M., Ferrari, S., Genovese, A., Piuri, V., & Scotti, F. (2020). Unsupervised Learning From Limited Available Data by β-NMF and Dual Autoencoder. *2020 IEEE International Conference on Image Processing (ICIP)*, 81–85. https://doi.org/10.1109/ICIP40778.2020.9191252

Adebiyi, M. O., Arowolo, M. O., Mshelia, M. D., & Olugbara, O. O. (2022). A Linear Discriminant Analysis and Classification Model for Breast Cancer Diagnosis. *Applied Sciences*, *12*(22), 11455. https://doi.org/10.3390/APP122211455

Adedigba, A. P., Adeshina, S. A., & Aibinu, A. M. (2022). Performance Evaluation of Deep Learning Models on Mammogram Classification Using Small Dataset. *Bioengineering*, *9*(4), 161. https://doi.org/10.3390/bioengineering9040161

Al Qadire, M., Alsaraireh, M., Alomari, K., Aldiabat, K. M., Al-Sabei, S., Al-Rawajfah, O., & Aljezawi, M. (2021). Symptom Clusters Predictive of Quality of Life Among Jordanian Women with Breast Cancer. Seminars in Oncology Nursing, 37(2), 151144. https://doi.org/https://doi.org/10.1016/j.soncn.2021.151144

Alfian, G., Syafrudin, M., Fahrurrozi, I., Fitriyani, N. L., Atmaji, F. T. D., Widodo, T., Bahiyah, N., Benes, F., & Rhee, J. (2022). Predicting Breast Cancer from Risk Factors Using SVM and Extra-Trees-Based Feature Selection Method. Computers, 11(9), 136. https://doi.org/10.3390/COMPUTERS11090136

AlWaheidi, S., McPherson, K., Chalmers, I., Sullivan, R., & Davies, E. A. (2020). Mammographic Screening in the Occupied Palestinian Territory: A Critical Analysis of Its Promotion, Claimed Benefits, and Safety in Palestinian Health Research. JCO Global Oncology, 6, 1772–1790. https://doi.org/10.1200/JGO.19.00383

American Cancer Society. (2022). What Does the Doctor Look for on a Mammogram? American Cancer Society.

Apaydin, H., Feizi, H., Sattari, M. T., Colak, M. S., Shamshirband, S., & Chau, K.-W. (2020). Comparative Analysis of Recurrent Neural Network Architectures for Reservoir Inflow Forecasting. *Water*, *12*(5), 1500. https://doi.org/10.3390/w12051500

Ayad, A., & Abdulmunim, M. E. (2023). Detecting Abnormal Driving Behavior Using Modified DenseNet. *Iraqi Journal for Computer Science and Mathematics*, *4*(3), 48–65. https://doi.org/10.52866/IJCSM.2023.02.03.005

Bagchi, S., Tay, K. G., Huong, A., & Debnath, S. K. (2020). Image processing and machine learning techniques used in computer-aided detection system for mammogram screening - a review. *International Journal of Electrical and Computer Engineering (IJECE)*, *10*(3), 2336–2348. https://doi.org/10.11591/ijece.v10i3.pp2336-2348

Baheti, B., Innani, S., Gajre, S., & Talbar, S. (2020). Eff-UNet: A Novel Architecture for Semantic Segmentation in Unstructured Environment. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 1473–1481. https://doi.org/10.1109/CVPRW50498.2020.00187

Banerjee, I., Bozkurt, S., Alkim, E., Sagreiya, H., Kurian, A. W., & Rubin, D. L. (2019). Automatic inference of BI-RADS final assessment categories from narrative mammography report findings. *Journal of Biomedical Informatics*, *92*, 103137. https://doi.org/10.1016/j.jbi.2019.103137

Batchu, S., Liu, F., Amireh, A., Waller, J., & Umair, M. (2021). A Review of Applications of Machine Learning in Mammography and Future Challenges. *Oncology*, *99*(8), 483–490. https://doi.org/10.1159/000515698

Benning, L., Peintner, A., & Peintner, L. (2022). Advances in and the Applicability of Machine Learning-Based Screening and Early Detection Approaches for Cancer: A Primer. *Cancers*, *14*(3), 623. https://doi.org/10.3390/cancers14030623

Bi, W. L., Hosny, A., Schabath, M. B., Giger, M. L., Birkbak, N. J., Mehrtash, A., Allison, T., Arnaout, O., Abbosh, C., Dunn, I. F., Mak, R. H., Tamimi, R. M., Tempany, C. M., Swanton, C., Hoffmann, U., Schwartz, L. H., Gillies, R. J., Huang, R. Y., & Aerts, H. J. W. L. (2019).

Artificial intelligence in cancer imaging: Clinical challenges and applications. *CA: A Cancer Journal for Clinicians*, *69*(2), 127–157. https://doi.org/10.3322/caac.21552

Biswas, A., Saran, I., & Wilson, F. P. (2021). Introduction to Supervised Machine Learning. *Kidney360*, *2*(5), 878–880. https://doi.org/10.34067/KID.0000182021

Bitencourt, A. G., Ferreira, E. V., Bastos, D. C., Sperandio, V. A., Graziano, L., Guatelli, C. S., Albuquerque, M. L., Souza, J. A., & Marques, E. F. (2019). Intramammary lymph nodes: normal and abnormal multimodality imaging features. *The British Journal of Radiology*, *92*(1103), 20190517. https://doi.org/10.1259/bjr.20190517

Bodepudi, H. (2021). Credit Card Fraud Detection Using Unsupervised Machine Learning Algorithms. *International Journal of Computer Trends and Technology*, *69*(8), 1–3. https://doi.org/10.14445/22312803/IJCTT-V69I8P101

Bokhare, A., & Jha, P. (2023). Machine learning models applied in analyzing breast cancer classification accuracy. *IAES International Journal of Artificial Intelligence*, *12*(3), 1370–1377. https://doi.org/10.11591/IJAI.V12.I3.PP1370-1377

Bommert, A., Sun, X., Bischl, B., Rahnenführer, J., & Lang, M. (2020). Benchmark for filter methods for feature selection in high-dimensional classification data. *Computational Statistics & Data Analysis*, *143*, 106839. https://doi.org/10.1016/j.csda.2019.106839

Bommert, A., Welchowski, T., Schmid, M., & Rahnenführer, J. (2022). Benchmark of filter methods for feature selection in high-dimensional gene expression survival data. *Briefings in Bioinformatics*, *23*(1). https://doi.org/10.1093/bib/bbab354

Castro-Tapia, S., Castañeda-Miranda, C. L., Olvera-Olvera, C. A., Guerrero-Osuna, H. A., Ortiz-Rodriguez, J. M., Martínez-Blanco, Ma. del R., Díaz-Florez, G., Mendiola-Santibañez, J. D., & Solís-Sánchez, L. O. (2021). Classification of Breast Cancer in Mammograms with Deep Learning Adding a Fifth Class. *Applied Sciences*, *11*(23), 11398. https://doi.org/10.3390/app112311398

Çayir, A., Yenidoğan, I., & Dağ, H. (2018). Feature Extraction Based on Deep Learning for Some Traditional Machine Learning Methods. *2018 3rd International Conference on*

Computer Science and Engineering (UBMK), 494–497. https://doi.org/10.1109/UBMK.2018.8566383

Cechanoviciute, V., & Cechanovičienė, I. (2021). Overview of the epidemiology, risk factors, clinical features, diagnostics and prevention of breast cancer. *Journal of Medical Sciences*, *10*(1), 71–79.

Ceylan, Z. (2020). Diagnosis of Breast Cancer Using Improved Machine Learning Algorithms Based on Bayesian Optimization. *International Journal of Intelligent Systems and Applications in Engineering*, *8*(3), 121–130. https://doi.org/10.18201/ijisae.2020363531

Chang, Y.-W., Chen, Y.-R., Ko, C.-C., Lin, W.-Y., & Lin, K.-P. (2020). A Novel Computer-Aided-Diagnosis System for Breast Ultrasound Images Based on BI-RADS Categories. *Applied Sciences*, *10*(5), 1830. https://doi.org/10.3390/app10051830

Charan, S., Khan, M. J., & Khurshid, K. (2018). Breast cancer detection in mammograms using convolutional neural network. *2018 International Conference on Computing, Mathematics and Engineering Technologies (ICoMET)*, 1–5. https://doi.org/10.1109/ICOMET.2018.8346384

Chen, C.-W., Tsai, Y.-H., Chang, F.-R., & Lin, W.-C. (2020). Ensemble feature selection in medical datasets: Combining filter, wrapper, and embedded feature selection results. *Expert Systems*, *37*(5), e12553. https://doi.org/https://doi.org/10.1111/exsy.12553

Cherrington, M., Thabtah, F., Lu, J., & Xu, Q. (2019). Feature Selection: Filter Methods Performance Challenges. *2019 International Conference on Computer and Information Sciences (ICCIS)*, 1–4. https://doi.org/10.1109/ICCISci.2019.8716478

Chollet, F. (2017). Xception: Deep Learning with Depthwise Separable Convolutions. *IEEE Conference on Computer Vision and Pattern Recognition*, 1251–1258. https://doi.org/10.48550/arXiv.1610.02357

Chougrad, H., Zouaki, H., & Alheyane, O. (2018). Deep Convolutional Neural Networks for breast cancer screening. *Computer Methods and Programs in Biomedicine*, *157*, 19–30. https://doi.org/10.1016/j.cmpb.2018.01.011

Ciuba, A., Wnuk, K., Nitsch-Osuch, A., & Kulpa, M. (2022). Health Care Accessibility and Breast Cancer Mortality in Europe. *International Journal of Environmental Research and Public Health*, *19*(20), 13605. https://doi.org/10.3390/ijerph192013605

Clunie, D. A. (2021). DICOM Format and Protocol Standardization—A Core Requirement for Digital Pathology Success. *Toxicologic Pathology*, *49*(4), 738–749. https://doi.org/10.1177/0192623320965893

Das, H., Naik, B., & Behera, H. S. (2022). A Jaya algorithm based wrapper method for optimal feature selection in supervised classification. *Journal of King Saud University - Computer and Information Sciences*, *34*(6), 3851–3863. https://doi.org/10.1016/j.jksuci.2020.05.002

Das, P., & Mazumder, D. H. (2023). An extensive survey on the use of supervised machine learning techniques in the past two decades for prediction of drug side effects. *Artificial Intelligence Review*, *56*(9), 9809–9836. https://doi.org/10.1007/s10462-023-10413-7

Deb, S., Rahman, A., & Jha, R. K. (2023). Breast cancer diagnosis using modified Xception and stacked generalization ensemble classifier. *Research on Biomedical Engineering*, *39*(23), 937–947. https://doi.org/10.1007/s42600-023-00317-4

Dhahri, H., Al Maghayreh, E., Mahmood, A., Elkilani, W., & Faisal Nagi, M. (2019). Automated Breast Cancer Diagnosis Based on Machine Learning Algorithms. *Journal of Healthcare Engineering*, *2019*(4253641), 1–11. https://doi.org/10.1155/2019/4253641

Djunaidi, K., Bedi Agtriadi, H., Kuswardani, D., & S. Purwanto, Y. (2021). Gray level co-occurrence matrix feature extraction and histogram in breast cancer classification with ultrasonographic imagery. *Indonesian Journal of Electrical Engineering and Computer Science*, *22*(2), 795–800. https://doi.org/10.11591/ijeecs.v22.i2.pp795-800

Dritsas, E., & Trigka, M. (2022). Lung Cancer Risk Prediction with Machine Learning Models. *Big Data and Cognitive Computing*, *6*(4), 139. https://doi.org/10.3390/bdcc6040139

Du, X., Si, L., Li, P., & Yun, Z. (2023). A method for detecting the quality of cotton seeds based on an improved ResNet50 model. *PLOS ONE*, *18*(2), 1–19. https://doi.org/10.1371/journal.pone.0273057

Eberl, M. M., Fox, C. H., Edge, S. B., Carter, C. A., & Mahoney, M. C. (2006). BI-RADS Classification for Management of Abnormal Mammograms. *The Journal of the American Board of Family Medicine*, *19*(2), 161–164. https://doi.org/10.3122/jabfm.19.2.161

Ebrahim, M., Sedky, A., & Mesbah, S. (2023). Accuracy Assessment of Machine Learning Algorithms Used to Predict Breast Cancer. *Data*, *8*(2), 35. https://doi.org/10.3390/DATA8020035

Eghtedari, M., Chong, A., Rakow-Penner, R., & Ojeda-Fournier, H. (2021). Current Status and Future of BI-RADS in Multimodality Imaging, From the AJR Special Series on Radiology Reporting and Data Systems. *American Journal of Roentgenology*, *216*(4), 860–873. https://doi.org/10.2214/AJR.20.24894

Epperlein, J. P., Overko, R., Zhuk, S., King, C., Bouneffouf, D., Cullen, A., & Shorten, R. (2022). Reinforcement learning with algorithms from probabilistic structure estimation. *Automatica*, *144*(1), 110483. https://doi.org/10.1016/j.automatica.2022.110483

Es-sabery, F., Hair, A., Qadir, J., de Abajo, B., Garcia-Zapirain, B., & la Torre Díez, I. (2021). Sentence-Level Classification Using Parallel Fuzzy Deep Learning Classifier. *IEEE Access*, *PP*(99), 1. https://doi.org/10.1109/ACCESS.2021.3053917

Falconí, L., Pérez, M., Aguilar, W., & Conci, A. (2020). Transfer Learning and Fine Tuning in Mammogram BI-RADS Classification. *2020 IEEE 33rd International Symposium on Computer-Based Medical Systems (CBMS)*, 475–480. https://doi.org/10.1109/CBMS49503.2020.00096

Fanizzi, A., Basile, T. M., Losurdo, L., Bellotti, R., Bottigli, U., Campobasso, F., Didonna, V., Fausto, A., Massafra, R., Tagliafico, A., Tamborra, P., Tangaro, S., Lorusso, V., & La Forgia, D. (2019). Ensemble Discrete Wavelet Transform and Gray-Level Co-Occurrence Matrix for Microcalcification Cluster Classification in Digital Mammography. *Applied Sciences*, *9*(24), 5388. https://doi.org/10.3390/app9245388

Geras, K. J., Mann, R. M., & Moy, L. (2019). Artificial Intelligence for Mammography and Digital Breast Tomosynthesis: Current Concepts and Future Perspectives. *Radiology*, *293*(2), 246–259. https://doi.org/10.1148/radiol.2019182627

Ghosh, M., Guha, R., Sarkar, R., & Abraham, A. (2020). A wrapper-filter feature selection technique based on ant colony optimization. *Neural Computing and Applications*, *32*(12), 7839–7857. https://doi.org/10.1007/s00521-019-04171-3

Giaquinto, A. N., Sung, H., Miller, K. D., Kramer, J. L., Newman, L. A., Minihan, A., Jemal, A., & Siegel, R. L. (2022). Breast Cancer Statistics, 2022. *CA: A Cancer Journal for Clinicians*, *72*(6), 524–541. https://doi.org/https://doi.org/10.3322/caac.21754

Gilroy, A. (2021). *Anatomy An Essential Textbook* (3rd ed.). Thieme.

Ginsburg, O., Yip, C.-H., Brooks, A., Cabanes, A., Caleffi, M., Dunstan Yataco, J. A., Gyawali, B., McCormack, V., de Anderson, M., Mehrotra, R., Mohar, A., Murillo, R., Pace, L. E., Paskett, E. D., Romanoff, A., Rositch, A. F., Scheel, J. R., Schneidman, M., Unger-Saldaña, K., … Anderson, B. O. (2020). Breast cancer early detection: A phased approach to implementation. *Cancer*, *126*(S10), 2379–2393. https://doi.org/https://doi.org/10.1002/cncr.32887

Girik Allo, C. B., Putra, L. S. A., Paranoan, N. R., & Gunawan, V. A. (2023). Comparing Logistic Regression and Support Vector Machine in Breast Cancer Problem. *Jambura Journal of Probability and Statistics*, *4*(1), 1–8. https://doi.org/10.34312/JJPS.V4I1.19246

Gnanasekaran, V. S., Joypaul, S., Meenakshi Sundaram, P., & Chairman, D. D. (2020). Deep learning algorithm for breast masses classification in mammograms. *IET Image Processing*, *14*(12), 2860–2868. https://doi.org/10.1049/iet-ipr.2020.0070

González, J., Ortega, J., Damas, M., Martín-Smith, P., & Gan, J. Q. (2019). A new multi-objective wrapper method for feature selection – Accuracy and stability analysis for BCI. *Neurocomputing*, *333*, 407–418. https://doi.org/10.1016/j.neucom.2019.01.017

He, K., Zhang, X., Ren, S., & Sun, J. (2015). Deep Residual Learning for Image Recognition. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 770–778. https://api.semanticscholar.org/CorpusID:206594692

Hindarto, D., Afarini, N., & Esthi, E. (2023). Comparison Efficacy of VGG16 and VGG19 Insect Classification Models. *JIKO (Jurnal Informatika Dan Komputer)*, *6*(3), 189–195. https://doi.org/10.33387/jiko.v6i3.7008

Hoque, M. E., Kipli, K., Zulcaffle, T. M. A., Al-Hababi, A. Y. S., Awang Mat, D. A., Sapawi, R., & Joseph, A. A. (2021). A Deep Learning Approach for Retinal Image Feature Extraction. *Pertanika Journal of Science and Technology*, *29*(4). https://doi.org/10.47836/pjst.29.4.17

Hossin, M., & M.N, S. (2015). A Review on Evaluation Metrics for Data Classification Evaluations. *International Journal of Data Mining & Knowledge Management Process*, *5*(2), 1–11. https://doi.org/10.5121/ijdkp.2015.5201

Hossin, M., Shamrat, F., Bhuiyan, M. R., Hira, R., Khan, T., & Molla, S. (2023). Breast cancer detection: an effective comparison of different machine learning algorithms on the Wisconsin dataset. *Bulletin of Electrical Engineering and Informatics*, *12*, 2446–2456. https://doi.org/10.11591/beei.v12i4.4448

Huang, G., Liu, Z., van der Maaten, L., & Weinberger, K. Q. (2017). Densely Connected Convolutional Networks. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 4700–4708. https://doi.org/10.48550/arXiv.1608.06993

Ibeni, W. N. L. W. H., Salikon, M. Z. M., Mustapha, A., Daud, S. A., & Salleh, M. N. M. (2019). Comparative analysis on bayesian classification for breast cancer problem. *Bulletin of Electrical Engineering and Informatics*, *8*(4), 1303–1311. https://doi.org/10.11591/eei.v8i4.1628

Ibrokhimov, B., & Kang, J.-Y. (2022). Two-Stage Deep Learning Method for Breast Cancer Detection Using High-Resolution Mammogram Images. *Applied Sciences*, *12*(9), 4616. https://doi.org/10.3390/app12094616

İz, F. B., & Tümer, A. (2016). Assessment of Breast Cancer Risk and Belief in Breast Cancer Screening Among the Primary Healthcare Nurses. *Journal of Cancer Education*, *31*(3), 575–581. https://doi.org/10.1007/s13187-015-0977-y

Jain, P. (2022). Detection of Breast Cancer Using Machine Learning Algorithms. *International Journal for Research in Applied Science and Engineering Technology*, *10*(6), 3484–3487. https://doi.org/10.22214/ijraset.2022.44676

Jones, A. M., & Headley, W. C. (2022). Considerations of Reinforcement Learning within Real-Time Wireless Communication Systems. *MILCOM 2022 - 2022 IEEE Military Communications Conference (MILCOM)*, 418–425. https://doi.org/10.1109/MILCOM55135.2022.10017303

Jovic, A., Brkic, K., & Bogunovic, N. (2015). A review of feature selection methods with applications. *2015 38th International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO)*, 1200–1205. https://doi.org/10.1109/MIPRO.2015.7160458

Kamboj Akanshaand, Tanay Prashmitand, Sinha Akashand, & Kumar Prabhat. (2021). Breast Cancer Detection Using Supervised Machine Learning: A Comparative Analysis. In V. E. and M. A. and H. C. de A. V. and G. A. K. Goyal Dinesh and Bălaş (Ed.), *Information Management and Machine Intelligence* (pp. 263–269). Springer Singapore. https://doi.org/10.1007/978-981-15-4936-6_29

Kamil, M. Y., & Jassam, A.-L. A. (2020). Analysis of Tissue Abnormality in Mammography Images Using Gray Level Co-occurrence Matrix Method. *Journal of Physics: Conference Series*, *1530*(1), 12101. https://doi.org/10.1088/1742-6596/1530/1/012101

Karadag Arli, S., Bakan, A. B., & Aslan, G. (2019). Distribution of cervical and breast cancer risk factors in women and their screening behaviours. *European Journal of Cancer Care*, *28*(2), e12960. https://doi.org/https://doi.org/10.1111/ecc.12960

Khan, R., Usman, M., & Moinuddin, M. (2024). From Raw Data to Actionable Insights: Navigating the World of Data Analytics. *International Journal of Advanced Engineering Technologies and Innovations*, *1*(4), 142–166.

Khan, T. M., Xu, S., Khan, Z. G., & Uzair chishti, M. (2021). Implementing Multilabeling, ADASYN, and ReliefF Techniques for Classification of Breast Cancer Diagnostic through Machine Learning: Efficient Computer-Aided Diagnostic System. *Journal of Healthcare Engineering*, *2021*, 1–15. https://doi.org/10.1155/2021/5577636

Kopitar, L., Kocbek, P., Cilar, L., Sheikh, A., & Stiglic, G. (2020). Early detection of type 2 diabetes mellitus using machine learning-based prediction models. *Scientific Reports*, *10*(1), 11981. https://doi.org/10.1038/s41598-020-68771-z

Kulakli, A., & Osmanaj, V. (2020). Global Research on Big Data in Relation with Artificial Intelligence (A Bibliometric Study: 2008-2019). *International Journal of Online and Biomedical Engineering (IJOE)*, *16*(2), 31. https://doi.org/10.3991/ijoe.v16i02.12617

Kumar Bhowmik, T. (2015). Naive Bayes vs Logistic Regression: Theory, Implementation and Experimental Validation. *Inteligencia Artificial*, *18*(56), 14–30. https://doi.org/10.4114/INTARTIF.VOL18ISS56PP14-30

Kumari, P., Kumar, S., Shukla, B., & Dubey, A. (2022). An overview on breast cancer. *International Journal of Medical and All Body Health Research*, *2*(3), 59–65.

Kurian, B., & Jyothi, V. L. (2022). Comparative Analysis of Machine Learning Methods for Breast Cancer Classification in Genetic Sequences. *Journal of Environmental and Public Health*, *2022*, 1–6. https://doi.org/10.1155/2022/7199290

Kwekkeboom, K. L. (2016). Cancer Symptom Cluster Management. *Seminars in Oncology Nursing*, *32*(4), 373–382. https://doi.org/https://doi.org/10.1016/j.soncn.2016.08.004

Lehman, C. D., Yala, A., Schuster, T., Dontchos, B., Bahl, M., Swanson, K., & Barzilay, R. (2019). Mammographic Breast Density Assessment Using Deep Learning: Clinical Implementation. *Radiology*, *290*(1), 52–58. https://doi.org/10.1148/radiol.2018180694

Li, W., Yu, S., Yang, R., Tian, Y., Zhu, T., Liu, H., Jiao, D., Zhang, F., Liu, X., Tao, L., Gao, Y., Li, Q., Zhang, J., & Guo, X. (2023). Machine Learning Model of ResNet50-Ensemble Voting for Malignant–Benign Small Pulmonary Nodule Classification on

Computed Tomography Images. *Cancers*, *15*(22). https://doi.org/10.3390/cancers15225417

Liu, H., Zhou, M., & Liu, Q. (2019). An embedded feature selection method for imbalanced data classification. *IEEE/CAA Journal of Automatica Sinica*, *6*(3), 703–715. https://doi.org/10.1109/JAS.2019.1911447

Lu, M. (2019). Embedded feature selection accounting for unknown data heterogeneity. *Expert Systems with Applications*, *119*, 350–361. https://doi.org/10.1016/j.eswa.2018.11.006

Maanav, A., Mithun, K., Naparajith, T. L., Suraj, K. M. A., Bose, R., & Brearley, B. J. (2023). Investigating the Prediction of Breast Cancer Diagnosis by Use of Support Vector Machines. *International Journal of Healthcare Information Systems and Informatics*, *18*(1). https://doi.org/10.4018/IJHISI.325219

Mafarja, M., Heidari, A. A., Faris, H., Mirjalili, S., & Aljarah, I. (2020). Dragonfly Algorithm: Theory, Literature Review, and Application in Feature Selection. In S. Mirjalili, J. Song Dong, & A. Lewis (Eds.), *Nature-Inspired Optimizers: Theories, Literature Reviews and Applications* (pp. 47–67). Springer International Publishing. https://doi.org/10.1007/978-3-030-12127-3_4

Mahmood, A. R., Korenkevych, D., Vasan, G., Ma, W., & Bergstra, J. (2018). Benchmarking Reinforcement Learning Algorithms on Real-World Robots. *CoRR*, *abs/1809.07731*. https://doi.org/10.48550/arXiv.1809.07731

Mahmood, T., Li, J., Pei, Y., & Akhtar, F. (2021). An Automated In-Depth Feature Learning Algorithm for Breast Abnormality Prognosis and Robust Characterization from Mammography Images Using Deep Transfer Learning. *Biology*, *10*(9). https://doi.org/10.3390/biology10090859

Mahmood, T., Li, J., Pei, Y., Akhtar, F., Ur Rehman, M., & Wasti, S. (2022). Breast lesions classifications of mammographic images using a deep convolutional neural network-based approach. *PLOS ONE*, *17*(1), 1–25. https://doi.org/10.1371/journal.pone.0263126

Manjunathan, N., Gomathi, N., & Muthulingam, S. (2023). Early Detection of Breast Cancer using Machine Learning. *2023 International Conference on Sustainable Computing and Smart Systems (ICSCSS)*, 165–169. https://doi.org/10.1109/ICSCSS57650.2023.10169777

Mantri, M., Taran, S., & Sunder, G. (2022). DICOM Integration Libraries for Medical Image Interoperability: A Technical Review. *IEEE Reviews in Biomedical Engineering*, *15*, 247–259. https://doi.org/10.1109/RBME.2020.3042642

Martín-Payo, R., Martínez-Urquijo, A., Zabaleta-del-Olmo, E., & del Mar Fernandez-Alvarez, M. (2023). Use a web-app to improve breast cancer risk factors and symptoms knowledge and adherence to healthy diet and physical activity in women without breast cancer diagnosis (Precam project). *Cancer Causes & Control*, *34*(2), 113–122. https://doi.org/10.1007/s10552-022-01647-x

Mayo Clinic. (2023). *Breast calcifications*. Mayo Clinic.

Medeiros, A., Ohata, E. F., Silva, F. H. S., Rego, P. A. L., & Reboucas Filho, P. P. (2020). An Approach to BI-RADS Uncertainty Levels Classification Via Deep Learning with Transfer Learning Technique. *2020 IEEE 33rd International Symposium on Computer-Based Medical Systems (CBMS)*, 603–608. https://doi.org/10.1109/CBMS49503.2020.00119

Ministry of Health. (2022). *Health Annual Report Palestine*.

Mohamed, A. A., Berg, W. A., Peng, H., Luo, Y., Jankowitz, R. C., & Wu, S. (2018). A deep learning method for classifying mammographic breast density categories. *Medical Physics*, *45*(1), 314–321. https://doi.org/10.1002/mp.12683

Mohamed, A. A., Luo, Y., Peng, H., Jankowitz, R. C., & Wu, S. (2018). Understanding Clinical Mammographic Breast Density Assessment: a Deep Learning Perspective. *Journal of Digital Imaging*, *31*(4), 387–392. https://doi.org/10.1007/s10278-017-0022-2

Mohamed Abdel-Nasser, A. M., & Puig, D. (2016). Towards cost reduction of breast cancer diagnosis using mammography texture analysis. *Journal of Experimental & Theoretical Artificial Intelligence*, *28*(1–2), 385–402. https://doi.org/10.1080/0952813X.2015.1024496

Mohanty, F., Rup, S., Dash, B., Majhi, B., & Swamy, M. N. S. (2020). An improved scheme for digital mammogram classification using weighted chaotic salp swarm algorithm-based kernel extreme learning machine. *Applied Soft Computing*, *91*, 106266. https://doi.org/https://doi.org/10.1016/j.asoc.2020.106266

Mohd Ali, N., Ab Aziz, N. A., & Besar, R. (2020). Comparison of microarray breast cancer classification using support vector machine and logistic regression with LASSO and boruta feature selection. *Indonesian Journal of Electrical Engineering and Computer Science*, *20*(2), 712. https://doi.org/10.11591/ijeecs.v20.i2.pp712-719

Montavon, G., Kauffmann, J., Samek, W., & Müller, K.-R. (2022). Explaining the Predictions of Unsupervised Learning Models. In *xxAI - Beyond Explainable AI* (pp. 117–138). https://doi.org/10.1007/978-3-031-04083-2_7

Moore, K., Dalley, A., & Agur, A. (2018). THORAX. In *Clinically Oriented Anatomy* (8th ed., pp. 772–779). Wolters Kluwer.

Murtaza, G., Shuib, L., Abdul Wahab, A. W., Mujtaba, G., Mujtaba, G., Nweke, H. F., Al-garadi, M. A., Zulfiqar, F., Raza, G., & Azmi, N. A. (2020). Deep learning-based breast cancer classification through medical imaging modalities: state of the art and research challenges. *Artificial Intelligence Review*, *53*(3), 1655–1720. https://doi.org/10.1007/s10462-019-09716-5

Nakayama, L. F., Ribeiro, L. Z., Gonçalves, M. B., Ferraz, D. A., dos Santos, H. N. V., Malerbi, F. K., Morales, P. H., Maia, M., Regatieri, C. V. S., & Mattos, R. B. (2022). Diabetic retinopathy classification for supervised machine learning algorithms. *International Journal of Retina and Vitreous*, *8*(1), 1. https://doi.org/10.1186/s40942-021-00352-2

Narula, P. (2023). Analysis of Common Supervised Learning Algorithms Through Application. *Advanced Computational Intelligence: An International Journal (ACII)*, *10*(1), 29–48. https://doi.org/10.5121/acii.2023.10303

Nassif, A. B., Shahin, I., Attili, I., Azzeh, M., & Shaalan, K. (2019). Speech Recognition Using Deep Neural Networks: A Systematic Review. *IEEE Access*, *7*, 19143–19165. https://doi.org/10.1109/ACCESS.2019.2896880

National Cancer Institute. (2023). *Breast Health: Follow-up after an Abnormal Mammogram*. NIH.

Nazzal, Z., Sholi, H., Sholi, S. B., Sholi, M. B., & Lahaseh, R. (2018). Motivators and barriers to mammography screening uptake by female health-care workers in primary health-care centres: a cross-sectional study. *Lancet (London, England)*, *391*, S51. https://doi.org/10.1016/S0140-6736(18)30417-3

Nwadike, U. I., Eze, C. U., Agwuna, K., & Mouka, C. (2018). Mammographic classification of breast lesions amongst women in Enugu, South East Nigeria. *African Health Sciences*, *17*(4), 1044. https://doi.org/10.4314/ahs.v17i4.12

Obenauer, S., Hermann, K. P., & Grabbe, E. (2005). Applications and literature review of the BI-RADS classification. *European Radiology*, *15*(5), 1027–1036. https://doi.org/10.1007/s00330-004-2593-9

Odhiambo Omuya, E., Onyango Okeyo, G., & Waema Kimwele, M. (2021). Feature Selection for Classification using Principal Component Analysis and Information Gain. *Expert Systems with Applications*, *174*, 114765. https://doi.org/10.1016/j.eswa.2021.114765

Ogutu, J. O., Schulz-Streeck, T., & Piepho, H.-P. (2012). Genomic selection using regularized linear regression models: ridge regression, lasso, elastic net and their extensions. *BMC Proceedings*, *6*(S2), S10. https://doi.org/10.1186/1753-6561-6-S2-S10

Ongsulee, P. (2017). Artificial intelligence, machine learning and deep learning. *2017 15th International Conference on ICT and Knowledge Engineering (ICT&KE)*, 1–6. https://doi.org/10.1109/ICTKE.2017.8259629

Oostwal, E., Straat, M., & Biehl, M. (2021). Hidden unit specialization in layered neural networks: ReLU vs. sigmoidal activation. *Physica A: Statistical Mechanics and Its Applications*, *564*, 125517. https://doi.org/10.1016/j.physa.2020.125517

Palo, M., Schubert, B., Wei, J., Liu, W., Polenghi, M., & Ogliari, E. (2022). Unsupervised Machine Learning for Blind Separation of Multiple PD Sources. *2022*

IEEE 4th International Conference on Dielectrics (ICD), 197–201. https://doi.org/10.1109/ICD53806.2022.9863584

Pan American Health Organization. (2016). *Early detection: Breast health awareness and early detection strategies cancer*.

Pattanayak, S. (2017). Unsupervised Learning with Restricted Boltzmann Machines and Auto-encoders. In *Pro Deep Learning with TensorFlow* (pp. 279–343). Apress. https://doi.org/10.1007/978-1-4842-3096-1_5

Pes, B. (2020). Ensemble feature selection for high-dimensional data: a stability analysis across multiple domains. *Neural Computing and Applications*, *32*(10), 5951–5973. https://doi.org/10.1007/s00521-019-04082-3

Ponraj, N., Poongodi, & Mercy, M. (2017). Texture analysis of mammogram for the detection of breast cancer using LBP and LGP: A comparison. *2016 Eighth International Conference on Advanced Computing (ICoAC)*, 182–185. https://doi.org/10.1109/ICoAC.2017.7951766

Poornajaf, M., & Yosefi, S. (2023). Improvement of the Performance of Machine Learning Algorithms in Predicting Breast Cancer. *Frontiers in Health Informatics*, *12*, 132. https://doi.org/10.30699/fhi.v12i0.400

Prabhala, S., Srirambhatla, A., & Pasula, S. (2022). Comparison of BIRADS lexicon to breast biopsy findings in low resource countries. *ScienceRise: Medical Science*, *4*(49), 55–60. https://doi.org/10.15587/2519-4798.2022.262145

Prastyo, P. H., Paramartha, I. G. Y., Pakpahan, M. S. M., & Ardiyanto, I. (2020). Predicting Breast Cancer: A Comparative Analysis of Machine Learning Algorithms. *Proceeding International Conference on Science and Engineering*, *3*, 455–459. https://doi.org/10.14421/icse.v3.545

Pratiwi, M., Alexander, Harefa, J., & Nanda, S. (2015). Mammograms Classification Using Gray-level Co-occurrence Matrix and Radial Basis Function Neural Network. *Procedia Computer Science*, *59*, 83–91. https://doi.org/10.1016/j.procs.2015.07.340

Remeseiro, B., & Bolon-Canedo, V. (2019). A review of feature selection methods in medical applications. *Computers in Biology and Medicine*, *112*, 103375. https://doi.org/10.1016/j.compbiomed.2019.103375

Reyad, M., Sarhan, A. M., & Arafa, M. (2023). A modified Adam algorithm for deep neural network optimization. *Neural Computing and Applications*, *35*(23), 17095–17112. https://doi.org/10.1007/s00521-023-08568-z

Ribli, D., Horváth, A., Unger, Z., Pollner, P., & Csabai, I. (2018). Detecting and classifying lesions in mammograms with Deep Learning. *Scientific Reports*, *8*(1), 4165. https://doi.org/10.1038/s41598-018-22437-z

Samardar, P., de Paredes, E. S., Grimes, M. M., & Wilson, J. D. (2002). Focal Asymmetric Densities Seen at Mammography: US and Pathologic Correlation. *RadioGraphics*, *22*(1), 19–33. https://doi.org/10.1148/radiographics.22.1.g02ja2219

Samee, N. A., Alhussan, A. A., Ghoneim, V. F., Atteia, G., Alkanhel, R., Al-antari, M. A., & Kadah, Y. M. (2022). A Hybrid Deep Transfer Learning of CNN-Based LR-PCA for Breast Lesion Diagnosis via Medical Breast Mammograms. *Sensors*, *22*(13), 4938. https://doi.org/10.3390/s22134938

Santos, C. F. G. Dos, & Papa, J. P. (2022). Avoiding Overfitting: A Survey on Regularization Methods for Convolutional Neural Networks. *ACM Computing Surveys*, *54*(10s), 1–25. https://doi.org/10.1145/3510413

Saravanakumar, M., & Kannan, Dr. S. (2023). Pattern Recognition in Breast Cancer Using Machine Learning. *INTERANTIONAL JOURNAL OF SCIENTIFIC RESEARCH IN ENGINEERING AND MANAGEMENT*, *07*(03). https://doi.org/10.55041/IJSREM18255

Sarker, I. H. (2021). Machine Learning: Algorithms, Real-World Applications and Research Directions. *SN Computer Science*, *2*, 160. https://doi.org/10.1007/s42979-021-00592-x

Sasirekha, N., Karuppaiah, J., Shekhar, H., & Naga Saranya, N. (2023). Breast cancer detection using Histopathology Image with Mini-Batch Stochastic Gradient Descent and

Convolutional Neural Network. *J. Intell. Fuzzy Syst.*, *45*(3), 4651–4667. https://doi.org/10.3233/JIFS-231480

Sawssen, B., Taouali, O., & Liouane, N. (2021). A mammographic images classification technique via the Gaussian Radial Basis Kernel ELM and KPCA. *International Journal of Applied Mathematics Computational Science and Systems Engineering*, *2*(2766–9823).

Shafique, R., Rustam, F., Choi, G. S., Díez, I. de la T., Mahmood, A., Lipari, V., Velasco, C. L. R., & Ashraf, I. (2023). Breast Cancer Prediction Using Fine Needle Aspiration Features and Upsampling with Supervised Machine Learning. *Cancers*, *15*(3), 681. https://doi.org/10.3390/cancers15030681

Shen, L., Margolies, L. R., Rothstein, J. H., Fluder, E., McBride, R., & Sieh, W. (2019). Deep Learning to Improve Breast Cancer Detection on Screening Mammography. *Scientific Reports*, *9*(1), 12495. https://doi.org/10.1038/s41598-019-48995-4

Siddeeq, S., Li, J., Bhatti, H. M. A., Manzoor, A., & Subhan Malhi, U. (2021). Deep Learning RN-BCNN Model for Breast Cancer BI-RADS Classification. *2021 The 4th International Conference on Image and Graphics Processing*, 219–225. https://doi.org/10.1145/3447587.3447620

Simonyan, K., & Zisserman, A. (2014). Very Deep Convolutional Networks for Large-Scale Image Recognition. *ArXiv Preprint*, 1409–1556.

Singh, P., Gupta, S., & Gupta, V. (2023). Multi-objective hyperparameter optimization on gradient-boosting for breast cancer detection. *International Journal of System Assurance Engineering and Management*. https://doi.org/10.1007/s13198-023-01955-8

Sistaninejhad, M., Rasi, H., & Nayeri, P. (2023). A Review Paper about Deep Learning for Medical Image Analysis. *Computational and Mathematical Methods in Medicine*, *2023*, 7091301. https://doi.org/10.1155/2023/7091301

Smiti, A. (2020). When machine learning meets medical world: Current status and future challenges. *Computer Science Review*, *37*, 100280. https://doi.org/10.1016/j.cosrev.2020.100280

Snell, R. (2011). The Upper Limb. In *Clinical Anatomy By Regions* (9th ed.). Lippincott Williams & Wilkins.

Spak, D. A., Plaxco, J. S., Santiago, L., Dryden, M. J., & Dogan, B. E. (2017). BI-RADS® fifth edition: A summary of changes. *Diagnostic and Interventional Imaging*, *98*(3), 179–190. https://doi.org/https://doi.org/10.1016/j.diii.2017.01.001

Srinivas, P., Guggari, S., Darapaneni, N., Paduri, A. R., & Sudha, B. G. (2023). Feature Selection Algorithms: A Comparative Study. In A. Abraham, T. Hanne, N. Gandhi, P. Manghirmalani Mishra, A. Bajaj, & P. Siarry (Eds.), *Proceedings of the 14th International Conference on Soft Computing and Pattern Recognition (SoCPaR 2022)* (pp. 402–412). Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-27524-1_38

Sriram, G., Babu, T. R. G., Praveena, R., & Anand, J. V. (2022). Classification of Leukemia and Leukemoid Using VGG-16 Convolutional Neural Network Architecture. *Molecular & Cellular Biomechanics*, *19*(1), 29–40. https://doi.org/10.32604/mcb.2022.016966

Tahir Iqbal, Asif Farooq, Sarwar, N., Mohsin Ashraf, & Asma Irshad. (2022). Prediction of Breast Cancer Using Machine Learning Techniques. *BioScientific Review*, *4*(1), 59–75. https://doi.org/10.32350/BSR.0401.04

Tahmooresi, M., Afshar, A., Bashari Rad, B., Nowshath, K., & Bamiah, M. (2018). Early Detection of Breast Cancer Using Machine Learning Techniques. *Journal of Telecommunication, Electronic and Computer Engineering*, *10*(3–2), 21–27.

Tan, M., & Le, Q. V. (2019). EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks. *International Conference on Machine Learning*, *PMLR*, 6105–6114. https://doi.org/10.48550/arXiv.1905.11946

Tao, L. R., Ye, Y., & Zhao, H. (2023). Early breast cancer risk detection: a novel framework leveraging polygenic risk scores and machine learning. *Journal of Medical Genetics*, *60*(10), 960–964. https://doi.org/10.1136/jmg-2022-108582

Thang, N. D., Dung, N. V., Duc, T. V., Nguyen, A., Nguyen, Q. H., Anh, N. T., Cuong, N. N., Linh, L. T., Hanh, B. M., Phu, P. H., & Phuong, N. H. (2021). *Building a X-ray Database for Mammography on Vietnamese Patients and automatic Detecting ROI Using Mask-RCNN* (Vol. 899, pp. 315–329). Springer. https://doi.org/10.1007/978-3-030-49536-7_27

Tiwari, A. (2022). Supervised learning: From theory to applications. In A. Tiwari (Ed.), *Artificial Intelligence and Machine Learning for EDGE Computing* (pp. 23–32). Academic Press. https://doi.org/10.1016/B978-0-12-824054-0.00026-5

Tsai, K.-J., Chou, M.-C., Li, H.-M., Liu, S.-T., Hsu, J.-H., Yeh, W.-C., Hung, C.-M., Yeh, C.-Y., & Hwang, S.-H. (2022). A High-Performance Deep Neural Network Model for BI-RADS Classification of Screening Mammography. *Sensors*, *22*(3). https://doi.org/10.3390/s22031160

Tsochatzidis, L., Costaridou, L., & Pratikakis, I. (2019). Deep Learning for Breast Cancer Diagnosis from Mammograms—A Comparative Study. *Journal of Imaging*, *5*(3). https://doi.org/10.3390/jimaging5030037

Venkatesh, B., & Anuradha, J. (2019). A Review of Feature Selection and Its Methods. *Cybernetics and Information Technologies*, *19*(1), 3–26. https://doi.org/10.2478/cait-2019-0001

Vinutha, D. C., Sushma, S. J., Assegie, T. A., Paramesha, K., & Geethashree, A. (2022). Optimization of RFM for automated breast cancer detection. *International Journal of Health Sciences*, *6*(S1), 5973–5982. https://doi.org/10.53730/ijhs.v6nS1.6218

Wajeed, M. A., Tiwari, S., Gupta, R., Ahmad, A. J., Agarwal, S., Jamal, S. S., & Hinga, S. K. (2023). A Breast Cancer Image Classification Algorithm with 2c Multiclass Support Vector Machine. *Journal of Healthcare Engineering*, *2023*(1), 1–12. https://doi.org/10.1155/2023/3875525

Wan, Y., Tong, Y., Liu, Y., Huang, Y., Yao, G., Chen, D. Q., & Liu, B. (2022). Evaluation of the Combination of Artificial Intelligence and Radiologist Assessments to Interpret Malignant Architectural Distortion on Mammography. *Frontiers in Oncology*, *12*. https://doi.org/10.3389/fonc.2022.880150

Wang, J., Yang, X., Cai, H., Tan, W., Jin, C., & Li, L. (2016). Discrimination of Breast Cancer with Microcalcifications on Mammography by Deep Learning. *Scientific Reports*, *6*(1), 27327. https://doi.org/10.1038/srep27327

Wang, T., & Fan, Q. (2022). Machine learning-based prediction of dual primary breast cancer. *Archives of Medical Science*. https://doi.org/10.5114/aoms/151976

Wassim, A., Elarbi, E., & Khadija, R. (2022). Application of Machine Learning Approaches in Health Care Sector to The Diagnosis of Breast Cancer. *Journal of Physics: Conference Series*, *2224*(1), 12012. https://doi.org/10.1088/1742-6596/2224/1/012012

WHO. (2023). *Breast cancer*. World Health Organization.

World Health Organization. (2007). *Cancer Control: Knowledge into Action: WHO Guide for Effective Programmes*.

World Health Organization. (2016). Policy statement and recommended actions for early detection of breast cancer in the Eastern Mediterranean Region. In *World Health Organization*. Regional Office for the Eastern Mediterranean.

World Health Organization. (2023). *Global breast cancer initiative implementation framework: assessing, strengthening and scaling up of services for the early detection and management of breast cancer: executive summary*.

Yala, A., Lehman, C., Schuster, T., Portnoi, T., & Barzilay, R. (2019). A Deep Learning Mammography-based Model for Improved Breast Cancer Risk Prediction. *Radiology*, *292*(1), 60–66. https://doi.org/10.1148/radiol.2019182716

Yang, L., & Xu, Z. (2019). Feature extraction by PCA and diagnosis of breast tumors using SVM with DE-based parameter tuning. *International Journal of Machine Learning and Cybernetics*, *10*(3), 591–601. https://doi.org/10.1007/s13042-017-0741-1

Yoon, J. H., & Kim, E.-K. (2021). Deep Learning-Based Artificial Intelligence for Mammography. *Korean J Radiol*, *22*(8), 1225–1239. https://doi.org/10.3348/kjr.2020.1210

Zakareya, S., Izadkhah, H., & Karimpour, J. (2023). A New Deep-Learning-Based Model for Breast Cancer Diagnosis from Medical Images. *Diagnostics*, *13*(11), 1944. https://doi.org/10.3390/diagnostics13111944

Zebari, R., Abdulazeez, A., Zeebaree, D., Zebari, D., & Saeed, J. (2020). A Comprehensive Review of Dimensionality Reduction Techniques for Feature Selection and Feature Extraction. *Journal of Applied Science and Technology Trends*, *1*(1), 56–70. https://doi.org/10.38094/jastt1224

Zeiler, M. D., & Fergus, R. (2013). Visualizing and Understanding Convolutional Networks. *ArXiv*, *abs/1311.2901*. https://api.semanticscholar.org/CorpusID:3960646

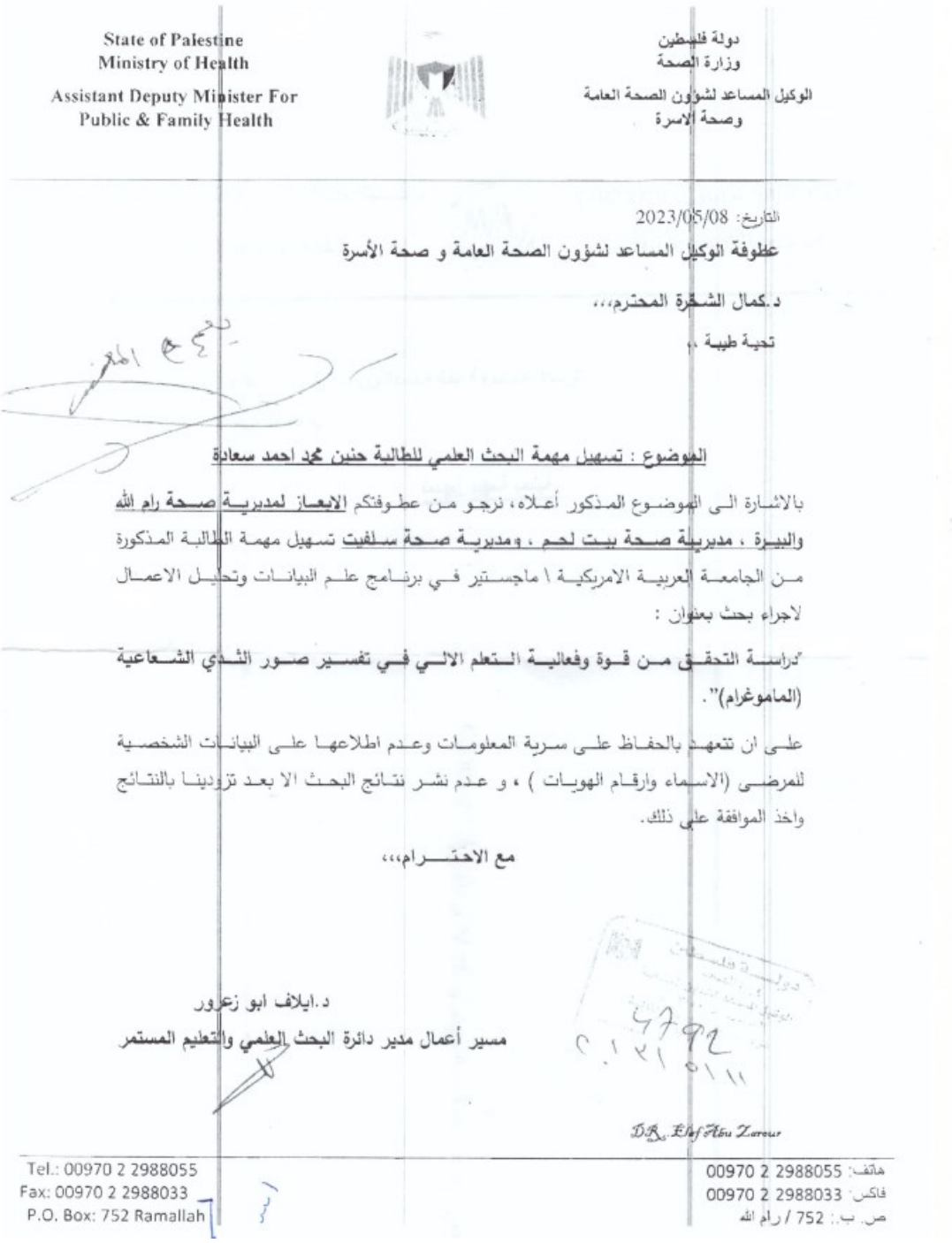
Zhang, H., Wang, J., Sun, Z., Zurada, J. M., & Pal, N. R. (2020). Feature Selection for Neural Networks Using Group Lasso Regularization. *IEEE Transactions on Knowledge and Data Engineering*, *32*(4), 659–673. https://doi.org/10.1109/TKDE.2019.2893266

Zhang, J. (2022). Mining imaging and clinical data with machine learning approaches for the diagnosis and early detection of Parkinson's disease. *Npj Parkinson's Disease*, *8*(1), 13. https://doi.org/10.1038/s41531-021-00266-8

Zhu, J. W., Charkhchi, P., Adekunte, S., & Akbari, M. R. (2023). What Is Known about Breast Cancer in Young Women? *Cancers*, *15*(6). https://doi.org/10.3390/cancers15061917

Appendices

Appendix (1) Data collection approval letter



State of Palestine
Ministry of Health
Assistant Deputy Minister For
Public & Family Health

دولة فلسطين
وزارة الصحة
الوكيل المساعد لشؤون الصحة العامة
وصحة الاسرة

التاريخ: 2023/05/08

عطوفة الوكيل المساعد لشؤون الصحة العامة و صحة الأسرة

د.كمال الشقرة المحترم،،،

تحية طيبة ،،

الموضوع : تسهيل مهمة البحث العلمي للطالبة حنين محمد احمد سعادة

بالاشارة الى الموضوع المذكور أعلاه، نرجو من عطوفتكم الايعاز لمديرية صحة رام الله والبيرة ، مديرية صحة بيت لحم ، ومديرية صحة سلفيت تسهيل مهمة الطالبة المذكورة من الجامعة العربية الامريكية \ ماجستير في برنامج علم البيانات وتحليل الاعمال لاجراء بحث بعنوان :

"دراسة التحقق من قوة وفعالية التعلم الالي في تفسير صور الشدي الشعاعية (الماموغرام)".

على ان تتعهد بالحفاظ على سرية المعلومات وعدم اطلاعها على البيانات الشخصية للمرضى (الاسماء وارقام الهويات) ، و عدم نشر نتائج البحث الا بعد تزودينا بالنتائج واخذ الموافقة على ذلك.

مع الاحتـــرام،،،

د.ايلاف ابو زعرور
مسير أعمال مدير دائرة البحث العلمي والتعليم المستمر

Dr. Elaf Abu Zarour

Tel.: 00970 2 2988055
Fax: 00970 2 2988033
P.O. Box: 752 Ramallah

هاتف: 00970 2 2988055
فاكس: 00970 2 2988033
ص. ب.: 752 / رام الله

Appendix (2) Medical History Report

PALESTINIAN NATIONAL AUTHORITY
MINISTRY OF HEALTH
BETHLEHEM

Name: **Address:** **Tel. #**
Date: March 20, 2023 **ID #:** **DOB:** Nov 01,1948

Referred by: *Dr.*
Learned by activities: *Medical* / Lecture / Friend / Literature /
Marital Status: ☐ Single ☐ Married ☐ Divorce ☐ Widow

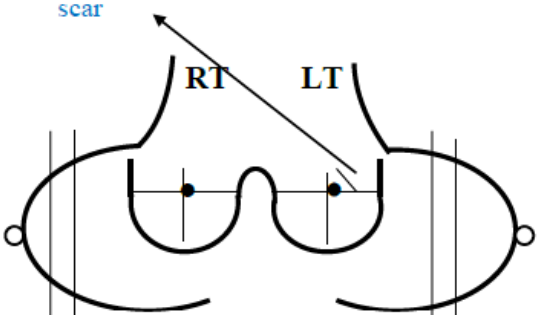
****

Medical History:

YES	NO		
☐	X	Previous problems	
X	☐	Previous Mammography	Date: Before 20 years Place:
		Result: Lt.Breast cysts.	
☐	X	Contraceptive Pill / Hormones :	
X	☐	Breast pain / tenderness.	Lt. / Rt. Since when ?
☐	X	Breast discharge /color/	Lt. / Rt. Since when ?
X	☐	Breast lump / Mass.	Lt. / Rt. Since when ?
☐	X	Previous breast OP / Biopsy.	Lt. / Rt. Date excision cysts
X	☐	Other disease- HT / Operation-	
☐	X	Family history of breast cancer Sister 50 / other Cancer disease.	
☐	X	Breast feeding now.	
☐	X	Pregnant. First menses 13 Marriage age 23 delivery age 24 # 4	
☐	X	Previous menses. Menopause 49	
		Weight 69	

COMMENTS: *No mass felt*
No nipple retraction.

scar



Performed By:

الملخص

ينتشر سرطان الثدي على نطاق واسع في جميع أنحاء العالم. إنه السبب الرئيسي المسبب في وفيات السرطان. بحسب التقرير السنوي لوزارة الصحة الفلسطينية، احتل سرطان الثدي المرتبة الثالثة من بين جميع وفيات السرطان في الضفة الغربية. لسرطان الثدي العديد من الأعراض مثل ألم الثدي، والإفرازات، والكتل. كما أن هنالك عوامل خطر يمكن تصنيفها إلى عوامل وراثية وعوامل تتعلق بنمط الحياة التي تساعد في الإصابة بالسرطان. بالإضافة إلى ذلك، يعد فحص تصوير الثدي بالأشعة السينية (الماموغرام) هو الأسلوب الأكثر شيوعاً لتشخيص تشوهات الثدي، ولكن هناك تحدياً يتمثل في عدم وجود خبراء ماهرين قادرين على تفسير صور الماموغرام بدقة. يلعب التعلم الآلي دوراً مهماً في معالجة الصور الطبية، خاصة في الكشف المبكر عن السرطان عندما يكون العلاج متاحاً وبأقل تكلفة. تقدم هذه الأطروحة طريقتين للكشف عن تشوهات الثدي اعتماداً على الشبكة العصبية التلافيفية (CNN). الطريقة الأولى صنفت الصور إلى طبيعية وغير طبيعية، في حين صنفتها الطريقة الثانية إلى درجات نظام تقارير وبيانات تصوير الثدي BI-RADS. بالإضافة، تم تطبيق ستة نماذج CNN في كلا المنهجين، وهم VGG16، VGG19، DenseNet121، ResNet50، Xception، وEfficientNetB7. مجموعة البيانات المستخدمة هي مجموعة بيانات فريدة وغير مستخدمة تم جمعها من وزارة الصحة الفلسطينية. بناءً على النتائج، تفوق DenseNet121 على النماذج الأخرى في الطريقة الأولى بـ 0.83 و0.85 لدقة الاختبار والمساحة تحت المنحنى (AUC) على التوالي. وفي المقابل، حصل كل من Xception وEfficientNetB7 على أفضل النتائج في الطريقة الثانية بدقة تزيد عن 90%. في المستقبل، يمكن دمج النماذج التي أظهرت أفضل النتائج مع بيانات المريض الأخرى مثل المعلومات الوراثية والتاريخ الطبي وعوامل نمط الحياة لتقييم خطر الإصابة بأمراض معينة مثل السرطان. وهذا من شأنه أن يزيد من معدل البقاء على قيد الحياة ويتيح اتخاذ تدابير استباقية. وأخيراً، ينبغي جمع مجموعة بيانات أكبر من صور الماموغرام لتحسين النتائج وتعميم النماذج على المهمات المشابهة.