Breast Abnormality Diagnosis Using Transfer and Ensemble Learning

by

Farnoosh Azour

A thesis
submitted to the University of Ottawa
in partial fulfillment of the
thesis requirement for the degree of
Master of Science
in
Biomedical Engineering

Ottawa, Ontario, Canada, 2022

Declaration of Authorship

I hereby confirm that this thesis and the work presented in it is entirely my own. Where I have consulted the work of others this is always clearly stated. All statements taken literally from other writings or referred to by analogy are marked and the source is always given. This paper has not yet been submitted to another examination office, either in the same or similar form. I agree that the present work may be verified with an anti-plagiarism software

Abstract

Breast cancer is the second fatal disease among cancers both in Canada and across the globe. However, in the case of early detection, it can raise the survival rate. Thus, researchers and scientists have been practicing to develop Computer-Aided Diagnosis (CAD)x systems. Traditional CAD systems depend on manual feature extraction, which has provided radiologists with poor detection and diagnosis tools. However, recently the application of Convolutional Neural Networks (CNN)s as one of the most impressive deep learning-based methods and one of its interesting techniques, Transfer Learning, has revolutionized the performance and development of these systems.

In medical diagnosis, one issue is distinguishing between breast mass lesions and calcifications (little deposits of calcium). This work offers a solution using transfer learning and ensemble learning (majority voting) at the first stage and later replacing the voting strategy with soft voting. Also, regardless of the abnormality's type (mass or calcification), the severeness of the abnormality plays a key role.

Nevertheless, in this study, we went further and made an effort to create a (CAD)x pathology diagnosis system. More specifically, after comparing multi-classification results with a two-staged abnormality diagnosis system, we propose the two-staged binary classifier as our final model. Thus, we offer a novel breast cancer diagnosis system using a wide range of pre-trained models in this study. To the best of our knowledge, we are the first who integrate the application of a wide range of state-of-the-art pre-trained models, particularly including EfficientNet for the transfer learning part, and subsequently, employ ensemble learning. With the application of pre-trained CNN-based models or transfer learning, we are able to overcome the lack of large-size datasets. Moreover, with the EfficientNet family offering better results with fewer parameters, we achieved promising results in terms of accuracy and AUC-score, and later ensemble learning was applied to provide robustness for the network. After performing 10-fold cross-validation, our experiments yielded promising results; while constructing the breast abnormality classifier 0.96 ± 0.03 and 0.96 for accuracy and AUC-score, respectively. Similarly, it resulted in 0.85 ± 0.08 for accuracy and 0.81 for AUC-score when constructing pathology diagnosis.

Acknowledgements

First, I am indebted to my supervisor; I would like to express my sincere gratitude to Prof. Azzedine Boukerche, a Canada Research Chair Tier-1 position with the University of Ottawa, for his effort and perfect supervision, continuous support, immense knowledge, continuous motivation, and patience during the whole time. Mainly, I would like to thank my supervisor for his extended financial support. Indeed, without this support, I would not be able to dedicate my whole time to research.

My gratitude extends to a gentleman who has significantly contributed in more than one way to bring this project to reality—Mr. Ali Hejazi for helpful discussions around my code and technical support.

I also extend my thanks to members of the examining committee, including Dr. Burak Kantarci and Dr. Richard Yu for critically evaluating my thesis and providing me with valuable comments about my research.

A heartfelt deepest gratitude is expressed to my first teachers Mojgan and Maziar (my dear mother and father), my siblings, Mehrnoosh, Amir Hossein, Payam, Yudum, my lovely nephews, Fardad, Barad, Arel, for their unconditional love and support. Without them, this journey would not have been possible, and to them.

Also, I have used the help of some friends and families, that I only can name a few; Mozhgan Nasr, Bedir Tapkan, Nikoo Aghaei, Amir Moslemi, Amir Hossein Azour, Yudum Karal, Leonardo Lai, and Lenin Falconi.

And to those lovely and supportive friends who helped me throughout this journey, including Dina, Salman, Bita, Parnian, Maziar Sh, Diba, Farzane Z and many others who I had the chance to spend my time with physically, and my lovely friends who live in my home-country, Iran, including Pegah, Mahroo, Hossein, Taha, Mohammad T, who gave me their warm and kind encouragement over their videocalls.

Dedication

My blessings are literally uncountable and, I would also like to dedicate this thesis to all the members of my caring and supportive and of course beautiful family, and the loveliest friends.

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

1.1 Breast Cancer and Statistics

Worldwide research has reported that the second leading cause of death among women is breast cancer [10]. Another recent article suggests that breast cancer is the most common type of cancer among women across the globe [130].. More specifically, the Canadian Cancer Statistics predicted estimates of cancer in Canada in 2020 in the form of a peer-reviewed article. According to this article that indicates cancer mortality and incidence estimates for 23 types of cancer, among women in Canada, breast cancer is the most widely diagnosed cancer. Similarly, in Canada, this type of cancer is the most prevalent among women [81]. Prior to that, in another study, it had been projected that in the year 2020, approximately a total number of 27700 would be diagnosed with breast cancer, among which 27400 and 240 people will be female and male, respectively. Furthermore, it is expected that 5100 people with breast cancer in Canada will lose their lives in 2020 [27]. Similarly, breast cancer is known to be the first leading cause of early death among women under 75 years old in Unites states [132].

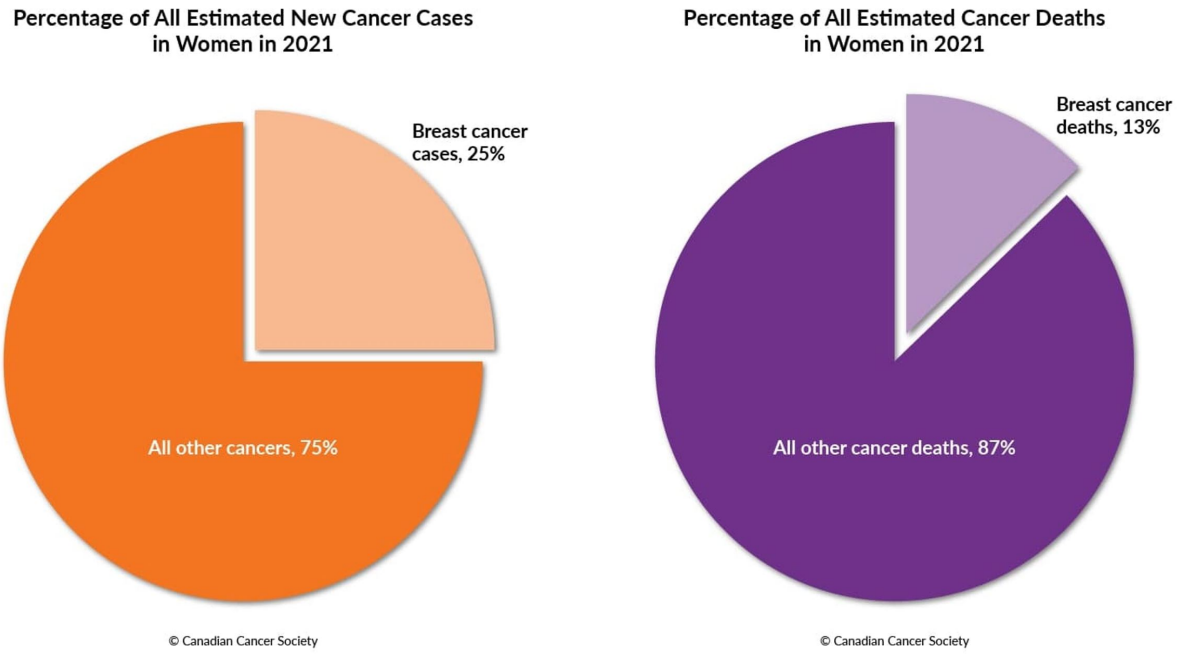

Figure 1.1: Canadian Cancer Society for breast cancer in 2021 [132]

1.1.1 What is Computer-Aided Diagnosis?

With the advancement of technology, medical data records are becoming more and more available. It has been reported that human beings tend to show high error rates in the clinical detection of breast cancer [123] [18]. The main underlying reason for this high rate of error is the low contrast which is apparent in mammograms [136]. This is sometimes known as the problem of low signal-to-noise ratio [94]. However, this rate was reduced where radiologists were asked to read the images a second time, yet it requires a higher workload and cost [152]. More specifically, around 10 to 30 percent of these cancers are missed by radiologists (False Negative Rates), and around 65 to 85 percent of women undergoing mammography are recalled for redoing the procedure (False Positive Rate), where they may even have to go under unnecessary biopsy [55]. Yet, eliciting meaningful information out of this large data set could be a cumbersome task.

1.2 Objective of this work

The prime goal of this research is to classify (mass or calcification) and diagnose (benign or malignant) abnormality mammogram patches using a range of pre-trained Convolutional Neural Networks (CNN)s or transfer learning. Unlike the conventional way of using hand-engineered features to develop a Computer-Aided Diagnosis (CAD)x, CNNs tend to find a feature vector automatically while the training occurs. Nevertheless, a large size of labeled data is required to train CNN-based models. Since we are performing our experiments on a limited dataset while carrying out our sets of experiments, this research seeks Transfer Learning to boost the results.

1.3 Contribution of this Thesis

As the second worldwide fatal disease, breast cancer requires Computer-Aided Diagnosis (CADx) systems to help radiologists. To this end, artificial intelligence-based methods have been applied by many researchers. The article contributes to the research community by:

1. giving a clear insight on how to leverage the power of artificial intelligence for Computer-Aided Detection/Diagnosis Systems using CNN-based architectures. This critical review of the state-of-the-art techniques is presented, which we believe can serve as a valuable source for research scientists investigating deep learning-based breast mammogram classification. .
2. proposing a two-stages classifer using the state-of-the-art pre-trained CNN-based models to develop a performant (CADx) system.

1.4 Dissertation Outline and Contributions

The thesis is outlined as follows:

1.4.0.1 Chapter 1: Introduction

This chapter provides the reader with information about breast cancer statistics and the need for a Computer-Aided Diagnosis system. In addition, we propose the objective of this thesis and summarize the contributions.

1.4.0.2 Chapter 2: Background

This chapter first familiarizes the reader with the required knowledge for understanding the thesis. We introduce different breast imaging modalities and then take an overview of deep learning transfer learning architecture. Later, Computer-Aided Detection (CADe) and Computer-Aided Diagnosis (CADx) are introduced. Also, we provide an overview of different publicly available data resources and available datasets for mammogram breast imaging. Then we conduct a literature survey where a critical review of the state-of-the-art techniques is presented.

1.4.0.3 Chapter 3: Methodology

This chapter shows the approach employed in the present study. We propose a two-staged abnormality diagnosis system using the transfer learning and ensemble learning technique, and we discuss the metrics by which we will evaluate our results.

1.4.0.4 Chapter 4: Results and Analysis

This chapter demonstrates the results yielded by conducting several experiments and evaluates them using the related evaluation metrics introduced in the chapter before. Tables and Figures are provided to compare the effect of each pre-trained model, transfer learning, and different ensemble learning strategies.

1.4.0.5 Chapter 5: Conclusion

This chapter summarizes the achieved results and offers ideas for future studies while concluding the underlying problem.

Chapter 2

Background

2.1 Overview

To find the gap between the application of the state-of-the-art deep learning models in breast cancer classification is, we conducted a survey on the works that had investigated breast mammogram classification in the past six years using publicly available datasets. However, since this research seeks to evaluate deep learning-based or CNN-based methods to perform breast cancer classification problems, we did not include the works that included machine learning techniques.

The remainder of this chapter is organized as follows. After describing breast cancer and different breast abnormality types in Section 2.2. We briefly review breast imaging methods and discuss their advantages and disadvantages in Section 2.3,. Then, we describe the cutting-edge approach and introduce a revolutionary subset of deep learning known as Convolutional Neural Networks (CNN), which has proven to help solve medical imaging problems in Section 2.4. Section 2.7 explores the main publicly available datasets used for this survey, and Section 2.8 will discuss CNN-based applications in breast mammogram classification. Eventually, Section 2.9 provides a number of concluding remarks using the surveyed articles.

In the field of breast cancer research, great literature surveys are been published in the past years. Particularly, for the application of (CNN)s, reference [10] recorded detailed CNN-based methods in mammography before 2017, when the evolution of pre-trained models had not yet flourished. Also, the survey did not contain any application of other deep learning-based methods. Another survey [104] published in 2019, provided a comprehensive overview of deep learning-based breast cancer classification using different modalities.

However, despite the considerable impact of CNN-based architectures and transfer learning and the huge attention researchers are paying to these architectures, the study did not seem exceptionally elaborate on pre-trained models. Compared with them, our survey not only describes the commonly used applications of deep neural networks in breast mammography classification, but also describes introduces state-of-the-art pre-trained CNN architectures.

2.1.1 Survey Scope Identification

The emergence of deep learning has once again pushed the image classification and, in particular, medical imaging classification task to a climax, and consequently, many research scientists have focused on that. We do not believe that a single survey can cover all research papers in this area. Therefore, we chose the impact of a sub-topic in breast cancer, so we decided on a sub-topic in computer-assisted detection and diagnosis systems that provide more promising results, namely, deep learning. The main objective of this survey is to assist the research scientists in constructing a novel and robust CAD system which is more computationally efficient and accurate in breast abnormalities detection and classification. Therefore, in this research, we aim to provide comprehensive and detailed answers to the following questions:

1. What are the medical imaging modalities, and why do we choose to focus on mammography?
2. What are the most common deep learning-based methods that play a crucial role in breast cancer classification?
3. What are the publicly available mammogram datasets used to develop deep learning-based classification models?
4. What are the concluding remarks based on deep learning's application in breast mammogram classification?

2.1.2 Inclusion/Exclusion Criteria

This section provides an insight into what studies have been used for the purpose of this survey.

2.1.2.1 Inclusion Criteria

The literature review is written based on the studies that included the application of CNN-based methods using the publicly available datasets, either to develop Computer-Aided Detection (CADe) or Computer-Aided Diagnosis (CADx), since 2015.

A few studies from histopathology were chosen for better explanation and comparison, even though more than 90% of the studies are chosen on mammogram imaging modality. Articles are manually selected from well-respected publications such as the Institute of Electrical and Electronics Engineers (IEEE), Elsevier, SemantiScholar, Society of Photo-optical Instrumentation Engineers (SPIE), Journal of Medical and Biological Engineering, Springer, PubMed, and Molecular Diversity Preservation International (MDPI).

2.1.2.2 Exclusion Criteria

Although we may have briefly referred to the wide application of deep learning methods in other modalities, this survey excludes studies involving other modalities such as ultrasound, MRI, PET scans, infrared, and histopathological images for our literature review. Moreover, articles investigating segmentation, prediction, and image retrieval are not of our main interest. Similarly, any literature performed prior to 2015 is not used for summarizing and presenting statistical information.

2.1.2.3 Selection Design

Figure 2.1 shows the workflow of a modern CAD system that works based on mammograms. Typically, such systems include multiple stages, namely segmentation, feature extraction, feature selection, and classification. However, on account of using deep learning's impressive ability to automatically extract features, many stages are embedded in the architecture.

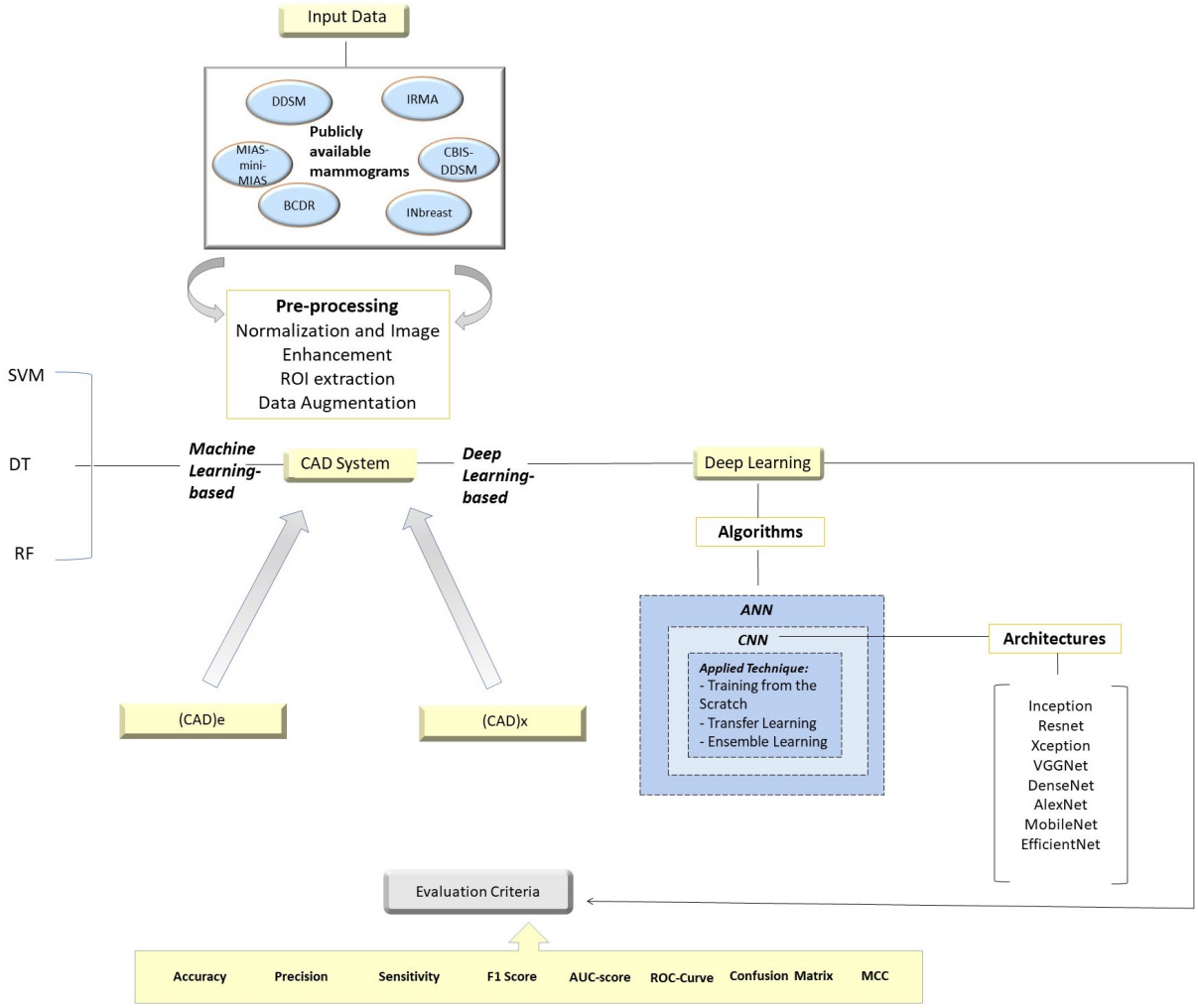

Figure 2.1: The general workflow of designing this survey-Support Vector Machine (SVM), Decision Tree (DT), and Random Forest (RF) are machine learning algorithms

This section begins with an overview of the most common breast imaging modalities and provides a detailed comparison. Subsequently, the knowledge required for later sections is provided. It is noteworthy to mention that, in the field of computer vision, classification refers to annotating the object categories, framing objects with bounding boxes that are colored for object localization, and object segmentation is defined by separating the edges of the object from the background [25]. Although numerous articles have been published on breast cancer segmentation, classification, and detection, this survey focuses on the

classification using publicly available mammogram datasets.

2.2 Breast Cancer and Types of Abnormalities

This type of cancer occurs due to the uncontrolled growth of breast cells and the nature of human anatomy; women are more vulnerable to that. The reasons behind the development of this disease are usually referred to as risk factors in the medical domain, and among them, smoking, genetics, working in an environment contaminated by chemicals, and infection agents has proven to be related to cancer development [82].

A women's breast anatomy consists of ducts, nipples, lobules, and fatty tissues [126]. According to scientific research, epithelial tumours or cysts usually grow in the ducts or inside the lobes and later change into a lump, which is responsible for cancer generation [82]. The first type of abnormalities is referred to the tumours or mass lesions. On the other hand, the second type of abnormalities is called microcalcifications, small granular calcium deposits. They may appear in cluster patterns (in linear or circular shapes) and reveal themselves as bright spots in the mammogram. Benign calcifications are often larger in size and coarser with smooth and round contours. However, while witnessing malignant calcifications, they appear to be numerous, small, clustered with various sizes, angular, or even irregular shapes.

2.2.1 BI-RADS numerical scoring tool

Knowing that breast abnormalities are twofold, both masses and calcifications can be either benign or malignant. Therefore, the American College of Radiology (ACR) has coined a term for a risk assessment tool, known as Breast Imaging Reporting and Database System score or BI-RADS [3]. Radiologists then will use this score to describe the results, using 6 categories of BI-RADS. According to this tool, the number 0 is given to patients who need to redo the test, and after the more, the number is, the more severe an abnormality can be [4].

2.3 Breast Imaging Modalities

The ultimate goal of medical imaging is to provide a more accurate diagnosis, treatment, and therapy for patients who are potentially diagnosed with a disease while trying to

alleviate the harmful side effects of the clinical procedure [43]. Consequently, this section summarizes the five uniquely different imaging modalities that play a major role in breast cancer analysis.

2.3.0.1 Mammograms

Mammograms are the low-dose X-ray image from the patient's breast and are often captured using two standard views: bilateral craniocaudal (CC) and mediolateral oblique (MLO). The images help radiologists check for any anomalies [51] [69]. The past twenty years have seen an increase in the popularity of this modality. Therefore, radiologists investigate the possible presence of mass, which could be either lump, cyst, or small calcium deposits that often appear in an irregular shape and are called micro-calcification. Currently, mammography falls under three main categories, Screen Film Mammography (SFM), Full Field Digital Mammograms (FFDM), and Digital Breast Tomosynthesis (DBT). The FFDM category of mammograms is also known simply as Digital Mammograms or DM. One downside of mammography is that its results are dependent on the lesion type, patient's age, and breast density. To be more specific, when it comes to denser breasts, they are "radiographically" harder to see and exhibit a lower contrast between cancerous parts and their backgrounds [82].

2.3.0.2 Ultrasound

The ultrasound imaging modality, also called a sonogram, is another popular imaging modality for breast cancer detection. In this clinical procedure, a high-frequency sound wave is sent into the breast, and the echo of the sound wave is employed for image generation. Unlike mammography or Magnetic Resonance Imaging (MRI), the application of ultrasound involves no radiation is, which seems to be a safe imaging modality [77]. Nonetheless, according to the Radiological Society of North America [2], when a significant amount of tissue is imaged, the quality of the observed image is poor. Furthermore, despite the advantages associated with this modality, it is an operator-dependent type of imaging, thus requiring an expert and knowledgeable radiologist for interpretation [77].

2.3.0.3 Magnetic Resonance Imaging

Magnetic Resonance Imaging (MRI) is a technology that applies magnetic fields and radio waves in order to capture a detailed view of the soft tissue, such as the lung, liver, and

breast. Consequently, the clear images captured by MRI machines can potentially show more information compared to mammograms, ultrasound, or even CT images. In addition to the diagnosis of suspicious areas, MRI can be used for biopsy, i.e., a process known as MRI-guided biopsy. However, since breast biopsy is rather invasive, MRI involves higher risks than other modalities. More specifically, when using MRI for a long time, a patient's body temperature could increase [89]. Due to clear and detailed MRI images that are often captured as a result of breast images, MRI is often requested after cancer has been diagnosed by a physician [104].

2.3.0.4 Histopathological Imaging

In this modality, namely histopathology biopsy imaging, tissue samples are collected from the suspicious breast regions and fixed on glass microscopes slides. The procedure works by using hematoxylin-eosin to stave these slides and examine them under a microscope by experts to diagnose the potential cancer tissue. Furthermore, these stained slides are scanned and subsequently converted to digital colored images known as Whole Slide Imaging (WSI) [71]. Pathologists, also known as cytologists, often extract regions of interest patches from these images using various zoom factors; this feature is not possible when using grayscale images. Apart from breast cancer detection, this imaging by this modality is considered the gold standard for many types of cancer due to these images' tissue level analysis feature. These cancer types may include liver, bladder, and lung [50] [71]. Thus, employing histopathology images has been quite popular for researchers who want to accurately perform multi-class breast cancer classification [65]. In addition to being time-consuming, the examination of the slides requires profound knowledge and expertise [56].

2.3.0.5 Other Imaging Methods

Table 2.1 presents a summary of the advantages and disadvantages of each of the above-mentioned breast imaging modalities. There are some rare breast imaging modalities such as Positron Emission Tomography (PET), Computed Tomography (CT), and infrared or thermal imaging. Nevertheless, according to some surveys focusing on several different modalities, [104] [100] these rarely used modalities have not been used for CAD development.

According to research studies, mammography is the most common type of breast imaging in (CADe) and (CADx) development [78]. In fact, many scholars refer to mammography as the gold standard [97]. Therefore, we decided to choose studies that have integrated publicly available mammograms and deep learning [146].

Table 2.1: Summary of the most popular breast imaging modalities

Medical Imaging Modality	Advantages	Disadvantages
Screen Film Mammography (SFM) [51] [114]	1) Highly sensitive for breast's fatty tissue 2) Cost efficient compared with HP and DM	1) Unsuitable for dense breasts 2) High dose of radiation 3) The imaging modality is not digital (inability to further process and enhance the images)
Full Field Digital Mammography (FFDM) Digital Mammography (DM) [53] [52]	1) Helpful in early cancer detection 2) Image resolution methods and improvements can be applied 3) Easier viewing, storage , and print 4) More effective modality than SFM (higher sensitivity)	1) Higher cost compared with SFM (approximately 1.5 to 4 times higher) 2) High dose of radiation 3) The potential for unnecessary biopsy due to high false positive results
Digital Breast Tomosynthesis (DBT) 3D Mammography [35] [138] [6] [156]	1) Significantly improves the performance including (accuracy, recall, and specificity) compared with DM, US, MRI 2) Provides radiologists with the ability to tackle overlapping breast tissue (having the ability to scroll through images) 3) lower rate of patients having recall	1) Entail higher cost relative to mammography and MRI 2) Higher amount of radiation compared with 2-D mammography due to being combined with regular mammography.
Ultrasound [77] [78] [155]	1) No radiation is involved (recommended during pregnancy) 2) Cost efficient 3) More suitable for dense breasts (compared with mammograms)	1) Depends on the operator's level of expertise (chance of misinterpretation if excessive probe compression is applied) 2) Less sensitivity (recall) compared with mammograms
MRI [117] [89] [77]	1) Can be used for higher risk patients 2) higher sensitivity compared with DM an US	1) Can increase the patient's body temperature 2) More expensive than DM and US 3) results in high rate of false positive rate (low specifity)
Histopathology Images [56] [50] [65]	1) Provides a comprehensive study for the tissue 2) Suitable for multi-classification purposes	1) Requires high proficiency while performing manual analysis 2) An invasive due to being associated with biopsy

2.3.1 CADe vs. CADx

While both the (CADe) and (CADx) are popular systems that assist doctors in interpreting their patient's disease, they hold distinct purposes. More specifically, (CADe) systems are geared to mark abnormal regions of an image, which in this paper is a mammogram [112]. They provide for the location of abnormal tissues. However, a (CADx) system determines the likelihood of an abnormality or the distinction between benign and mass lesions [101].

2.3.1.1 Traditional vs Modern CAD

Traditional CAD methods use image features mainly designed based on manually extracted features. Generally, detection of calcification, in this approach, is performed by image enhancement, frequency decomposition, stochastic modelling, and machine learning. Similarly, mass detection is followed by pixeled-based and region-based approaches [57]. Due to the advancement of deep neural networks, their exceptional ability for automatic feature learning from large training data has become possible. This has provided an end-to-end system for feature extraction to classifier building. Additionally, this learning technique is robust to dataset noise, which makes it suitable for abnormality detection and diagnosis in mammogram images [57] [131] [141]. Figure 2.2 compares the traditional and modern

CAD workflow.

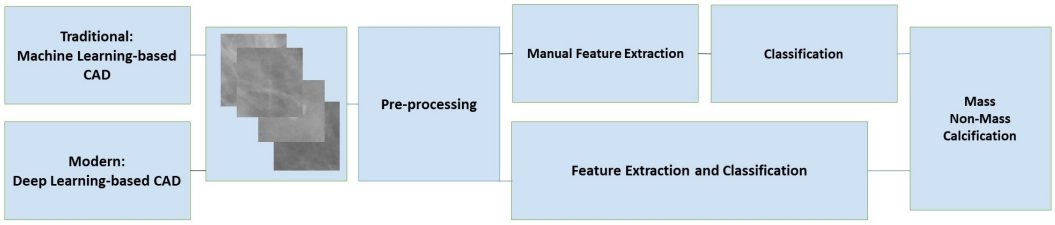

Figure 2.2: Comparison of traditional and modern CAD systems

2.3.2 Pre-processing

This section summarizes the best practices to improve deep learning performance while performing pre-processing prior to implementing the algorithms.

2.3.2.1 Normalization and Image Enhancement

Evidence suggests that the digitalization and acquisition of medical images depend heavily on the color and lighting condition. The goal of pre-processing is to improve the image quality and reduce noise while the diagnostic images remain intact. Various noises are likely to appear in mammogram images, such as Gaussian, salt and pepper, Poisson, and Speckle noise. Scientists tend to adopt two techniques to normalize the pixel values: global and local image normalization techniques. In the global method, the same operation is performed on all pixels of an image; however, using local normalization, an operation is performed on each pixel based on the intensity or contrast of the neighbouring pixels. Nevertheless, it has been recorded that global contrast normalization does not provide higher-quality image datasets [111]. Such noises can be reduced commonly through the application of the median filter, adaptive mean filter, Histogram Equalization (HE) [49], and Contrast Limited Adaptive Histogram Equalization (CLAHE) [48] [30].

2.3.2.2 ROI extraction

Segregation of the normal and abnormal tissues in mammograms is known as Region of Interest (ROI) extraction. Not only does this process increase the size of the dataset

for training, but it also provides the deep neural network with only normal and abnormal regions rather than irrelevant parts. For example, in the reference [11] the authors extracted ROIs using the sliding window approach and provided all the possible patches.

2.3.2.3 Data Augmentation

Data augmentation is one of the most effective techniques to reduce overfitting and improve the model's generalization, ultimately boosting its performance. Overfitting is defined as a situation where the model learns the details too well from the training data. However, it does not generalize well based on the training data to predict the unseen future data well. Therefore, the performance of the trained model is poor for the testing data despite having high validation accuracy. The mentioned situation usually occurs when the size of the dataset is small compared to the parameters of the model. Data augmentation occurs by applying transformations such as flipping, highlighting, rotation, focusing on some of them, and creating new images artificially. This approach has improved the accuracy in many recent studies, and further information is discussed in [148] [96] [23].

2.3.3 Deep Learning Background

Investigations on Artificial Neural Networks (ANN) dates back to the 1940 and 50s [36]. According to S.Pattanayak [109], deep learning is created due to ANN evolution, and Y.Guo et al. [64] claim it to be a fast-growing sub-field of machine learning that leverages the power of hierarchical structure architectures to learn high-level abstractions. On the other hand, deep learning today is considered as a sub-field of representation learning, which attempts to find features automatically through a general learning procedure to learn functions of high complexity [62]; therefore, the following definition for deep learning is proposed: deep learning is one of the representation learning methods, which utilizes a general-purpose procedure for learning in order to adjust the parameters (weights). This is performed to automatically extract the features or representations to solve artificial intelligence problems. [10]. This emerging new approach is utilized in various traditional artificial intelligence domains. Such domains include; semantic parsing [83], Natural Language Processing (NLP) [108] [121] [118], and last but certainly not least, computer vision [84] [103] and many others.

Generally, three reasons are said to be behind the booming of deep learning in today's world:

1. Chip processing abilities have dramatically increased (e.g., availability of GPUs).

2. Computing hardware cost has reduced [64].

3. Machine learning algorithms have advanced considerably [38] .

2.3.3.1 Deep Learning in Medical Imaging

Considering the increased availability of medical images and the pivotal developments of deep learning techniques, research scientists are replacing traditional machine learning-based strategies with the-state-of-the-art deep learning approaches. According to a survey conducted by Litjens et al. [96] on the application of deep learning in medical image analysis, the beginning of this trend started with unsupervised methods (parse auto-encoder and restricted boltzmann machine) [28] [113] [135]. Nonetheless, a clear shift towards applying CNNs was apparent after a while. CNN architecture was first implemented for image processing; consequently, applying CNN for image screening would provide a Computer-Aided solution for breast cancer detection and classification. Yet, the medical image analysis community has employed other supervised deep learning methods; perhaps one other potentially practical deep learning approach that plays a significant role is Recurrent Neural Networks (RNN)s. Despite having been applied for breast image analysis [154] in some research articles, we did not find a relevant article integrating the use of publicly available mammogram datasets and RNNs; thus, these studies are not summarized in this article.

The application of this state-of-the-art technology, namely CNN, has been the power to make enormous contributions, such as (CADe) or (CADx) tools the uncontrollable pandemic of COVID-19 [85]. The following section will summarize deep learning categories that have been the focus of research scientists integrating deep learning and (CADe)/(CADx) development.

2.4 Convolutional Neural Networks

Convolutional Neural Network (CNN) is the most impressive technique among all the different types of deep learning approaches for studying images. The idea was first proposed by LeCun et al. [91] in the 1990s and is inspired through analyzing animal's visual cortex [74] [58]. This method has worked completely effectively and is also the most widely used technique [87]. In general, three main neural layers make the CNN structure. These layers are convolutional layers, pooling layers, and fully connected layers, where each layer plays a specific role. As the name suggests in the convolutional layers, the convolution is

the main operation performed. In a clarifying example, this operation can be defined using the formula 2.1.

In a simple convolutional layer, a CNN uses various with a 1Dimensional input x and a 1Dimensional kernel size of k:

Kernels are for convolving the whole image along with the intermediate feature maps; this will result in the creation of new feature maps; three main advantages are associated with the convolution operation [64].

$$y[n] = (x * k)[n] = \sum_{-\infty}^{\infty} x[m]k[n-m] \tag{2.1}$$

- The weight sharing process for every same feature map decreases the parameter numbers compared with fully connected layers.
- Using local connectivity, correlations are learned among the neighbouring pixels.
- The translation invariance property resulted by incorporation of pooling layers.

There are two main steps regarding the training process: First, a forward step and then a backward step. Firstly, the main idea of having a forward step is to show the input image with the current parameters such as weight and bias in every layer. After that, the output for prediction is used for computing the loss cost along with the ground truth labels. Subsequently, the backward stage computes the gradient for every parameter with chain rules. All parameters will be updated derived from the gradients, and then these parameters are available for computing the next level. After iterating several times, the network can be stopped. The next thing is to introduce the activation functions according to recent development for each layer [64]. In other words, a CNN is a hierarchical neural network that has convolutional layers that change based on their pooling layers (sub-sampling layers) and then are followed by fully connected layers. The architecture of a simple CNN is shown using Figure 2.3.

2.4.0.1 Pooling Layers

In general, a pooling layer is a layer that follows the convolutional layer to reduce the dimensions of network parameters and the feature map. Similar to convolutional layers, these pooling layers are invariant to the translation. Two of the most commonly used

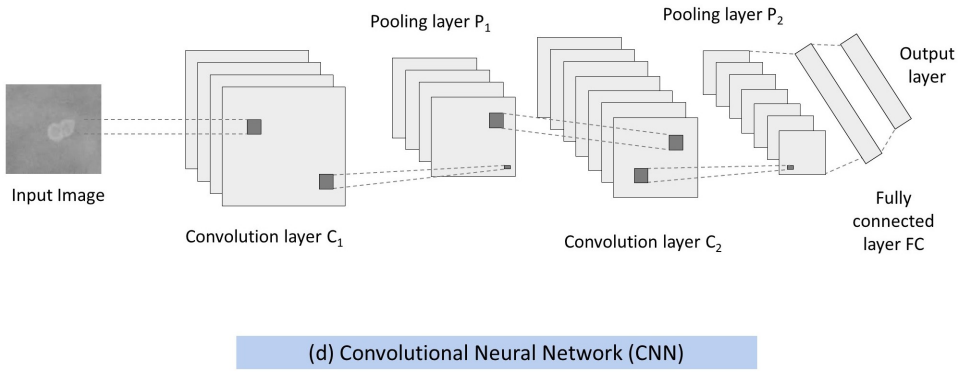

Figure 2.3: Simple architecture of a convolutional neural network

strategies are called max-pooling and average pooling. In a max-pooling example, for some 8×8 feature maps, the output map is made smaller into 4×4 dimensions, along with a max-pooling operator with the size of 2×2 and stride of 2. Regarding the max-pooling and average pooling, Boureau et al. [26]proposed a detailed theoretical analysis of the performances of these layers. Scherer et al. [124] later performed a comparison between the two common pooling strategies. They then stated that using max-pooling will help to achieve a higher convergence and choose superior invariant features and benefit generalization. In recent studies, CNN variants with different fast GPU implementations were introduced, and in all of them, the max-pooling strategy was used. The operation of this layer, namely, the pooling layer, was the most extensively studied by researchers. Other three famous approaches for the pooling layers, each one serving a different purpose, can be summarized in Stochastic Pooling, Spatial pyramid pooling (SPP), and Def-pooling.

2.4.0.2 Activation Function

For each feature map, an activation function is needed to bring non-linearity into the CNNs. Although Rectified Linear Unit (ReLU) seems to be the most common type among the activation functions, many variations of that have been applied recently. They include leaky ReLU, randomized ReLU, and parametric ReLU [61]. Also, regarding other types of activation functions, we can refer to tanh and sigmoid. Yet, the Sigmoid function has a serious disadvantage, which is known as the vanishing gradient problem. In this problem, the gradient that belongs to inputs of small values to the sigmoid function has a tendency to become small (nearly 0). Since the gradients are computed during the backpropagation, it results in a slow learning rate in earlier layers of the network. The slow

learning rate is not quite popular simply because it increases computation and cost [110]. ReLU Activation function, however, has almost become successful in solving the problem of vanishing gradient [61]. For positive and negative inputs, ReLU has a gradient of 1 and 0, respectively. Meaning that when the input is above zero, the gradient of the activation function is 1, and learning can be made. This is how ReLU addresses the vanishing gradient problem, which was apparent in sigmoid function. However, the disadvantage for the ReLU should not be neglected. This means that when the gradient is 0, the corresponding nodes would no longer have any influence on the network, and this problem is called the "dying ReLU" problem [98]. Instead of having zero as the output value when the input is negative, a Leaky ReLU was proposed. This type of ReLU function was able to offer a small negative slope (around of 0.01). This slope reduces sparsity; nevertheless, it makes the gradient optimize as the weight will be adjusted for the nodes, which were inactive using ReLU. In addition, if the value of is not constant, the function will be known as randomized ReLU [10].

2.4.0.3 Fully-Connected Layers

This layer is followed by having the last pooling layer, which was previously discussed. There are several fully connected layers that convert the two-dimensional features into a one-dimensional feature vector for representing later features. These networks act similar to a neural network and include approximately %90 of the parameters in the CNN. It makes us able to feed forward the network into a network with a length, which has been defined previously. Then, we can either feed-forward the vector into several categories for the purpose of image classification [144]. Or, we could take it as a feature vector for later processing [60]. Altering the structure of the fully connected layer is not very common. Nevertheless, a new example was presented regarding the transferring approach. It preserved parameters learned through ImageNet but used the last fully-connected layer as a replacement. This was performed for doing the new visual recognition tasks. The main drawback associated with the mentioned layers is that they include a large number of parameters, which requires high computational power for training. Thus, a promising approach that has proved to be successful for decreasing the number of parameters is to remove these layers. A good example could be GoogleNet, which was designed using a deep and wide network while keeping the budget dedicated to computation constant through switching from fully-connected architecture to the sparsely connected architecture [141].

2.4.1 The Strategies to Mitigate Overfitting Issue

In comparison to shallow learning, one advantage of deep learning is that it has the potential to create deep architectures so as to learn more abstract information. Nevertheless, having a large number of parameters causes another problem, overfitting. Recently, there have been many numerous regularization methods proposed to tackle the overfitting issue. One of the most famous techniques is to use stochastic pooling, which was previously mentioned. Here some regularization techniques will be introduced [64].

2.4.1.1 Hyperparameters

The variables that determine the network structure are known as a hyperparameter, such as the number of hidden layers. They also would include variables that determine the training of the network, for example, the learning rate. Hyperparameters should be chosen manually before training the network.

2.4.1.2 Dropout and DropConnect

Using dropout layers was proposed by Hinton et al. and was later explained more by Baladi et al. [21]. While doing this process, the algorithm will remove half of the feature detectors to prevent complicated co-adaptations on the data for training and, consequently, enhance the generalizing ability. This method was improved later in [150] [147] [151]. Particularly, Warde-Farley et al. [151]examined the efficiency of these layers and demonstrated that using dropout layers is an ensemble method, which proves to be effective. One famous method for generalization is derived from dropout layers is known as Drop Connect. Applying this technique will drop weights on random bases rather than activations. Observations have proven that with this method, the network can achieve competitive or even better results in terms of different benchmarks, even though it makes the training process slower.

2.5 Convolutional Neural Networks Training

Generally, CNNs are trained using two different approaches, namely training from scratch and tansfer learning.

2.5.1 Training from scratch

If the CNNs are provided with enough size training samples, they can be trained from scratch. This approach requires a considerable set of skills and experience. The most challenging tasks are selecting the hyperparameters such as the number of layers, the drop-out rate for each layer, the filter size for the convolutional layers, learning, regularization parameters, and activation function type. Consequently, the entire training process tends to be time-consuming and requires high GPU processing power.

2.5.2 Transfer learning

To propose a definition for using a framework, Pan and Yang [107] used domain, task, and marginal probabilities. Where a domain is presented with $D = \{\chi, P(X)\}$ that is consisted of two parts; the feature space is represented by χ and the marginal probability distribution is shown with $P(x)$. which is summarized $X = \{x_1, x_2..., x_n\} \in \chi$. A Task can be presented using T $= \{y, f(x)\}$ where y and $f(x)$ represent the label space and target prediction function, respectively. $f(x)$ can represent a conditional probability as well $P(y \mid x)$. Thus, transfer learning is formally defined as follows:

When D_s and T_s are given as source domain, and corresponding source task, and D_t and T_t are given as target domain and corresponding task, respectively, transfer learning refers to the process to improve the target predictive function $f_t(0)$ using the related information from D_s and T_s, knowing that $D_s \neq D_t$ or $T_s \neq T_t$.

2.5.2.1 The Application of Transfer Learning

The medical image community has taken considerable notice to employ transfer learning instead of training an entire CNN from scratch with random initialization. Applying transfer learning can be done either through fine-tuning a pre-trained network [125] [75] [44] [80] or through using a pre-trained network as feature extractor [54] [88] [106] [22]. The main reason behind using transfer learning in the medical domain is that data is expensive and scarce, and in many cases, it is not available to the public. Furthermore, collecting and labelling data by radiologists could be time-consuming [70] [33]. In addition, for a deep CNN to be trained, huge computational and memory resources are needed [141] [90] [63].

2.5.3 Pre-trained CNN-based Networks

The Large-Scale Visual Recognition Challenge (ILSVRC) and ImageNet dataset [37] are widely used by researchers as two main benchmarks over the past two decades [122] to evaluate the ability of deep neural networks. This challenge has evaluated several challenges, including image classification and object detection. The ILSVRC challenge resulted in several models pre-trained by the ImageNet dataset. Consequently, the year 2012 can be considered an essential landmark in this challenge, where Krizhevsky demonstrated the capability of CNN-based networks and his other colleagues [90]. Developers have been designing various architectures to evaluate them and consequently minimize errors.

2.5.3.1 AlexNet

The first pre-trained CNN that performed better than state-of-the-art in classification and object detection tasks was AlexNet. This pre-trained model, which uses 60 million parameters and 650,000 neurons, was first proposed by Krizhevsky, who was studying at the University of Toronto in 2012 [90]. The network includes eight layers; five layers of convolutions, which were followed by three layers of fully connected layers. The first layer performs the filtering of the input image, which is 224 $\times$ 224. The number of classes determines the number of neurons in the third fully connected layer.

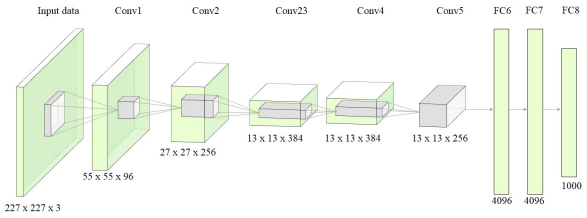

Figure 2.4: AlexNet Architecture

2.5.3.2 VGGNet

In 2014, the authors named Simonyan and Zisserman [131] from the University of Oxford studied the effect of network depth, considering the convolutions filters to be of a very small size. They proved that pushing depth up to 16-19 layers could significantly improve

the outcome. A fixed-size input of 224 × 224 is considered as the input of the convolution layer. VGG architecture created a more effective and enlarged receptive field of the network through stacking several layers of convolution, which have kernels of small size; at the same time, it reduces the number of parameters in comparison with kernels of large size. The authors examined several configurations with various depths (9,11,16, and 19 layers). In one of these configurations, 1 × 1 filters were used. This could also be seen in a linear transformation where the input channels are located. This method provides the decision function with more non-linearity without adversely affecting the receptive field of the convolutional layers. Furthermore, one of the configurations included the LRN layer as well. As it is mentioned in the paper, depths ranging from 16 to 19 yielded the best outcomes, where each one included 138 and 144 million parameters, respectively [131].

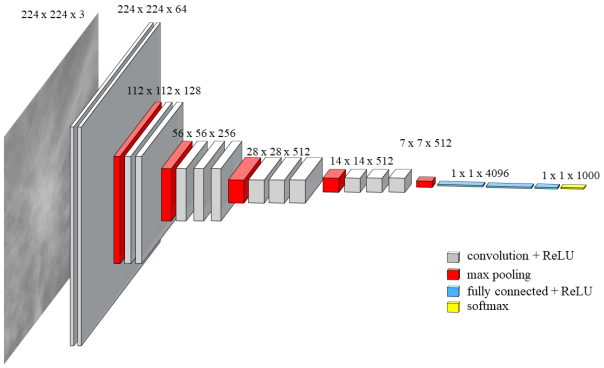

Figure 2.5: The VGG16 Architecture [145]

2.5.3.3 GOOGLENET/INCEPTION

In 2014, a model was introduced by Szegedy et al. and several other members of Google incorporation; GoogleNet [141] is the first implementation in which the Inception module is used. Finding how a network's dense components can estimate local sparse structure is the main idea behind GoogLeNet. The authors' goal is to achieve the optimal local structure and repeat it subsequently, thus creating a multi-layer network. This module, namely, Inception, is consisted of four branches that get the same input. These modules operate in a way that the first branch is responsible for filtering the input with a 1 × 1 convolution; this operation linearly transforms the input channels. The dimensionality is reduced for the second and third branches by performing 1 × 1 kernelled convolutions. Then, it is followed by other convolution layers with a kernel size of 3 × 3 and convolution layers of 5 × 5. The fourth one applies max-pooling, which is then followed by convolution

with a kernel size of 1 × 1. As for the final stage, the output of each branch is added and then fed into the next block as an input. GoogLeNet is built through stacking nine modules of Inception. To minimize the dimensionality of the feature maps, a max-pooling layer is placed in specific locations within the inception module. One noteworthy feature of GoogLeNet is its ability to incorporate auxiliary classifiers. Simple classifiers, two fully connected layers, and a softmax layer were added as middle layers of a CNN as it is assumed that these layers should produce discriminative features. These simple classifiers operate on the network, which is created as a result of an intermediate point of that one. After introducing the Inception-v1 in 2014, the authors upgraded their model and enhanced its performance by increasing the accuracy and reducing the time complexity. More specifically, Szegedy et al. [142] proposed InceptionNet-v3 in 2016 and introduced Inception-v4 and Inception-ResNet in 2017 [140].

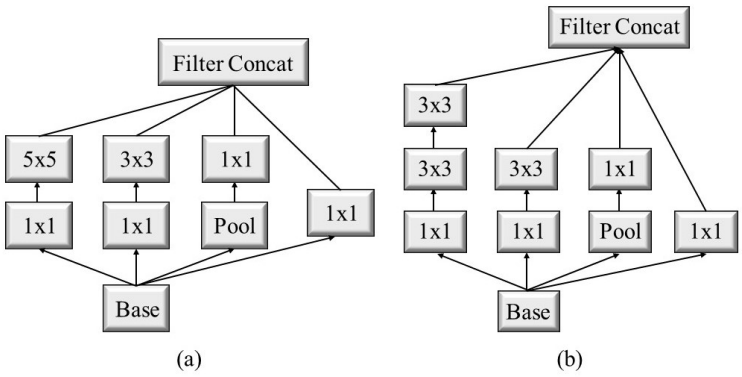

Figure 2.6: The architecture of GoogleNet with Inception Modules [141]

2.5.3.4 Residual Networks

Residual Networks (ResNets) [67], include convolutional layers that are reformulated in a way that they learn residual functions with reference to the inputs. The authors believe that using this type of network is more convenient to optimize and can be greatly increased in depth. Implementing a residual block seems to be straightforward. Meaning that for every few layers of convolutions, a layer which is known as shortcut connection is added; this runs in parallel to the other layers and provides the identity mapping. The output derived from the convolution layer will be added to the output resulting from the shortcut branch. After that, the result will be propagated to the next blow.

Figure 2.10 depicts this workflow. In addition to using shortcut connections, ResNet architecture is mostly inspired by the VGG network philosophy. All the layers of convolutions have kernels of small size around 3×3, and all follow simple rules for design:

1. If the output feature map has the same size, the layers should have the same number in terms of their filters.
2. the feature map size is split into halves (using convolutional layers with the stride of 2), these filter numbers are doubled to remain complexity of time per each layer. The authors, therefore, examined different architectures with varying depths, which ranged between 34 and 152 layers [67].

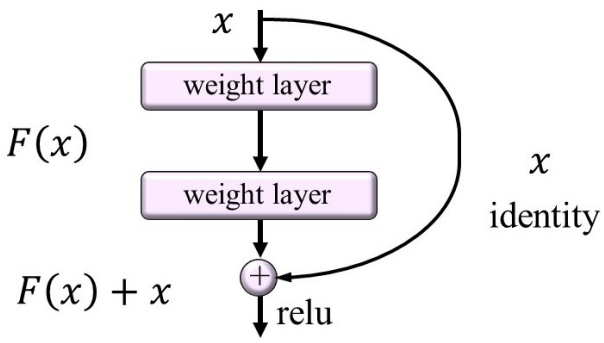

Figure 2.7: The skip connection architecture in ResNet [67]

2.5.3.5 DenseNet

Huang et al. [73], Facebook AI Research, proposed Densely Connected Convolutional Networks in 2016. The idea behind it was to create dense connectivity in the channel-wise concatenation. Each layer receives the preceding feature map for its input in this architecture, which helps mitigate the vanishing gradient problem. This problem often occurs in

the very deep networks that make the error 0 at some point during the backpropagation. The dense connections suggested by the authors also help in reusing the number of parameters. This phenomenon happens since the network receives the feature map information for the previous stage at each stage rather than creating more parameters. Remarkably, the Densely connected architecture of this network has been able to reduce the number of the parameters five times compared to the ResNet architecture while having the same number of layers.

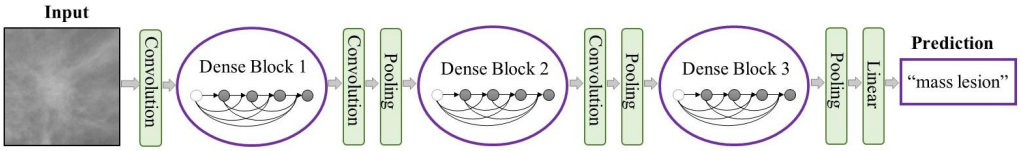

Figure 2.8: The dense blocks in a DenseNet architecture

2.5.3.6 MobileNet

A portable CNN architecture was first suggested by Howard et al. [72] in 2017. The authors offered depthwise separable convolutions instead of conventional convolutions used in the previous models to provide a lightweight model. The depthwise-separable convolution consists of two separable operations; depthwise convolution and pointwise convolution. By proposing this architecture, two global hyperparameters, namely, resolution multiplier and width multiplier, were introduced to control the input image resolution and channel-depth, respectively. These hyperparameters helped provide trade-off accuracy or latency for miniaturization and speed while developing based on the model developer's need.peed while developing based on the model developer's need.

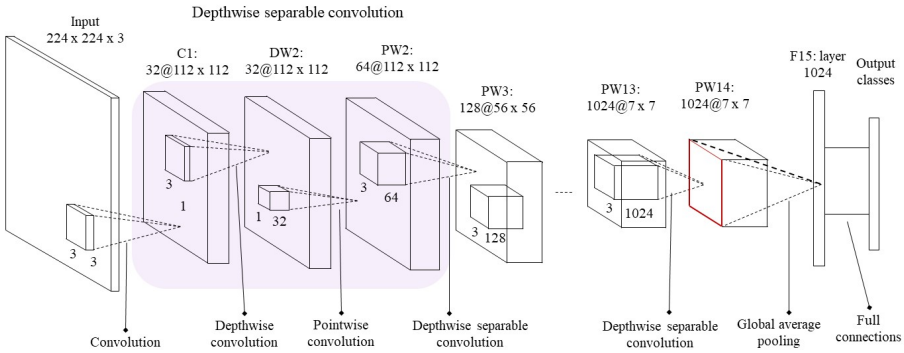

Figure 2.9: The implemented architecture in MobileNet

2.5.3.7 EfficientNet

In the same year, Tan and Le [143] offered what is known as compound scaling, which can set the EfficientNet apart from the other network architectures. By proposing this architecture, the authors demonstrated the effectiveness of applying compound coefficients to scale up ResNet and MobileNet architectures. To be more specific, the designers offered a strategy by which all the dimensions, namely depth (number of layers), image resolution (image size), or width (number of channels), are scaled while the balance between these dimensions and the network is maintained.

Table 2.2: Comparison of the pre-trained models

Model	Year	Model Variant	Param (Million)	Size (MB)	Depth	Top 5% Error Rates	Top 1% Error Rates	Input Size	Distinguishing Feature
VGG	2014	VGG-19	143.6 M	98	26	90.00%	71.30%	224 × 224	1) small size of convolutional filters 2) deeper architectures
		VGG-16	138 M	528	23	90.01%	71.13%		
ResNet	2015	ResNet-50	25.6 M	548	168	92.1%	74.9%	224 × 224	uses skip connections
		ResNet-152	60.4 M	232	152	94.29%	78.57%		
GoogleNet	2015	Inception-v3	23.8 M	92	159	93.70%	77.90%	299 × 299	brings the idea of multi-level feature extraction
DenseNet	2016	DenseNet-121	8.2 M	33	121	93.34%	76.39%	224 × 224	1) narrower layers 2) simplifying the connectivity
		DenseNet-169	14.3 M	57	169	93.2%	76.2%		
		DenseNet-201	20.2 M	80	201	93.6%	77.3%		
MobileNet	2017	MobileNet-v1	4.2 M	16	88	90.1%	74.90%	224 × 224	uses depthwise separable convolutions
		MobileNet-v2	3.5 M	14	88	92.10%	74.90%		
EfficientNet	2019	EfficientNet-b0	5.3 M	88	-	93.5%	76.3%	224 × 224	1) all dimensions are scaled in a uniform manner known as compound scaling
		EfficientNet-b3	12 M	29	-	95.6%	81.7%		

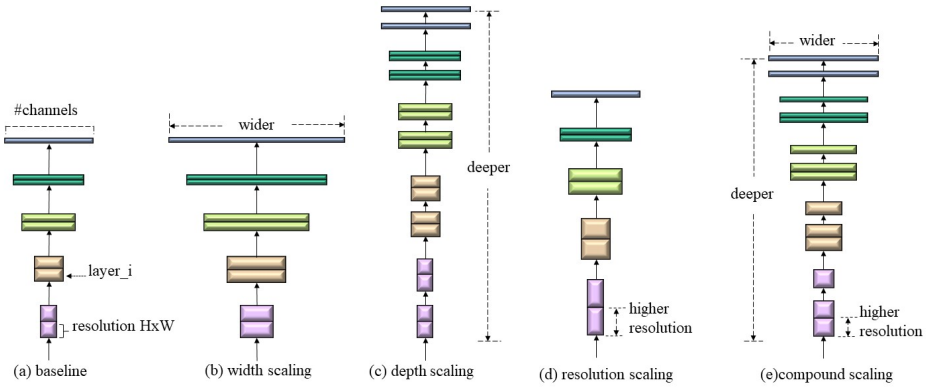

Figure 2.10: The idea behind compound scaling in EfficientNet architecture

Table 2.2 presents the above-mentioned pre-trained models about the year they have been released, the number of parameters, size, depth, and input size. Also, the top-1 and top-5 accuracy are indicators of the model's performance, where top-1 shows the accuracy for a single-class classification, and top-5 accuracy indicates that over five classes accuracy.

2.6 Ensemble Learning

A technique implemented on top of the other classifiers that is applied in this study is known as ensemble learning. In this approach, a number of classifiers are concatenated

so as to provide a more robust system. This technique has both industrial and academic application [157]. Voting has been introduced as one way of combining different models. Nevertheless, to have an appropriate ensemble model, individual classifiers need to be accurate, and divers [41]. The reason behind the necessity of diversity is that an ensemble classifier can only function if each individual classifier is different in terms of parameters, structure, and data [119] [99]. Voting is categorized into hard and soft voting, where hard voting is defined as the mode of the predicted classes that is also known as majority voting. In addition, in soft voting, the final vote is dependent on the probability score that is provided through each classifier.

2.6.1 Hard Voting

Consider performing standard supervised learning tasks, where the function ϕ needs to be found, which maps as features or the input space, to an output space Υ. With the use of CNN-based architectures as classification algorithms $y_i = \phi X_i$ for particular i as the test sample. Within the given frame of references, the vote by the majority or the mode of the predicted classes using weak classifiers defines the hard voting strategy in the ensemble learning universe. having n ensemble models, the hard voting strategy can be defined in:

$$y_i = mode\{\phi_1(X_i)\phi_2(X_i)...\phi_n(X_i)\} \tag{2.2}$$

2.6.2 Soft Voting

For the second strategy, consider the same supervised learning task from the probabilistic perspective. soft voting works in a similar manner to the hard voting with one important difference that is its dependence on the probability that each individual classifier returns. This time, $p_i = P_{(y_i|X_i)}$ is derived through the classification algorithm, where each classifier gives score p_i from the sample i that pertains to the y_i class. Therefore, soft voting can be defined using the formula below for n individual classifiers.

$$y_i = argmax \sum_{J=1}^{n} w_j p_{ij} \tag{2.3}$$

2.7 Publicly Available Datasets

Considering that most data algorithms have been tested on private data sets or other potentially insufficient subjects is an undeniable fact that providing large datasets can help the imaging research community. Several researchers have performed this task, particularly among all types of imaging modalities [120] [92].

Below is a brief description of the available datasets for the mammography research community, and Table 2.3 summarizes their characteristics. As it was previously mentioned in Section 2.3.0.1, currently, mammography is categorized into SFM, FFDM, and DBT [104]. In some cases of publicly available mammograms datasets such as the DDSM, CBIS-DDSM, and MIAS, the original version of mammogram images were SFM printed on large film sheets. They were later scanned for research purposes formed a valuable source for CAD development researchers. The last type, DBT, is also known as 3D mammography, which is likely to provide a more clear view for the radiologists; however, it is not available in all the imaging facilities [8]. Another point that is worthwhile to mention is that although images can be stored in different formats, for better and more efficient communication and management in the medical imaging field, Digital Imaging and Communications in Medicine (DICOM) is considered the standard format [5].

- Mammographic Image Analysis Society: A group of researchers collected this dataset in the UK named mammographic image analysis society (MIAS). The mentioned dataset is publicly available using the Pilot European Image Processing Archive (PEIPA), located at the University of Essex. The films are converted into a 200-µm pixel edge. Also, to make every image in the same size 1024×1024, images are padded/clipped. The dataset contains left and right breast images annotated by radiologists classified into normal and abnormal categories. The second category includes six sub-categories; calcification, asymmetry, well-defined, distortion, ill-defined and spiculated. Furthermore, the extend of severity in each abnormality is mentioned, i.e., benign, malignant. Nevertheless, since the number of normal, benign, and malignant samples is not normalized, in many studies, they have been used as a binary classification task [47].

- Digital Database for Screening Mammography: This remarkably huge dataset is a collection of mammogram images from several resources and was released in 1997. The Digital Database for Screening Mammography (DDSM) repository resources include Wake Forest University School of Medicine, Wake Forest University School of Medicine, Sacred Heart Hospital and Washington University of Sacred Heart Hospital, and Washington University of St Louis School of Medicine. The dataset contains

2620 scanned film mammograms, categorized into normal, benign, and malignant. For masses and calcification cases, annotations are provided and, the repository Also includes pathological patient information (BI-RADS) for the mass lesions [120]. Although the research community widely uses the DDSM mammogram images not saved in a standard format, requiring compression for research purposes. In addition, the lesion images are not precisely segmented, merely showing the general image position.

- Image Retrieval in Medical Applications: In 2008, The project Image Retrieval in Medical Applications (IRMA) [93] involved uniting several datasets MIAS, DDSM, and two other less popular datasets, namely the routine images from the Rheinisch-Westfälische Technische Hochschule (RWTH) Aachen and the Lawrence Livermore National Laboratory (LLNL). While the images of the IRMA dataset have a high-resolution and precise lesion segmentation, the mass lesions' shape variations are limited.

- Curated Breast Imaging Subset of DDSM: This dataset is an updated and standardized version of the DDSM dataset, including 861 and 753 images of masses and calcifications, respectively. The Curated Breast Imaging Subset of DDSM (CBIS-DDSM) provides the medical image analysis community with a powerful tool for CAD development. Although using DDSM has some substantial challenges, such as having scanned mammograms instead of full-field ones, R.Lee et al. [92] proposed that the scale of the data along with the verified pathology information makes it a suitable candidate to develop and test decision support systems. Other than providing an updated version of the DDSM, the authors improved ROI segmentation.

- Breast Cancer Digital Repository: This dataset includes several variations, and the BCDR-F03 dataset [1], the main subset of the Breast Cancer Repository, is being newly used in mammography classification and includes 736 lesions from 344 patients proven after undergoing biopsy. The clinical data from each patient is provided. This dataset, having both views, MLO and CC, which have made this possible to gather the coordinates in the counters of lesions. This is a binary class dataset consisting of benign and malignant mammogram findings. Table 2.3 summarizes the mentioned datasets in more details.

Although BCDR and INbreast datasets provide scientists with high-resolution images of FFDM, the number of instances is considered low, making them a relatively poor resource to be applied for training classifiers individually. Therefore, these datasets are often employed in combination with other potentially more extensive resources such as the

Table 2.3: Summary of publicly available datasets

Dataset	Year	Classes	Views	Type	Format	Numbers of patients	Overall number of images	Description
MIAS mini-Mias	2015	Normal Cancer	Only MLO	SFM	PGM	161	322	the size is relatively small. low-resolution images. only has one view. still widely used.
DDSM	1997	Normal Benign Malignant	MLO CC	SFM	JPEG	2620	10480	some image information is confusing the software is broken (unavailable) The format is not standard
IRMA	2008	Normal Benign Malignant	MLO CC	Both	PNG	-	3676	high -resolution images. precise lesion position available.
CBIS-DDSM	2017	Mass Clacification	MLO CC	SFM	DICOM	-	3012	Region of Interest (ROI)s are already extracted - pre-segmented
INbreast	2017	Normal Benign Malignant	MLO CC	FFDM	DICOM	115	419	limited size limited variations in mass shapes standard format of the files
BCDR	2012	Normal Benign Malignant	MLO CC	FFDM	DICOM	344	736	limited size precise lesion position available. having images of different resolutions standard format of the files

DDSM or CBIS-DDSM datasets; the experiments conducted by Alkhaleefah et al. [17] and Agarwal et al. [11] can be good exemplars of this fact. Having been released earlier than the others, the DDSM dataset has been chosen by many researchers. Nevertheless, its application has become less popular in the past few years, and we associate the emergence of the Curated Breast Imaging subset of the DDSM (CBIS-DDSM) dataset with that.

2.8 Applications of Deep Learning in Breast Cancer Classification

This section summarizes empirical analyses of different types of CNN-based (CADe/CADx) systems that have been developed using mammogram datasets. Most of the studies evaluated in this section include CNN compared with other deep neural networks.

Here is the summary of some related work in the computer vision domain for which various deep learning methods have been implemented. An effort has been made to summarize most of the works that have been done on mammogram classification using Deep Learning-based techniques, yet some of the studies that have been summarized are performed using other modalities such as histopathological.

2.8.0.1 CNN Applications in (CADe) and (CADx) Development

As explained in Section 2, with the advent of deep learning methods, particularly CNNs, the research focus for CAD development shifted from using hand-engineered features, machine learning to deep learning approaches, and as the trend continued, the application continued of CNNs began to flourish. The studies that are discussed in this section involve training CNNs from scratch. The first application of deep learning techniques for constructing a (CADx) system dates back to 2015 by Arevalo et al. [19] The authors proposed a CNN-based architecture including two convolution layers, two pooling layers, and one fully connected layer to extract meaningful features from the mass lesion images, and achieved 79.9% and 0.86 for the accuracy and AUC-score, respectively. Although their study defined a benchmark for other researchers, the training dataset was small (BCDR dataset has a limited size), increasing the possibility of overfitting. It is noteworthy to mention that research scientists have at the same time been conducting other experiments on other breast imaging modalities; for example, Spanhol et al. [133] applied AlexNet to create a CNN-based model for classifying the histopathological images into benign or malignant classes. Nonetheless, the researchers only explored one type of pre-trained model, AlexNet, one of the basic pre-trained models among others, making the study insufficient to generalize its findings. In the same year, Albayrak et al. [16] provided a feature extraction algorithm that was based on deep learning so as to detect mitosis in histopathological images. CNN model was applied for the purpose of feature extraction based on ImageNet dataset for mitosis classification.

In 2016, Jiao et al. [81] created an in-depth feature-based framework based on deep features by which they introduced a mass classification scheme. The scheme obtained 'middle-level' and 'high-level' features using a pre-trained CNN. The classifier was able to correspond to a coarser and a finer scale. Regarding the novelty of their technique, these features were mixed with the intensity information. That is to say; they combined intensity information for breast mass classification. In 2017, B.Swiderski et al. [139] argued that although the application of CNNs in image classification tasks is extremely useful, directly using them will not be fully successful. Therefore, they introduced a solution to solve the over-fitting problem created by limited training data. They applied Non-negative Matrix Factorization (NMF) and statistical self-similarity to richer training data. After performing 10-fold cross-validation, they achieved 85.82% accuracy in the case of normal and abnormal classification and 84.75% accuracy in the case of malignant and non-malignant categorization. In the same year, Dhungel et al. [40] proposed a fully automated multi-view ResNet (mResNet) to develop a (CADx) system. The authors demonstrated that combining both mammogram views (MLO, CC) of the INbreast dataset with the

lesion segmentation masks automatically generated during their experimentation yields promising results. The researchers reported the results of their experiments using five-fold cross-validation; however, owing to the limited size of INbreast dataset, the lack of data augmentation can be considered a drawback of their study, increasing the probability of overfitting.

Moya et al. [102] employed 1000 mammograms and achieved a recognition rate of 92.33% images from the DDSM dataset, which were divided into five categories based on their BI-RADS. Using these five different categories, they provided results for classifying breast cancer. Their results were later optimized using a grid search method that combined several parameter values, including batch size, loss function, optimizer, learning rate, and the number of layers frozen using the Inception-v3 model. The authors finally achieved 97% accuracy through using a stochastic gradient descent optimizer, batch size of 8 and 172 frozen layers. Even though their results seem promising, the lack of cross-validation or averaging makes the final result doubtful. Also, their work lacks data augmentation.

Hepsag et al. [68] employed 2-D convolutions on two datasets, namely, BCDR and mini-MIAS. Before that, they performed several pre-processing steps, including cropping, augmentation, and making the dataset balanced. When working on the BCDR dataset, the radius of circles was not given; therefore, the authors created masks to extract the ROIs. The researchers witnessed considerable achievements after performing these stepsand training CNNs. More specifically, the accuracy for BCDR rose from 65% to 85% after applying pre-processing steps.

In 2018, Al-masani et al. [15] implemented the "You Only Look Once" (YOLO) model to detect and classify mammogram images simultaneously. They applied transfer learning on the DDSM dataset to train their YOLO-based CAD system using the ImageNet datasets' pre-trained weights. Thus, their novel system could locate the masses with 99.7% accuracy using the YOLO model. Using the fully connected Neural Network, the proposed system distinguished the images to malignant with 97% accuracy after cross-validation.

In the same year, Ahmed and Salem [12] demonstrated the immense power of deep CNNs using the INbreast dataset. Also, they highlighted the importance of other evaluation metrics (such as recall) to construct a proper CAD system. After performing 5-fold cross-validation, they proposed their result to be 80.10% and 0.78 for accuracy and AUC-score, respectively. In addition, the authors provided a graphical user interface for the radiology community. Nonetheless, one main drawback of their work was that they only used the INbreast dataset, which has a limited number of images. More importantly, data augmentation could have been used to overcome small-sized dataset problems; thus, their experiments may suffer overfitting issues.

Table 2.4: CNNs for works for detection or diagnosis on publicly available mammograms

Author	Year	Dataset	Purpose	CNN-based Method	Aug	Cross Validation	Testing Accuracy	AUC
J.Arevalo et al. [19]	2015	BCDR	Diagnosis	CNN from scratch	yes	no	79.9%	0.86
Z.Jiao et al. [81]	2016	DDSM	Diagnosis	CNN from scratch	yes	yes	96.7%	-
B.Swiderski et al. [139]	2017	DDSM	Detection Diagnosis	CNN+ Non-Negative	yes	yes	85.82% 84.75%	0.91 0.90
E.Moya et al. [102]	2017	DDSM (only 1000 images)	Diagnosis	Matrix Factorization inception-v3	no	no	97%	-
N.Dunghel et al. [40]	2017	INbreast	Diagnosis	ResNet (mResNeta)	no	yes	-	0.8
K.Geras et al. [59]	2017	INbreast DDSM	Diagnosis	CNN from scratch	yes	no	68%	-
M.Jadoon et al. [76]	2017	IRMA, MIAS DDSM	Diagnosis	CNN+ Discrete Wavelet CNN+ Discrete Curvelet	yes	yes	81.83% 83.7%	0.83 0.83
P.Hepsag et al. [68]	2017	mini-MIAS BCDR	Diagnosis	CNN from scratch	yes	yes	85%	-
M.Al-masani et al. [14]	2017	DDSM	Diagnosis-Detection	CNN from scratch	yes	yes	97%	-
A.Ahmed et al. [12]	2018	INbreast	Detection	Deep CNN	no	yes	80.10%	0.78
A.Duggento et al. [42]	2019	CBI-DDSM	Diagnosis	CNN from scratch	yes	no	71%	0.77

Table 2.4 presents the summary of research studies that have employed CNNs for mammogram-based (CADe)/(CADx) development.

2.8.0.2 Transfer Learning Applications in CAD Development

Developing (CADe) based on the publicly available mammograms and transfer learning started in 2015 when Ertosun et al. [45] proposed a regional probabilistic system based on deep learning which was able to distinguish mammograms with a mass from the normal ones. Furthermore, this model had the ability to locate the masses after the image; some mass was detected in the mammogram. After applying data using cropping, flipping, scaling, and translation, they leveraged the power of transfer learning using AlexNet, GoogleNet, and VGGNet. Based on their experiments, the best accuracy, 85% result, was yielded while using GoogleNet. However, this accuracy is considered to be poor, specifically in the case of mass detection, where a false negative can be fatal.

In 2018, Xi et al. [153] investigated four different pre-trained CNNs to build a (CADe) system. This study was the first research that combined the CBIS-DDSM dataset and pre-trained models; thus, it established a baseline for the CAD research community. The

authors constructed a binary classifier that categorized the mammogram images to masses and calcifications using ResNet, VGGNet, GoogleNet, and AlexNet, and showed that VGGNet outperforms the others, having an average accuracy of 92.4%. Finally, they employed Class Activation Map (CAM) to illustrate to localize the lesions. Nevertheless, The authors could have investigated other pre-trained architectures (DenseNet and Xception) released prior to that had been proposed before.

In 2019, Agarwal et al. [11] proposed a CNN-based method to detect breast masses automatically. Their experiment included three pre-trained models, Inception-v3, VGG16, and ResNet-50. Overall, they performed three experiments; in the first one, using pre-trained models whose weights are obtained from the ImageNet database, they performed transfer learning on the CBIS-DDSM dataset. They then compared it against a randomly initialized CNN model. Inception-v3 outperformed these three models, with 84.16% as testing accuracy. In the second experiment, the authors claimed that since the INbreast and DDSM datasets are only different in their acquisition mode (SFM and FFDM), their CNN's feature space is highly likely to be relevant. Consequently, they chose the best model from the first experiment and fine-tuned it with the INbreast. After analyzing through five-fold cross-validation, the authors realized that the accuracy had improved to 88.86%. The third experiment of Agarwal involved mass detection using CNN patches. Although the authors' approach included some novelty in transfer learning, their final detection results were less than the previously published result by Xi et al. [153].

In 2018, Ragab et al. [115] investigated connecting the last fully connected layer of the AlexNet to an SVM to achieve better breast tumour classification results. Using this fine-tuned deep CNN-SVM, AlexNet architecture for feature extraction, 87.2% and 0.94 were obtained for accuracy and AUC-score, respectively. Also, their work experimented with two segmentation approaches, namely region-based and threshold segmentation. Despite the novel deep CNN-SVM approach they tested regarding CAD development, using a simple architecture like AlexNet can be considered a drawback. Table 2.5 presents (CADe) while demonstrating the experimentation results.

With regard to mammogram-based (CADx) systems, in 2016, Levy et al. [94] adopted a transfer learning method to classify the collected breast masses and achieved satisfactory results, which surpassed human performance. They carried out their experiments by carefully preprocessing and data augmentation and later applied AlexNet and GoogleNet pre-trained, models. Ultimately, they showed the interpretability of the model using saliency maps and achieved 92.9% and 93.4% for accuracy and recall, respectively, when performing transfer learning using GoogleNet. However, the researchers only investigated two pre-trained CNN architectures, making their results insufficient to generalize.

Table 2.5: Transfer Learning works for detection using mammograms

Author	**Year**	**Dataset**	**Transfer Learning Architecture**	**Aug**	**Cross Vali**	**Testing Accuracy**	**AUC**
M.Ertosun et al.	2015	DDSM	VggNet/AlexNet GoogleNet	yes	no	85%	-
P.Xi et al.	2018	CBIS-DDSM	AlexNet/VGGNet GoogleNet/ResNet	yes	no	92.53%	-
R.Agarwal et al.	2019	INbreast/CBIS-DDSM	VGG-16/ResNet Inception-v3	yes	yes	88.86%	-
D.Ragab et al.	2019	CBIS-DDSM DDSM	AlexNet (SVM)	yes	yes	87%	0.94

In the following year, Suzuki et al. [137] leveraged the power of deep CNN to prove that transfer learning has a promising potential. They investigated a pre-trained architecture called AlexNet for mass detection and claim to have proposed the first work on mammogram classification using deep CNN. The mammogram images chosen from the DDSM dataset were 198, consisting of 90 and 109, for normal and massed, respectively. Their experiments ultimately resulted in a sensitivity of 89.9%. Nevertheless, they are the authors who have proposed the only mammogram-based CAD system in our research without the use of accuracy as the evaluation metrics, which provides doubtful information on the False Positive response.

In 2017, Jiang et al. [79] employed two pre-trained architectures, AlexNet and GoogleNet, using data augmentation on the BCDR dataset. The objective of their study was to classify mass lesions into malignant and benign images. They achieved the AUC-score of 0.88 and 0.83 for GoogleNet and AlexNet, respectively, demonstrating the better ability of GoogleNet for a potential (CADx) system. Their results were improved compared with the similar research conducted by Arevalo in 2015; nevertheless, besides the limited dataset they used for their study, they could explore more impressive pre-trained models proposed to solve the ImageNet challenge.

Furthermore, in 2018, Chougrad et al. [34] evaluated VGG-16, ResNet-50, and Inception-v3 using three different datasets for training; DDSM, BCDR, and the INbreast, with the help of data augmentation. Also, they utilized MIAS dataset for testing the best model. The authors conducted one particular experiment to demonstrate that transfer learning or using the weights derived from ImageNet dataset classification outperforms random weight initialization. They finally proved that fine-tuned inception-v3 on the merged dataset. Therefore, the model was selected as the final classifier to categorize the mass lesion into malignant and benign.mass lesion is malignant or benign.

In 2019, Nasir Khan et al. [105] proposed a model known as a Multi-View Feature Fusion (MVFF) that was based on a number of CNN-based architectures, two well-known variants of VGG (VGG-16 and VGG-19), Inception-V3, and ResNet-50. Besides transfer learning using fine-tuning approach and reporting the results, one main novelty of the authors was that his work included four various views instead of single-viewed systems. Their final model consisted of three main stages; at the first stage, they proposed a model that was able to classify between normal and abnormal images using a mini-MIAS dataset. At the second one, their model could distinguish between masses and calcifications. The authors reported their results for each stage after averaging each metric five independent times. Therefore, the researchers were able to develop a wholes (CADx) system while raising more time complexity due to three-staged architecture. Although their paper presented a novel approach in the (CADx) development, the proposed accuracy for the third stage is considered relatively poor.

Chen et al. [31] conducted an empirical study on the DDSM dataset using fine-tuned ResNet. Ironically, all images in the CBIS-DDSM dataset were augmented, whereas only the train data should be augmented in the deep learning training process. Despite the final promising results, 93.15%, 93.83% for accuracy and sensitivity, respectively, their results keep us unconvinced since they tested their (CADx) on the augmented dataset.

Falconí et al. [49] have conducted extensive experiments to use transfer learning for the classification of transfer learning while investigating newly introduced pre-trained architectures. The authors explored the CBIS-DDSM dataset using two impressive CNN architectures that had not been thoroughly explored up to that time, known as MobileNet and NasNet, to develop a (CADx) system and compared their results with Inception-v3 and ResNet architectures. They demonstrated that the results obtained using ResNet-50 and MobileNet outperformed the others, 78.4%, and 74.3% for each, respectively. Not having applied data augmentation and observing the model's tendency to overfit, in the following year, 2020, they extended their work by classifying the dataset using ResNext, ResNet, VGG, and Xception. In a similar fashion to their previous work, the objective was to classify the mass on the CBIS-DDSM dataset. To improve the quality of dataset images, CLAHE was used, and in order to deal with the problem of overfitting, the authors employed several augmentation methods such as rotation, flipping, shearing, zooming, adjusting the brightness, and HE. Moreover, they used image filtering to deal with the noise and some artifacts presented on the images. Among all the models, they implemented VGG-16 with 0.844 was known to be the winner [48].

In 2020, Al-antari et al. [14] adopted the YOLO algorithm on DDSM and INbreast datasets, and after achieving detection accuracies of 99.17% and 97.27%, respectively, they classified these datasets using CNN. They performed the experiment by augmenting each

mammogram 22 times and applying transfer learning using pre-trained ResNet-50 and InceptionResNet-v2. Also, even though they proposed their paper in 2020, when many strong pre-trained models such as Inception-v3, DenseNet, had been released, they only tested their data using ResNet-50.

In the same year, Cao et al. [30] applied a Multi-Tasking U-shaped Network (MT-Unet). Such a U-shaped classification network was able to adapt to heterogeneous breast nature. They examined the performance of their model on the DDSM and INbreast datasets and demonstrated that in addition to transfer learning, which addresses the overfitting problem, label smoothing in the training process could greatly improve the classification accuracy. However, according to our survey, their research is one of the few studies that did not include data augmentation, and their result might suffer from overfitting.

Again in 2020, Alkhaleefah et al. [17] investigated double-shot transfer learning to improve the ability of their proposed (CADx) system. To this end, they argued that CNN networks that are trained on a large benchmark dataset like ImageNet would not yield good results. The authors believed that the ImageNet dataset lacks labelled mammogram images or any medical images. Thus, they decided to update the weights by feeding a large dataset such as CBIS-DDSM to a pre-trained model like AlexNet, and later testing these models on a target dataset, which in their case are MIAS and BCDR mammogram images. They performed their experiment through using the AlexNet, GoogleNet, VGG-16, VGG-19, MobileNet-v2, ResNet-50, ResNet-101, and ShuffleNet.

In 2020, Al-Antari et al. [13] employed a Full-resolution Convolutional Network (FrCN) to perform mammogram segmentation in a novel manner. Later, using three-CNN-based models, they categorized the already detected and segmented images into benign and malignant groups. By implementing CNN, ResNet-50, and InceptionResNet-V2, their experiments resulted in average overall accuracies of 88.74%, 92.56%, and 95.32%, respectively INbreast dataset. Tabel 2.6 shows the result of recent studies integrating the use of publicly available datasets and pre-trained models for (CADx) development.

2.8.0.3 Ensemble Learning applications in (CADe) and (CADx) development

In 2019, Shen et al. [127] employed CBIS-DDSM and INbreast datasets to compare the classification results using several well-known architectures, both individually and when averaging the top four models. They introduced a staged CNN-based scheme and explored different values for the learning rate. They witnessed that the AUC-score increased from 0.88 to 0.91 in the CBIS-DDSM dataset after four model averaging. Similarly, this value improved from 0.95 to 0.98 after the same operation on the INbreast dataset. Ultimately,

Table 2.6: Transfer Learning works for diagnosis on mammograms

Author	**Year**	**Dataset**	**Transfer Learning Architecture**	**Aug**	**Cross Vali**	**Testing Accuracy**	**AUC**
D.Levy et al.	2016	DDSM	AlexNet, GoogleNet	yes	no	92%	-
S.Suzuki et al.	2016	DDSM	AlexNet	no	no	89.9% sensitivity	-
F.Jiang et al.	2017	BCDR	AlexNet, GoogleNet	yes	no	-	0.88
H.Chougrad et al.	2018	DDSM INbreast BCDR	Inception-v3 ResNet-50 VGG-16	yes	yes	96.22% 92.00% 96.00%	0.98 0.97 0.96
N.Khan et al.	2019	mini-MIAS CBIS-DDSM	VGGNet/ResNet/ Inception-v3+ Feature Fusion technique	yes	no	93.73% 92.29% 77.66%	0.93 0.90 0.75
L.Falconi et al.	2019	CBIS-DDSM	NasNet, MobileNet InceptionNet-v3, ResNet	no	no	78.4%	-
L.Falcooni et al.	2020	CBIS-DDSM	VGG, ResNet, Resnext,Xception	yes	no	-	0.84
M.Alkhaleefah et al.	2020	BCDR CBIS-DDSM MIAS CBIS-DDSM	(Double Shot) AlexNet/VGGNet/ResNet/ GoogleNet/MobileNet-v2/ShuffleNet	yes	no	86.11% 93.86%	0.94 0.99
M.Al-antari et al.	2020	DDSM INbreast	Inception-v3-InceptionResNet CNN from the scratch	yes	yes	97.50% 95.32%	0.97 0.95
H.Cao et al.	2020	DDSM INbreast	U-Shaped Network ResNet-50	no	no	98.17% 93.91%	0.99 0.97
M.Al-antari et al.	2020	INbreast	CNN from the scratch ResNet-50 InceptionResNet	yes	yes	88.74% 92.33% 95.32%	0.87 0.97 0.93

the architecture designed by the researcher was able to perform malignant and benign classification in addition to localization.

The empirical study performed by Rampun et al. [116] in 2018 suggests that good results can be achieved by simply modifying the AlexNet architecture and applying ensemble learning. To be more specific, the modification was to replace the original Rectified Linear Unit (ReLU) with a more sophisticated activation function known as Parametric Rectified Linear Unit (PReLu) and add more drop-out layers. They later proposed an ensemble model based on the average probability of the sub-architectures. Their experiments resulted in 80.4% and 0.84 for accuracy and AUC-score, respectively. Nevertheless, as the authors mentioned, the achieved result does not surpass the radiologist's performance.

One of the most notable works that used ensemble learning rather than classifying the dataset was the study conducted by Arora et al. [20] in 2020. After applying some visualization improvement techniques (image-guided filters and HE), the authors employed several models such as VGG-16, AlexNet, GoogleNet, InceptionResNet-v2, and ResNet-18 as the sub-architectures of their model. After that, they proposed their final classifier using an ensemble learning technique called stacking, using an MLP architecture for their meta-model. They provided a (CADx) and reached an area under the curve of 0.88 with the classification accuracy of 88%. One drawback of their work was that they did not perform cross-validation, making their results less valid due to the possibility of overfitting. A summary of recent mammogram-based CAD research studies using ensemble learning is provided using Table 2.7.

Table 2.7: Summary of works that include ensemble learning for mammogram-based CAD development

Researcher	**Year**	**Sub-architectures**	**Ensemble Learning Strategy**	**Dataset**	**Purpose**	**Aug**	**Cross Val**	**Accuracy**	**AUC**
L.Shen et al	2017	VGG16 ResNet-50 and their combination	Model Averaging	CBIS-DDSM	Detection	yes	yes	92.30%	0.92
A.Rampun et al.	2018	AlexNet and Modified AlexNet	Model Averaging	CBIS-DDSM	Diagnosis	yes	yes	80.4%	-
R.Arora et al.	2020	VGG16 GoogleNet ResNet-18 InceptionResNet	Stacked Generalization	CBIS-DDSM	Diagnosis	yes	no	88%	0.88

2.9 Conclusion of the Survey

In this survey, we carried out a detailed review of the advantages, disadvantages, and performance of the main deep learning techniques that are popular with medical imaging researchers.

2.9.1 Overall methods

According to Figure 2.11, transfer learning, accounting for 57.57% of the literature reviewed, is by far the most popular method to solve mammogram classification problems. Also, model concatenation or ensemble learning was only employed in 9.09% of the reviewed materials, which was employed to construct the sub-architectures for the large ensemble model, where model averaging (soft voting) is found to be the most popular strategy.

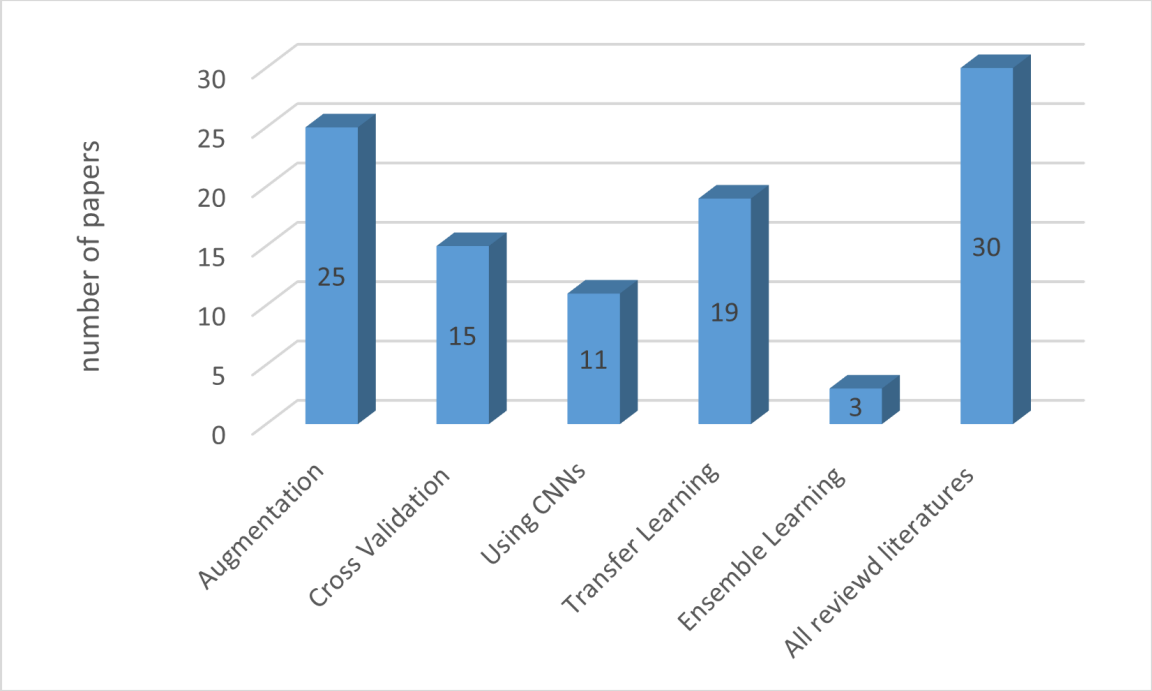

Figure 2.11: Comparing popularity rate of approaches and techniques among researchers

Data augmentation is a pre-processing method to avoid the overfitting issue was applied in 84.84% of the studies, introducing this technique as one of the best practices research

scientists can use while designing a (CADe) or (CADx). Moreover, to assure the absence of overfitting in 51.51% of the surveyed papers, cross-validation was applied. In addition, we analyzed the reviewed literature from their objective perspective, and the result is illustrated using Figure 2.13.

2.9.1.1 Investigating the pre-trained models comparison

While conducting the survey, we kept track of what pre-trained CNN architectures were used by researchers in three different categories. The first category referred to those who had only used the architecture and had not to employ the pre-trained weights. The second one was the works, and the third one was the ensemble learning. The results are illustrated using the Figure 2.12.

It is interesting to note that, by 2017, the pre-trained CNN model designeres presented more light-weighted architectures with fewer parameters such as MobileNet and EfficientNet. Nevertheless, these models were not employed as much for breast cancer classification, and they were, to our limited knowledge, never used for ensemble learning. For example, one study that involved examining newer models such as DesneNet and EfficientNet about the breast cancer problem was the research conducted by Wang et al. [149] in 2020. They classified lymph node metastases from histopathological images using EfficientNet-b3, ResNet-50, DenseNet-121, and they finally proposed their boosted EfficientNet approach and demonstrated the superiority of their model. Another example can be the employment of DenseNet-169 and EfficientNet-B5 on private mammogram datasets in the Hallym University Sacred Heart Hospital [134].

According to Figure 2.12, not all the state-of-the-art deep learning models were evaluated. For example, in many studies, AlexNet, Inception-v3, VGGNet, ResNet were explored, while only two in one study the application of MobileNet was evaluated [49], and DenseNet and EfficientNet were remained unexplored in breast mammogram classification using publicly available datasets after 2015.

2.9.1.2 Objective of the surveyed papers

The survey reveals another fact concerning each mammogram classification study's objective. Most studies were aimed to diagnose the pathology using the BI-RADS score (benign or malignant images). Such studies account for 70% of all the reviewed literature. Nonetheless, the purpose of 20% of the surveyed literature was to detect an abnormality; for some of them, the abnormalities were only mass lesions, and for the others, they also included

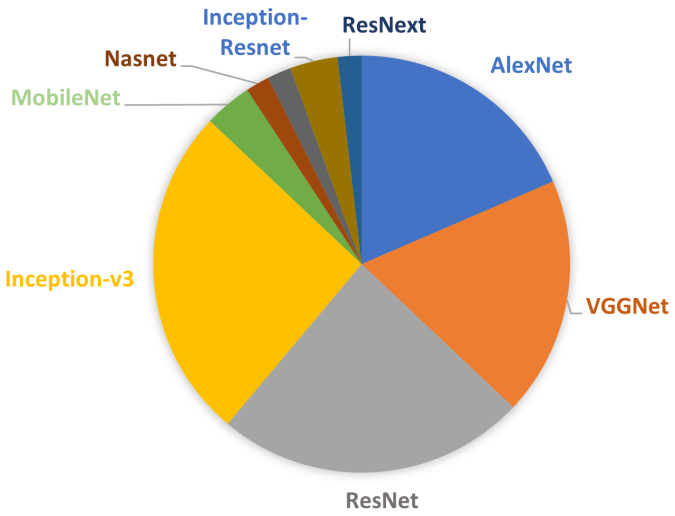

Figure 2.12: Distribution of pre-trained models employed by researchers

calcifications (minor deposits of calcium). Ironically, only a few studies were aimed for both detection and diagnosis; these studies make up for 10% of the studies in Figure 2.13.

2.9.1.3 Datasets employed

Using publicly available datasets, this survey on mammogram-based CAD revealed that 46.66% of works had employed the DDSM dataset for their classification tasks, which accounts for 14 of them out of 30. Figure 2.14 shows the distribution of datasets applied in this context. Although BCDR and INbreast datasets provide scientists with high-resolution images of FFDM, the number of instances is considered low, making them a relatively poor resource to be applied for training classifiers individually. Therefore, these datasets are often employed in combination with other potentially larger resources such as the DDSM or CBIS-DDSM datasets; the experiments conducted by Alkhaleefah et al. [17], and Agarwal et al. [11] can be good exemplars of this fact. Having been released earlier than the others, the DDSM dataset has been chosen by many researchers. Nevertheless, its application has become less popular in the past few years, and we associate the emergence of the CBIS-DDSM dataset with that.

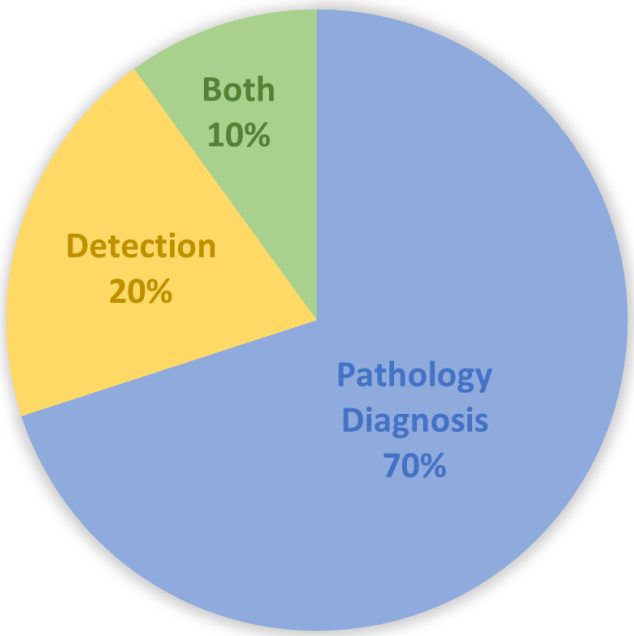

Figure 2.13: Comparing the objective's popularity rate for the researchers

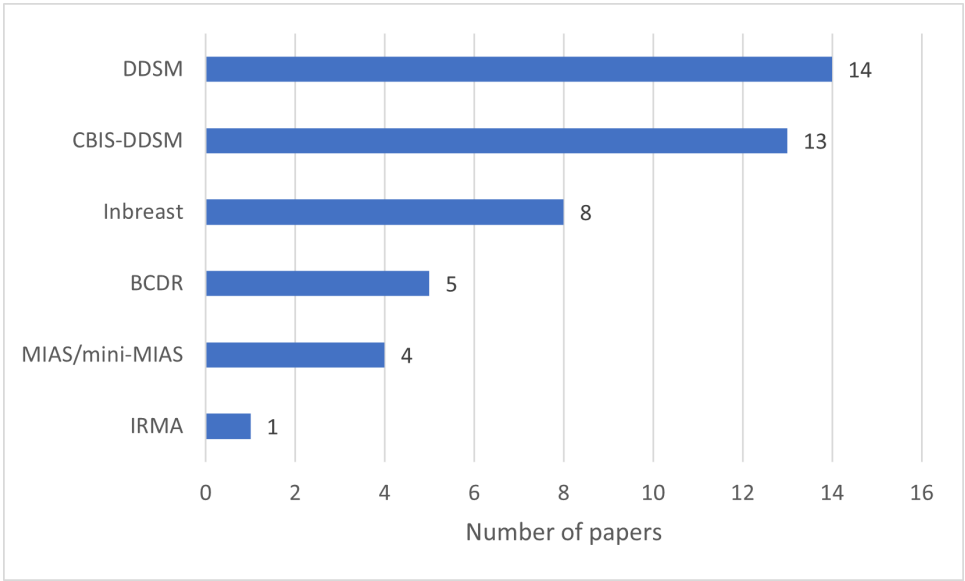

Figure 2.14: Distribution of datasets in the studies reviewed in this survey

Chapter 3

Methodology

3.1 Overview

In this chapter, we present the basic concepts behind each part of a (CADx) system 3.2. Next, our proposed model is described using visualizations 3.3, and finally, we introduce the evaluation metrics that are employed to evaluate our results 3.4.

This work seeks to combine data augmentation, transfer learning, and ensemble learning to provide a whole abnormality diagnosis system (CADx) that could assist radiologists in the healthcare system. Several experiments were conducted over the course of this work by implementing various fine-tuned CNN-based architecture and concatenating them with different strategies.

Previously, in Section 2.5.2, it was proposed that transfer learning can be significant when the target data is not large, such as in mammogram datasets. Nevertheless, some of the state-of-the-art pre-trained models were unexplored and only had a few studies performed further steps such as model concatenation or ensemble learning. Therefore, we systematically performed several rounds of the experiment to assess the accuracy of 11 fine-tuned pre-trained models from 8 different families in solving two problems known as abnormality classification and pathology classification.

3.2 Our Approach

The mammogram images can be looked at from different perspectives. To put it differently, detecting the type of an abnormality and diagnosing its severity are two distinct and

critical problems. By proposing this work of research, they have both been addressed in our research, which is only performed by a few people [105]. This section elaborates on each problem in more detail.

3.2.1 Classification of Abnormality

The first purpose of this empirical study was to evaluate the performance of pre-trained models individually in breast abnormality classification. To this end, various types of ImageNet based models were tested. These classifiers included; VGGNet-family, ResNet-family, EfficientNet-family-DenseNet-family-MobileNet-v2-Inception-v3.

From a statistical perspective [128], in the first problem, our null hypothesis is not finding anything significant, which is considered to be calcification. Therefore, when having a type I error, the null hypothesis is rejected, and we declare that something significantly different had been found, while there is nothing of significant difference. However, type II error occurs when we claim there is no difference between the null hypothesis and our achieved results when, in fact, there was. Although committing either type of error is problematic, in this case, type II error is of greater concern since failing to detect mass lesions can increase the mortality rate.

3.2.2 Classification of Pathology

From a biomedical point of view, the treatment of malignant and benign abnormalities is different from each other. In fact, the breast radiology community usually faces difficulty in mass lesion classification, which results in the unnecessary biopsy that adds expense and pressure on the patient as well as the healthcare facilities [34]. To address this concern, another solution known as Breast Pathology Diagnosis was proposed. Since in the CBIS-DDSM dataset, the pathological information is provided, data can be classified according to the severity of an abnormality.

Similar to the first problem, when statistically analyzing the errors, we propose the null hypothesis that the abnormality type is benign. The type I error refers to the problem where the mammogram is mistakenly claimed to be malignant while it is benign. Nevertheless, committing the type II error is more dangerous and results in losing the significant difference, which is not diagnosing the malignant abnormalities.

3.3 Final Proposed Model

We finally decided on having a two-staged classifier for both the abnormality classification and pathology classification parts. The two-staged architecture is depicted as follows. Figure 3.3 introduces two main stages and their objective, while Figure 3.2 shows the sub-components of the final ensemble models and the applied strategies which happen.

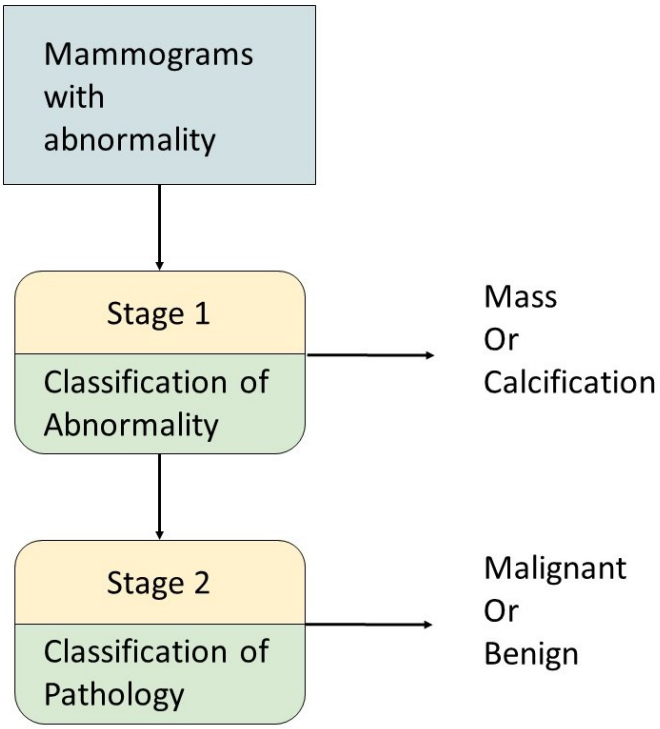

Figure 3.1: Constructing a two-staged (CADx)

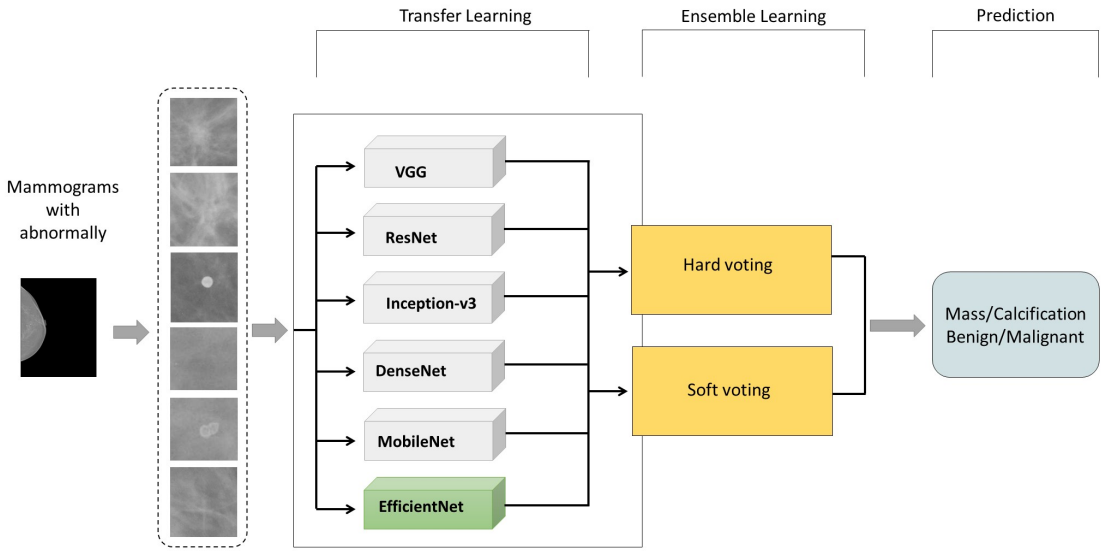

Figure 3.2: The high-level architecture behind each stage

3.3.1 Transfer Learning

The other strategy to tackle limited data sizes in the medical domain is called Transfer Learning, which is a CNN-based technique that powerfully helps the knowledge across various neural network tasks to be transferred. In this technique, a pre-trained network, which has learned features extracted from a certain image. In this case, images belong to a huge dataset known as ImageNet with 14 million images [7]. Such features are used as starting point to learn a new problem with small samples for training. The knowledge can be transferred through Feature Extraction.

3.3.2 Feature Extraction vs Fine Tuning

Feature Extraction is referred to the technique in which pre-trained deep features are extracted. Such features are the activations of one layer that, in most cases, is the last hidden

layer, or even multiple layers in an image. In the feature extraction method, the original network is not modified, and this benefits the new tasks based on the complex features that are learned using previous tasks. Nevertheless, this technique is put in opposition to fine-tuning, where the network parameters are often modified [95]. In this work, we fine-tuned all the pre-trained models by modifying their fully connected and classification layers. The output size parameter in each pre-trained model, in their fully connected and classification layer, was replaced from 1000 various objects to 2 abnormality types. The classification layer performs its task by taking the output from the softmax layer and assigning each output to k-mutually exclusive classes, using the cross-entropy loss function. The structure of the final fine-tuned layers is shown in Figure 3.3.

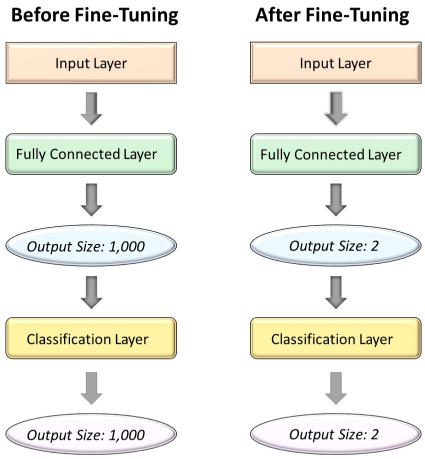

Figure 3.3: The fine-tuned architecture

3.3.3 CNN-based Pre-trained Models Participating in This Work

We implemented several types of CNN architectures for training and feature extraction in this work. Most of them have been employed in research studies related to breast cancer classification; however, some of them, like EfficientNet and MobileNet are introduced recently. Thus not many studies have utilized their power. It should be mentioned that

pre-trained ImageNet models have 1000 classes in their final layer. Nine models, pre-trained on natural images, ImageNet dataset [37] were explored for which the weights are available. Previously, the information related to each one was summarized in Table 4.3.

3.4 Evaluation Metrics

For each experiment several metrics are used to determine how promising the results are. Some of them can be employed with the same structure for both binary and multi-classification scenario such as accuracy, precision, recall. Nonetheless, other metrics like ROC curve need to be modified for multi-classification purpose.

- True positives (TP): The actual class is positive, and the predicted class is positive
- False positives (FP): The actual class is positive, but the predicted class is negative
- True negatives (TN): The actual class is negative, and the predicted class is negative
- False negatives (FN): The actual class is negative, but the predicted class is positive
- Accuracy: The accuracy of overall accuracy shows the number of the total instances that are correctly classified. In other words, here, this measure demonstrates how much of patients with mass lesions are correctly predicted or how much of the patients with calcification are correctly diagnosed. In the case of the second problem, it shows how many patients with malignant pathology are correctly classified or how many of the patients with a benign type of abnormality are correctly diagnosed.

$$ACC = \frac{(TP + TN)}{(TP + TN + FP + FN)} \tag{3.1}$$

- Sensitivity or Recall: This metric represents what number of the total positive instances are correctly classified. To put it simply, our project indicates how many of the patients with mass lesions are correctly diagnosed out of the total number of patients with mass. In the case of the second problem, it shows how much of patients with malignant pathology are correctly classified out of the total number of patients with malignant abnormality type.

$$Sn = \frac{(TP)}{(TP + FN)} \tag{3.2}$$

- Specificity: This measure indicates what number of the total negative predictions are correctly classified. Here, it demonstrates how many patients with calcification are correctly diagnosed out of the total number of patients with calcification. In the case of the second problem, it shows how much of patients with benign pathology are correctly classified out of the total number of patients with benign abnormality type.

$$Sp = \frac{(TN)}{(TN + FN)} \tag{3.3}$$

- Precision: This metric shows what number of the positive predictions are correctly classified. The measure is calculated using the formula below:

$$Pr = \frac{(TP)}{(TP + FP)} \tag{3.4}$$

- F_1-score: This measure shows the impact of both the precision and recall simultaneously using harmonic means through having more penalty in cases of extreme values. The metric can be calculated using the formula 3.5

$$F_1 - score = \frac{(2.percision \times recall)}{(percision + recall)} \tag{3.5}$$

- ROC curve (Receiver Operating Characteristic Curve): The ROC graph plots the curve of precision against the recall. The Area measured under this curve or (AUC) is a standard evaluation metric that offers an aggregate measure based on the classifier's performance at different classification thresholds. The value of the AUC score is always between one and zero, and higher value score is associated with a better classifier performance.

- Confusion Matrix: a confusion matrix consists of True Positives (TP), True Negatives (TN), False Positives (FP), and False Negatives (FN). A sample of confusion matrix is illustrated using Table 3.1.

Table 3.1: Confusion Matrix sample for the binary classification

Actual	**Predicted**		
		Negative	Positive
	Negative	True Negative	False Positive
	Positive	False Negative	True Positive

3.4.1 K-fold cross validation

Cross-validation refers to a re-sampling procedure employed to validate the classifier's performance of how well it can generalize to hidden data. In this particular type of cross-validation, the data is split to k equal or almost equal folds, then k iterations in the training and validation phase are performed. Therefore, a different data fold is held out to perform the validation task on it, which leaves k-1 folds for the training purpose [24].

Chapter 4

Results and Analysis

4.1 Overview

This chapter firstly describes the dataset that we employed in Section 4.2, later it explains the experimental settings in Section 4.3, experiments the results Section 4.4, and finally analyzes the results of implementing transfer learning and ensemble learning using the CBIS-DDSM dataset. Figure 4.1 briefly describes the steps that we took to develop a two-staged (CADx) system.

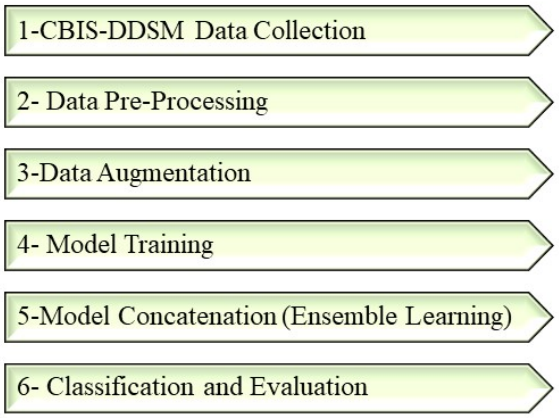

Figure 4.1: The overall development process for the proposed (CADx) model

4.2 Dataset Description

As it was earlier mentioned in Section 2.7 several publicly available datasets are available to the research community, one of the largest mammogram databases is known as Digital Database for Screening Mammography (DDSM). The DDSM dataset collected by the University of South Florida team members includes 2620 patients, which are put in 43 various volume categories. Both mammogram views, namely (CC) and (MLO), are provided using this dataset. With radiologists' assistance, the mammograms are categorized into normal, benign, and malignant using BIRADS [39].

In 2017, Lee et al. [92] introduced an updated version of the DDSM dataset is called Curated Breast Imaging Subset of DDSM or CBIS-DDSM. This dataset is used in this thesis for performance evaluation purposes. Unlike the original DDSM, this dataset does not offer normal patients and gives the abnormality types along with the pathology diagnosis and pixel-wise annotations of the (ROI)s. The images are changed into DICOM format and compressed.

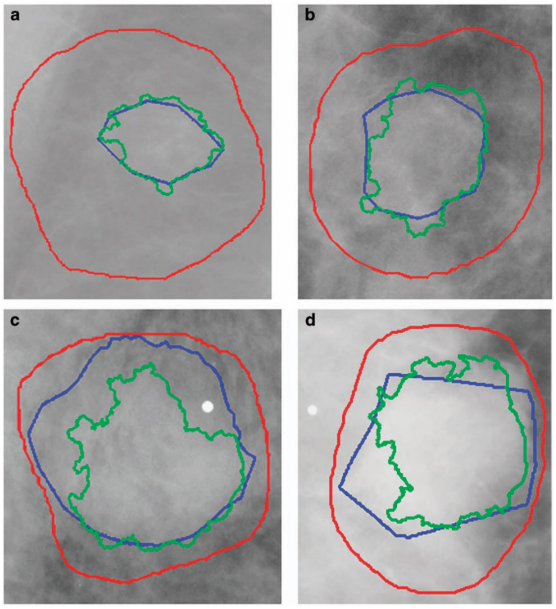

Figure 4.2: Sample of images provided in the CBIS-DDSM dataset. The red lines are depicted by the radiologists collecting DDSM, and the blue outlines were done by the other radiology experts who were collecting CBIS-DDSM. The reasearchers who were working on this dataset designed a semi-segmentation algorithm, which created the green outlines. [92]

4.2.1 Data pre-processing

The image patches were converted into NumPy arrays at the first stage, and the pixel values were normalized as they have to be compatible with the input of each model that is chosen. For example, the VGG pre-trained model works with images in (0,255). Then after shuffling the train and test split, we modified the input size. In other words, each pre-trained architecture requires a different input size; for example, inception-v3 needs the input image to be of 299×299 size, whereas the other models that were employed in our research required the input image of 224×244 the size for each image patch is 150×150, which needed to be changed using PyTorch resize function.

The training data was split into two subsets, "training" and "validation", where the training data will be only used to calculate the loss function that is being exploited by the optimizer, and the validation dataset will be used to evaluate the performance on an independent set of data while training.

4.2.2 Data Augmentation

One viable solution to this problem is image augmentation, by which the number of the training data is increased, more robust results are concluded. The other technique to mitigate the lack of sufficient data is to apply transfer learning. As it was elaborated in the literature review, many researchers have implemented image data augmentation methods. Nevertheless, some have discussed the angles of rotation, and most of them have demonstrated that using the values of 90, 180, -90 preserves the quality of the original images [34] [80]. In this research, we increased the number of training data using the TORCHVISION.TRANSFORMS function in PyTorch by flipping them both in the horizontal and vertical direction; rotation and shearing were employed as well, and the values are summarized in Table 4.1.

Table 4.1: The optimum values for augmentation

Augmentation Type	**Probability**
Rotation	50%
Horizontal Flipping	50%
Vertical Flipping	50%
Shearing	30%
Contrast	50%

4.2.3 Data Split

The distribution of masses and calcifications in the original version of the CBIS-DDSM dataset is illustrated in Figure 4.3. However, we did a further investigation to change the original splitting ratio.

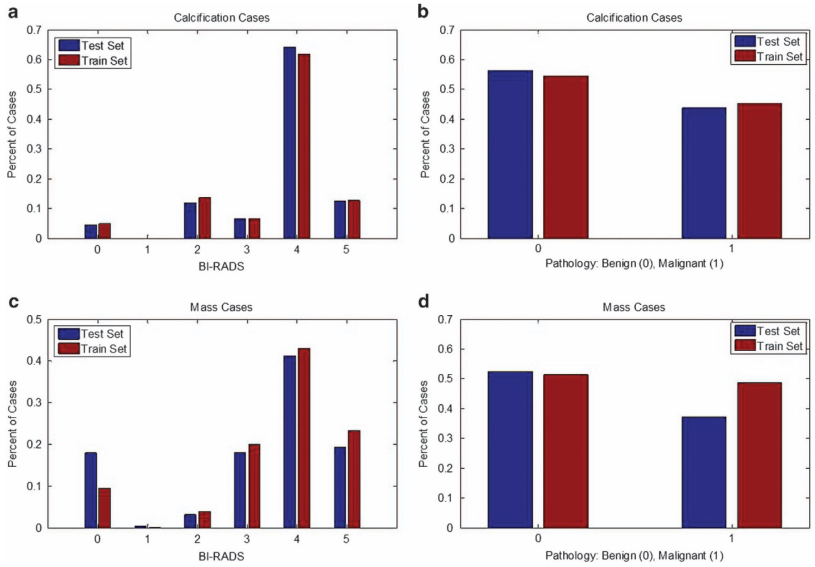

Figure 4.3: The pathology distribution in the original CBIS-DDSM dataset [92]

Table 4.2: The datasets employed in the literature

Dataset	**Data Augmentation**	**Train_Test _Validation Split**	**Classification Type**	**Purpose**
D_1	Augmentation	70%-15%-15%	Binary Classification	Mass and Calcification
D_2	Augmentation	80%-10%-10%	Binary Classification	Mass and Calcification
D_3	Augmentation	84%_8%_8%	Binary Classification	Mass and Calcification
D_4	Without Augmentation	80%-10%-10%	Binary Classification	Mass and Calcification
D_5	Augmentation	80%-10%-10%	Multi Classification	Benign\Malignant Mass Benign\Malignant Calcification
D_6	Augmentation	80%-10%-10%	Binary Classification	Benign and Malignant

The first round of experiments was conducted by testing three different train/test splitting ratios, with 70% for training and 30% for testing. Next, we carried out another round of experiments to 80% and 20%. Not surprisingly, when we increase the training data

size, the better the performance can become. The problem with large models such as VGG-19 was addressed, and generally, the models were better able to classify masses and calcifications. Thus, we went one step further and investigated another splitting ratio. We noticed that the overall accuracy tends to decrease mainly because of the small test size. We finally concluded that splitting the data to 80% for training and 20% for testing yields more promising results.

4.3 Experiments Settings

Tables 4.3 shows the training options and the hyper-parameters values used to implement the pre-trained models during the training process.

Table 4.3: Training options of the experiments

Training Options	**Configuration**
Optimizer	Adam
Mini Batch Size	32
Maximum Epochs	50
Initial Learning Rate	0.001
Execution Environment	GPU
Learning Rate Schedule	Constant

4.3.0.1 Loss Function

In a supervised neural network training task, dealing with labeled data, an objective function is required to evaluate the prediction errors that are made by the model. This function is needed during model training, and has different categories [86]:

- Binary Classification: Binary Cross-Entropy, Hinge Loss
- Multi-Class Classification: Categorical Cross-Entropy, Expectation Loss
- Regression: Euclidean Loss, Structural Similarity Measure (SSIM), l-1 Error
- Identity Verification: Contrastive Loss

Considering previous promising results achieved by other researchers in the medical imaging classifications studies [105], the binary and categorical cross entropy were used as the loss function in binary and multi-class classification problems, respectively. The formula 4.1 calculates the binary cross-entropy loss where $y\__i$ and $\hat{y}_i$ refer to the actual and predicted labels, respectively.

$$J(\omega) = \frac{1}{N}\sum_{i=1}^{N}[y_i log(\hat{y}_i) + (1\_y_i)log(1\_\hat{y}_i)] \tag{4.1}$$

Adaptive Gradient Algorithm or Adam optimizer, an extension of Stochastic Gradient Descent, was chosen, as it was previously applied on the DDSM dataset by Levy et al. [94]. All the experiments were conducted with the learning rate of 1e_1, and the number of samples in each iteration; the batch size was 32. Having a larger batch size would need more memory, which was not possible. While running extensive experiments, we noticed that owing to the limited nature of the CBIS-DDSM dataset; the training behaviour overfits nearly at epoch number 40. Therefore, we chose the number of epochs to be 50.

4.3.0.2 Execution Environment

Implementing a deep neural network system from scratch is a cumbersome and exhausting task, which is often behind the skill of researchers in the field of medical imaging. Therefore, researchers usually employ libraries and toolkits that are publicly available [10]. All the experiments were carried out on a Tesla K80 Graphical Processing Unit (GPU) from Cybera Cloud Computing Platform with a 12 GB RAM. Also, the python library used was PyTorch with CUDA support. Moreover, we used a 64-bit Ubuntu 18.04.3 as an operating system.

While training, the publicly available machine learning tool, Weight and Biases (Wandb) was employed. This tool provided the ability of real-time monitoring and facilitate performance evaluation [9].

4.4 Experimentation and Results

Recently the application of deep learning with their sub-field called CNNs has become quite popular. In fact, CNNs are potentially able to extract the objective features directly from the images, and there is no need to use manual selection, and feature extraction [46]. For

the model classifiers to have high accuracy, large-size data is needed. The small-sized data available in the medical domain usually results in a problem known as over-fitting [32].

Transfer Learning using CNN-based architectures is associated with some advantages: Firstly, using pre-trained weighted in the early layers, general features shared among all images, such as edges, are extracted. Secondly, since these models do not require all the layers to be re-trained, they help mitigate the overfitting problem. Third, the training time is reduced since only a few groups of layers at the end of the network require training.

With the help of the ImageNet dataset and the ILSVRC contest, many of the the-state-of-the-art models have been trained on natural images and proposed to the research community [37]. As it was previously elaborated in the literature review, one or several models have been applied in investigations related to mammograms. Nevertheless, most studies have only focused on solving one problem, either breast abnormality classification or breast pathology classification. Also, fewer experiments have been conducted to explore all the pre-trained models offered to the medical image analysis community. Therefore, this study is proposed to fill this gap by running extensive experiments with 11 different pre-trained networks, which were fine-tuned to be used in either binary or multi-class classification form. Later, the ensemble model was formed to boost performance by concatenating the best models. Particularly, we intended to evaluate the performance of networks with fewer number parameters such as MobileNet and EfficientNet, as was previously presented in Section 2.5.3.

4.4.1 Results Overveiw

We successfully improved the result of our previous architecture through model concatenation or ensemble learning. This objective was served by concatenating a wide range of pre-trained networks, including VGG-19, ResNet family, Inception-v3, DenseNet-201, and some smaller models such as MobileNet and EfficientNet family. Table 4.4 shows the overall results of our proposed models.

Table 4.4: Performance of different proposed models

Problem	**Transfer Learning**	**Accuracy**	AUC-score	**Ensemble Strategy**	**Accuracy** ($\mu \pm \sigma$)	**AUC-score** ($\mu \pm \sigma$)
Experiment 1	MobileNet-v1	0.91	0.91	Majority Voting	0.9402 ± 0.0508	0.9392
				Model Averaging	0.9602 ± 0.0377	0.9600
Experiment 2	EfficientNet-b3	0.72	0.71	Majority Voting	0.7737 ± 0.0837	0.7745
				Model Averaging	0.8571 ± 0.0827	0.8107

It should be mentioned that we tried to explore AlexNet,ShuffleNet, and NasNet. However, NasNet was no longer available for implementation, and AlexNet and ShuffleNet did

not yield promising results thus, they were not explored.

4.4.2 Experiment 1: Breast Abnormality Classification

Distinguishing tiny deposits of calcium and mass lesions from one another is a cumbersome task for radiology experts, and high error rates have been reported in the clinical domain. Thus, as our first goal, we chose to construct a classifier that is able to classify the mammogram images into masses and calcifications on the CBIS-DDSM dataset.

4.4.2.1 Experiment 1.1: Determining the Splitting Ratios in Individual Transfer Learning

Technical values that we set for the training of each individual modified model and ensemble learning models are summarized in Table 4.3, and the result of applying each pre-trained model has been proposed after. In the first round of experiments, dataset D_1 was chosen, where the train and test split were conducted to be 70 to 30 ratio as for the train and test image patches, respectively.

Since AlexNet was explored by other researchers previously, as it was elaborated in Section 2, it was not evaluated in any of the classification tasks. Thus, their results are not documented. We found that this splitting ratio may not provide the model with enough data to be trained with. This is mainly because large models such as VGG-19 and DenseNet, generally claimed to work well in medical image classification tasks, have not yielded promising results. Thus, another round of experimentation was performed in search of the optimum train/test splitting ratio. This time the test and train split was changed to 80/20, as it is provided in Table 4.2 with the dataset name D_2. It should be taken into account that the data needs to be shuffled so as to have balanced data classes after the change in the test set.

Not surprisingly, when we increase the training data size, the better the performance can become. The problem with large models such as VGG-19 was addressed, and generally, the models were better able to classify masses and calcifications; this improvement makes the results in this stage the best possible candidate for concatenation in learning through ensemble principle that will take place later.

Later, the third round of experiments was added to scrutinize the optimum point for train and split, and the results were subsequently evaluated. As for this round, the split ratio was changed to 84/16, where details of that can be found in Table 4.2 by the D_3 title.

By evaluating the final classification results, we noticed that the overall accuracy tends to decrease mainly because of the small test size.

4.4.2.2 Experiment 1.2: Determining the Effect of Data Augmentation

Regardless of training a CNN model from scratch or using the pre-trained models, data augmentation is one of the best practices considered helpful when training CNN-based models. We carried out two experiments with different scenarios to investigate this matter, one without and the other with image data augmentation. For this experiment, we employed the D_2 and D_4 datasets in Table 4.2. Figure 4.4 compares the effect of augmenting images while training EfficientNet-b0 with respect to their train_test,_validation accuracy, and train_test,_validation loss. In training neural networks having too many epochs can cause overfitting, similarly not having enough of that can lead to underfitting. Therefore, in our research, the best model was chosen by implementing the early stopping method, which stops the model training once it no longer improves on the validation dataset.

Figure 4.4: The effect of data augmentation

As it is evident in Figure 4.4, by applying data augmentation, the training process is performed in a smoother manner, and the test accuracy is generally more throughout

the experiment. Therefore, although the time needed to perform the experiment with data-augmentation was doubled compared with not applying that, we have employed this technique for all the experiments.

4.4.2.3 Experiment 1.3: Evaluation of Individual Pre-trained Models

This experiment compared the performance of 9 pre-trained CNN-based architectures using the transfer learning approach and reporting their accuracy and F_1-score. The output of every architecture was fine-tuned as it was explained in Section 3.3.2. The training behaviour for the DenseNet-201 has been illustrated in Figure 4.5 among all the pre-trained models. To choose the optimum accuracy, early stopping approach with a patience of 5 is used.

Table 4.8 summarizes the results of the first experiment. The test accuracy and F_1-score are reported in this table, and the best ones were selected for the next stage, which is ensemble learning. Although all the pre-trained models seem to provide us with promising accuracy, the MobileNet with the 91% test accuracy wins the competition. Nevertheless, to improve the current accuracy and develop our own state-of-the-art architecture, we will concatenate the models that will be reported in Tables 4.6, and 4.7.

Table 4.5: The accuracy and $F1$-Score for 80/20 Split

Model Name	**Variants**	**Optimum Accuracy**	**F1-Score**
VGGNet	VGG-19	0.83	0.84
ResNet	ResNet-50	0.91	0.91
	ResNet-121	0.87	0.87
DenseNet	DenseNet-201	0.90	0.89
Inception	Inception-v3	0.91	0.91
EfficientNet	EfficientNet-b0	0.90	0.90
	EfficientNet-b3	0.88	0.88
MobileNet	MobileNet-v1	0.91	0.91
	MobileNet-v2	0.88	0.88

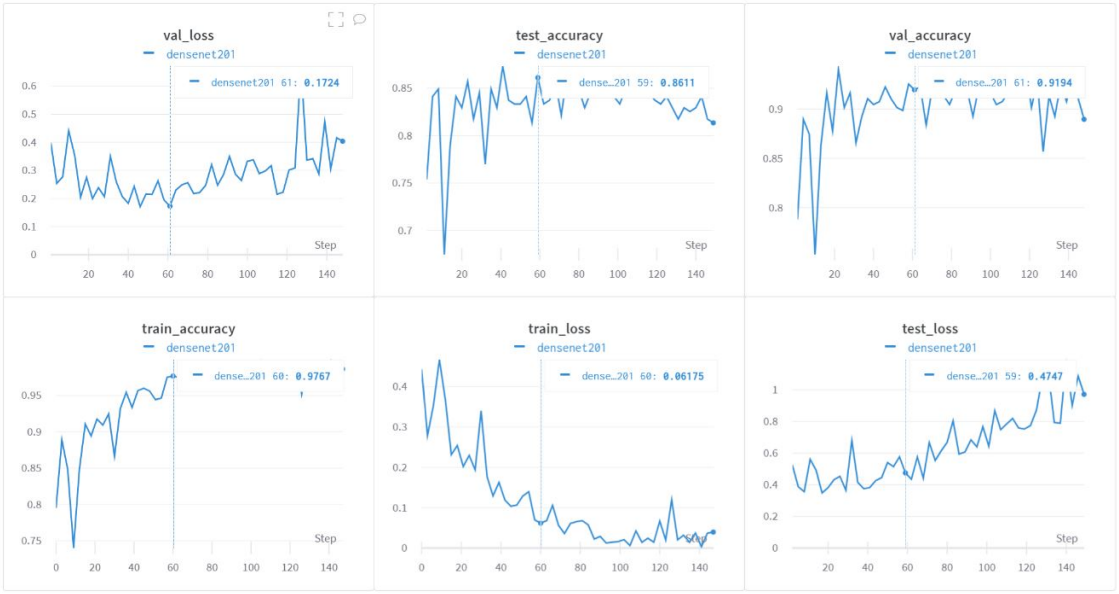

Figure 4.5: The training of DenseNet in breast abnormality classification

4.4.2.4 Experiment 1.4: Ensemble Learning for Breast Abnormality Classification

This section demonstrates the power of a technique that was implemented on top of our individual classifiers to improve the accuracy and generalization, as it was mentioned in Section 2.6. After running these three rounds of experiments, the best models were combined to propose an ensemble model. The results for the combination are provided through either concatenating all or some of the pre-trained models. Finally, the best models in each split set were concatenated in a novel manner. At first, the majority voting method was applied to report the final accuracy. The results of ensemble experiments are summarized in Table 4.6. To provide more information of how accurate the classifier works, AUC-score is calculated, and ROC plot is depicted in Figure 4.7 for all the models in the 80/20 split. Also, a confusion matrix is provided using Figure 4.6 to give a better illustration of our winner ensemble combination, ensemble2, in the hard voting scenario. It was then concluded that having all the models, VGGNet, ResNet, DenseNet, EfficientNet, MobileNet as the sub-components, outperform other concatenation forms.

Table 4.6: Ensemble models built upon hard voting

Models Combined Constructed	**Hard Voting**		
	Accuracy ($\mu \pm \sigma$)	**F1 score** ($\mu \pm \sigma$)	**AUC**
Ensemble 1 all the models VGG-19,VGG-16, ResNet-152/50, EfficientNet-b0-b3-DenseNet-201-MobileNet-v1-v2-Inception-v3	0.9291 ± 0.0436	0.9300 ± 0.0423	0.9289
Ensemble2 VGG-16, ResNet-50, EfficientNet-b0-b3-DenseNet-201-MobileNet-v1-v2-Inception-v3	**0.9402 ± 0.0508**	0.9397 ± 0.051	**0.9392**
Ensemble 3 VGG-19,VGG-16, ResNet-152, DenseNet-201, Inception-v3	0.9249 ± 0.0600	0.9247 ± 0.0604	0.9243

Figure 4.7 shows the superiority of the implemented ensemble model compared with other individual models. It clearly illustrates the more AUC_score that is present under the ROC curve for the ensemble model.

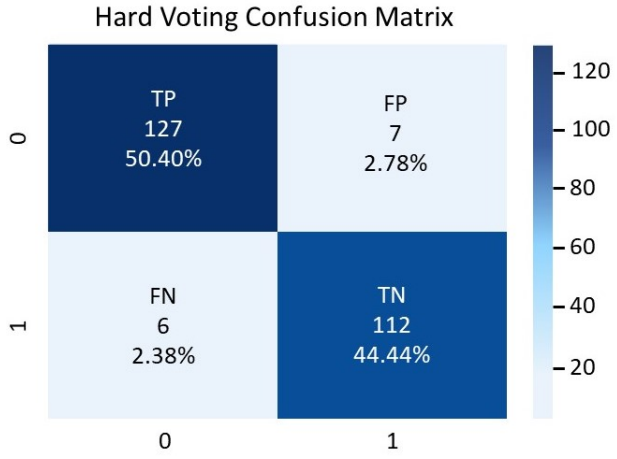

Figure 4.6: The confusion matrix for hard voting strategy in abnormality diagnosis

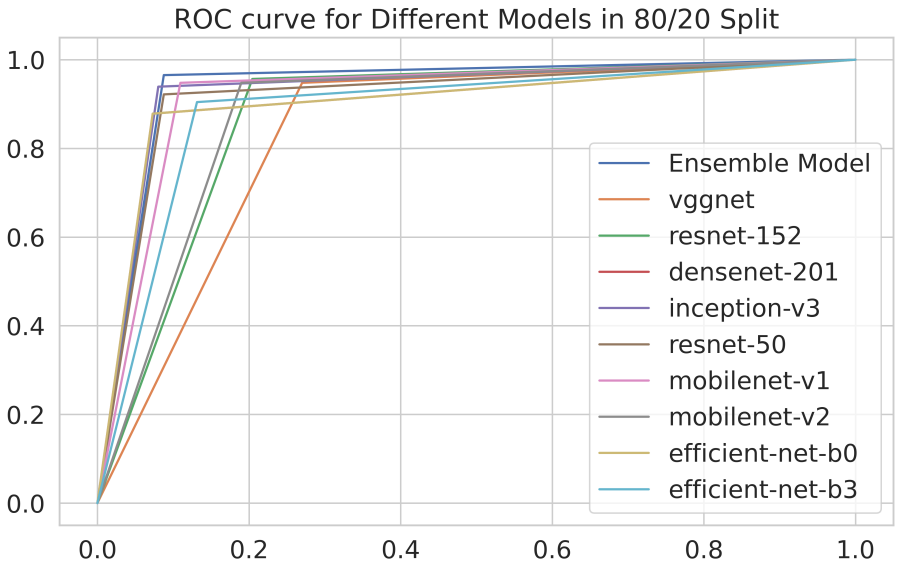

Figure 4.7: The ROC-curve for hard voting strategy in breast abnormality diagnosis (binary classifier)

The second strategy applied in the ensemble learning, was soft voting, and the results are presented in Table 4.7. Although the previously investigated hard voting technique yielded promising results, both the accuracy and AUC-score were improved by replacing the majority voting with soft voting.

Table 4.7: Ensemble models built upon soft voting

Models Combined Constructed	**Soft Voting**		
	Accuracy ($\mu \pm \sigma$)	**F1-score** ($\mu \pm \sigma$)	**AUC**
Ensemble 1 all the models VGG-19,VGG-16, ResNet-152/50, EfficientNet-b0-b3-DenseNet-201-MobileNet-v1-v2-Inception-v3	0.9329 ± 0.0366	0.9335 ± 0.0361	0.9322
Ensemble2 VGG-16, ResNet-50, EfficientNet-b0-b3-DenseNet-201-MobileNet-v1-v2-Inception-v3	**0.9602 ± 0.0377**	0.9600 ± 0.0378	**0.9600**
Ensemble 3 VGG-19,VGG-16, ResNet-152, DenseNet-201, Inception-v3	0.9208 ± 0.0703	0.9203 ± 0.0711	0.9212

Figures 4.6 and 4.8 display the confusion matrices when applying majority voting and

model averaging, respectively. When comparing these two figures, we realized a slight yet impressive reduction in the number of FP and FN responses, proving that the general accuracy was improved.

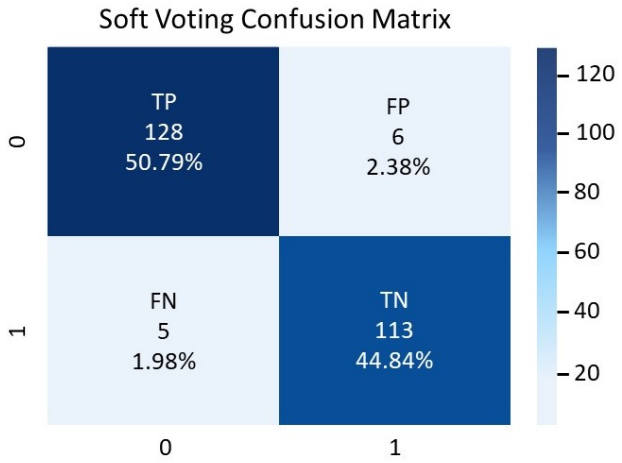

Figure 4.8: The confusion matrix for soft voting strategy in abnormality diagnosis

Figure 4.9 all the sub-components along with two best-performing ensemble models are provided. Looking at the soft and hard voting curves, we noticed that soft voting is slightly better in terms of classification.

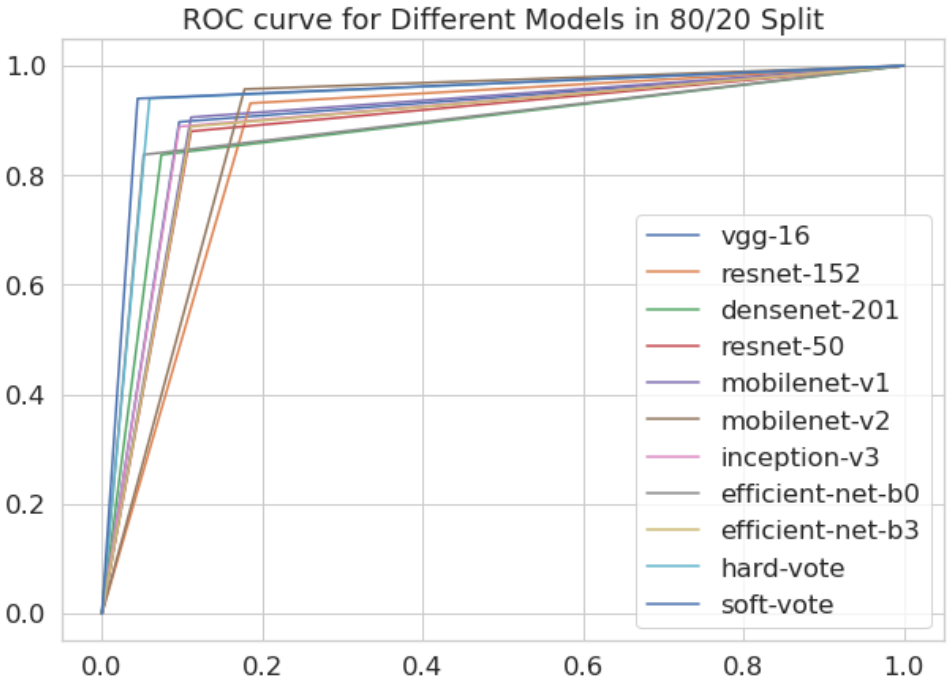

Figure 4.9: The ROC-curve for soft voting strategy in breast abnormality diagnosis (binary classifier)

4.4.3 Experiment 2: Breast Abnormality Diagnosis

According to the clinical reports, only 15% to 30% of biopsies end up being malignant [129]. In addition to the discomfort, anxiety and the chance of panic attack for the patient, this creates a burden on patients and the healthcare system [34]. Thus, breast abnormality diagnosis aims to provide the radiologist with a second opinion on how to sever a mammogram abnormality.

4.4.3.1 Experiment 2.1: Performing Multi-classification

The training procedure for the DenseNet-201 has been illustrated in Figure 4.17, like the previous stages, early stopping with a patience of 5 was employed to choose the best epoch. For this round of experiments, we used dataset D_5 as it was shown in Table 4.2. In a similar manner to the binary classification, the best model was selected and considered for concatenation. Although constructing such a device helps the radiologists to have a

second opinion about the pathology of an abnormality, the accuracy has dropped compared with the previous stage where the only task for the model is to classify into two categories.

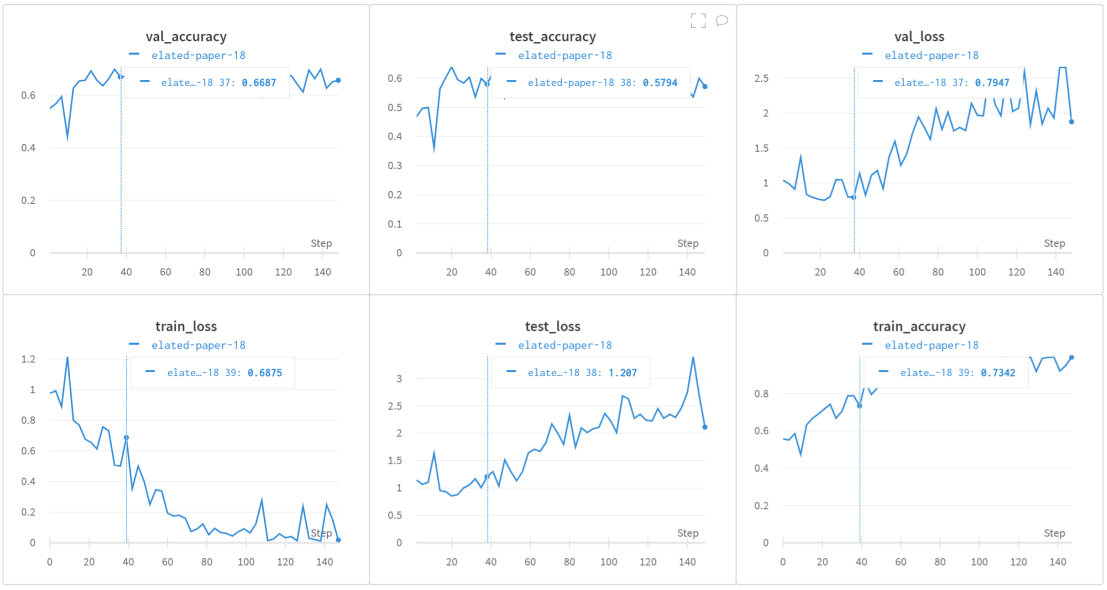

Figure 4.10: The training of DenseNet in breast abnormality diagnosis (multi classification)

Table 4.8: The accuracy and F1-Score for 80/20 split

Model Name	**Variants**	**Optimum Accuracy**	**F1-Score**
VGGNet	VGG-19	0.54	0.54
ResNet	ResNet-50	0.62	0.57
	ResNet-152	0.54	0.48
DenseNet	DenseNet-201	0.50	0.47
Inception	Inception-v3	0.58	0.53
MobileNet	MobileNet-v2	0.48	0.48
EfficientNet	EfficientNet-b0	0.71	0.71
	EfficientNet-b3	0.72	0.71

We began the experiment by fine-tuning the output of the pre-trained architectures to four classes and performing transfer learning individually to categorize the image patches in

four classes. Similar to the previous stage, we performed the two strategies in the ensemble learning (hard_soft voting) as is presented in Table 4.9.

Table 4.9: Multi-classification ensemble models built upon soft/hard voting

Models Combined Constructed	**Hard Voting**		**Soft Voting**	
	Accuracy	**F1 Score**	**Accuracy**	**F1-score**
Ensemble 1 all the models VGG-19, ResNet-152/50, EfficientNet-b0-b3- DenseNet-201-MobileNet-Inception-v3	0.64	0.64	0.66	0.64
Ensemble 2 VGG-19, ResNet-50, EfficientNet-b0-b3	0.65	0.61	0.69	0.68
Ensemble 3 VGG-19, ResNet-152, DenseNet-201, Inception-v3	0.61	0.54	0.59	0.54

As Table 4.17 represents, the ensemble2 that was formed by the concatenation of ResNet50/152, VGG-19 and the EfficientNet family as its sub-architectures yielded the best accuracy results, 65% and 69% for hard and soft voting, respectively.

Figures 4.11, 4.12, 4.13, and 4.14 illustrate the performance of the ensemble model compared with some of the best performing pre-trained architectures on this problem. As it is observed from the curves, the ensemble model in each class outperforms the others and encompasses a larger area under it. This accuracy is enhanced when soft voting is replaced with hard voting. Nevertheless, it is not still a promising performance for a (CADx) system.

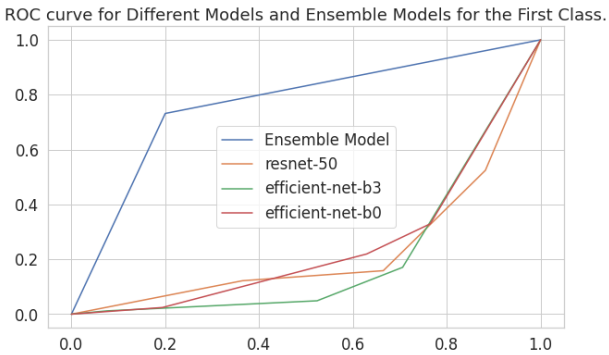

Figure 4.11: The ROC curve of ensemble models and its sub-component in the first class

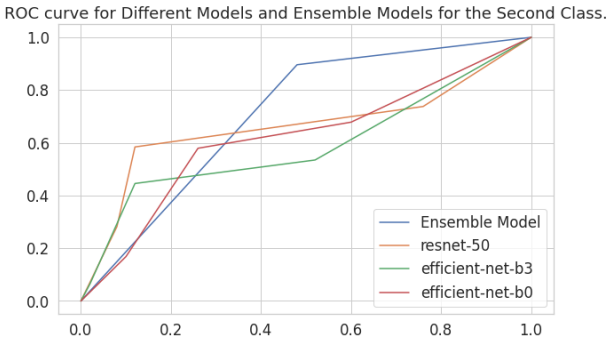

Figure 4.12: The ROC curve of ensemble models and its sub-component in the second class

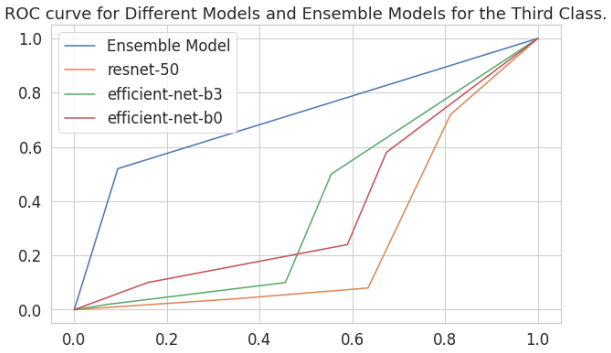

Figure 4.13: The ROC curve of ensemble models and its sub-component in the third class

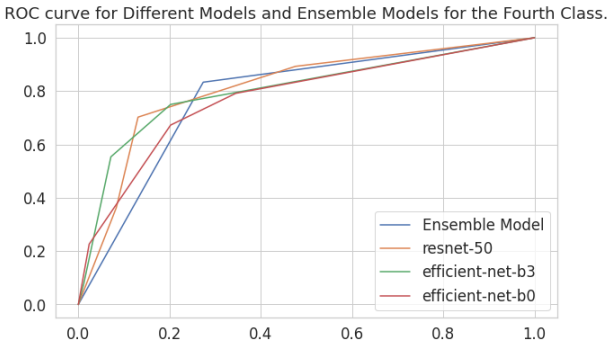

Figure 4.14: The ROC curve of ensemble models and its sub-component in the fourth class

Figure 4.15 and 4.16 show the classification matrices for hard and soft voting scenarios,

after model concatenation. As it is perceived from looking at the numbers located on the main diagonal of each confusion matrix, the correctly predicted samples, the soft voting outperforms hard voting.

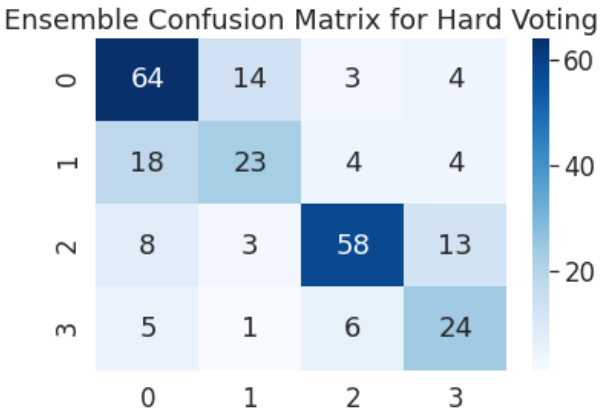

Figure 4.15: The confusion matrix for hard voting strategy (multi-classification)

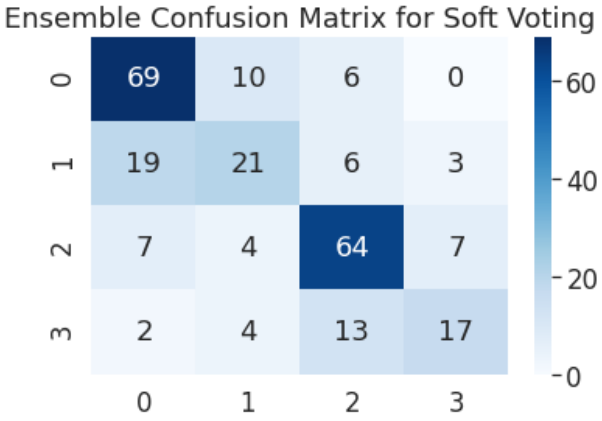

Figure 4.16: The confusion matrix for soft voting strategy (multi-classification)

After looking at the result of this experiment, we noticed that although having a multi-classifier provides the radiologists with better information about the pathology of an abnormality, it is not accurate in terms of solving the first problem (classifying masses and calcifications). Therefore, we were inspired by the research that was previously conducted by Khan et al. [105] in which they used a multi-staged structure to create a whole Breast

Cancer Diagnosis (CADx) system. In fact, they proposed the idea of having two independent, which in our case is ensemble independent together of classifying type of the disease (abnormality) and assess the severity level (pathology).

4.4.3.2 Experiment 2.2: Constructing the Second Stage

The aim of the first experiment carried out in this research was to perform binary classification on this dataset, meaning to utilize the power of transfer learning and ensemble learning to distinguish between mass and calcification mammograms. Using pre-trained models and majority voting, the AUC-score of 0.9388 and accuracy of 0.94, was received. After applying soft voting to the same structure, the accuracy and the AUC-score were improved nearly by 2%.

While performing the second round of the experiment, not only our proposed model is able to distinguish between masses and calcification mammograms, but it also can diagnose whether the particular abnormality is benign or malignant. Nevertheless, we noticed a decrease in the accuracy of a classifier in the case of multi-classification. Consequently, in the third round of experiments, we applied individual pre-trained models to binary classify the dataset into malignant and benign mammograms; we applied the D_6 dataset for as it was introduced in Table 4.2.

Figure 4.17 represents the behaviour of DenseNet-201 architecture while being trained for breast abnormality diagnosis, where using early stropping with a patience of 5 the best epoch was selected.

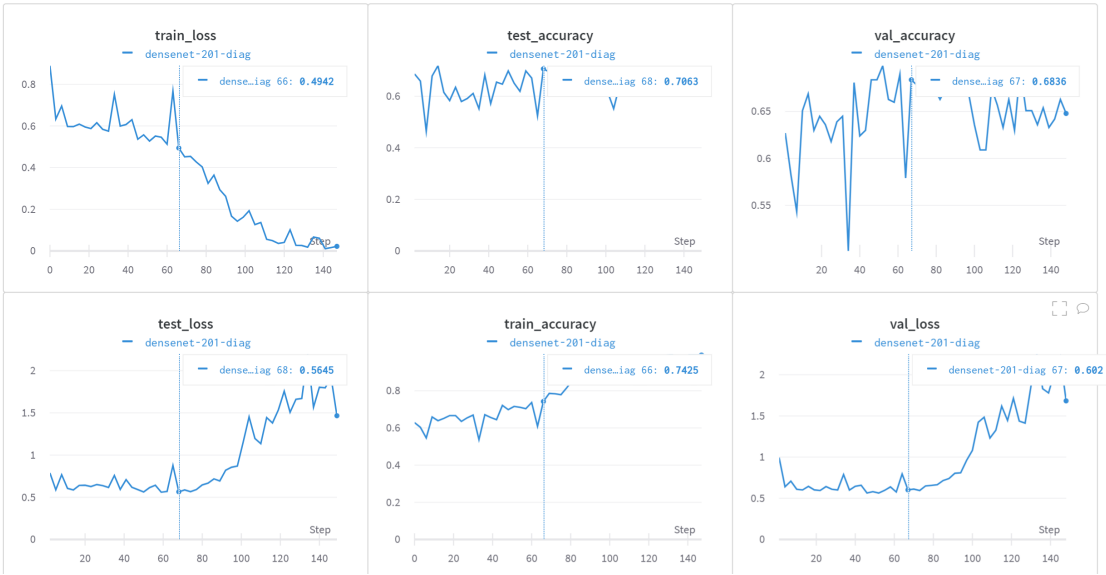

Figure 4.17: The training of DenseNet in breast pathology diagnosis (binary classification)

Table 4.10 represents the result of each model on the CBIS-DDSM dataset individually, where EfficientNet-b3 wins the competition with 72% as accuracy. Considering the performance of light-weighted models such as MobileNet and EfficientNet family, they are among the main potential main building blocks for the ensemble model.

Table 4.10: The accuracy and F1-Score for 80/20 Split

Model Name	**Variants**	**Optimum Accuracy**	**F1-Score**
VGGNet	VGG-19	0.54	0.54
ResNet	ResNet-50	0.67	0.68
	ResNet-152	0.64	0.63
DenseNet	DenseNet-201	0.68	0.67
Inception	Inception-v3	0.58	0.53
MobileNet	MobileNet-v2	0.70	0.69
EfficientNet	EfficientNet-b0	0.71	0.71
	EfficientNet-b3	0.72	0.71

4.4.3.3 Experiment 2.3: Ensemble Learning for Breast Pathology Classification

A more robust model is built upon all or some of the pre-trained models. Three ensemble models were constructed and evaluated, one resulting from combining all or some of the pre-trained models that were elaborated in the methodology section. The proposed model built upon light-weighted models outperformed the others, and not surprisingly, the models that only used them as their main ensemble components performed better. The reason for that can be the small size of the models such as MobileNet and EfficientNet. To be more specific, due to their smaller number of parameters, they perform compatible with our needs in the medical domain. The other experiment that was conducted by the research was the effect of applying soft voting instead of hard voting or majority voting. Our results were further validated through 10-fold cross-validation and reported in Tables 4.11 for hard voting strategy, and 4.12 for soft voting strategy.

Table 4.11: Breast abnormality diagnosis built upon hard voting

Models Combined Constructed	**Hard Voting**		
	Accuracy ($\mu \pm \sigma$)	**F1 score** ($\mu \pm \sigma$)	**AUC**
Ensemble 1 all the models VGG-19,VGG-16, ResNet-152/50, EfficientNet-b0-b3-DenseNet-201-MobileNet-v1-v2-Inception-v3	0.7420 ± 0.0574	0.7706 ± 0.0506	0.7345
Ensemble2 VGG-16, ResNet-50, EfficientNet-b0-b3-DenseNet-201-MobileNet-v1-v2-Inception-v3	**0.7737 ± 0.0837**	0.7864 ± 0.0724	**0.7745**
Ensemble 3 VGG-19,VGG-16, ResNet-152, DenseNet-201, Inception-v3	0.6785 ± 0.1109	0.7239 ± 0.1009	0.7243

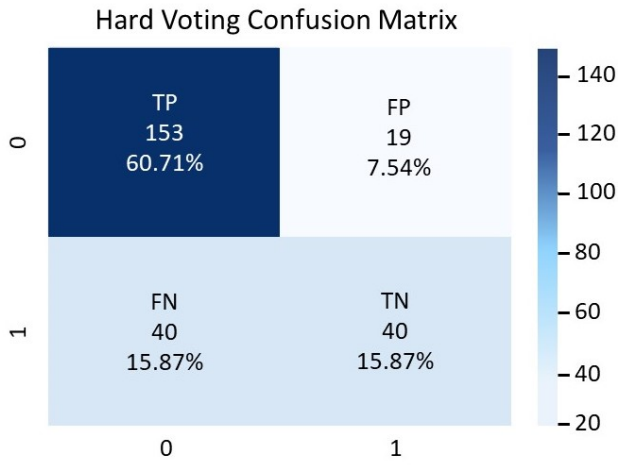

Figure 4.18: The confusion matrix for hard voting strategy in breast cancer diagnosis

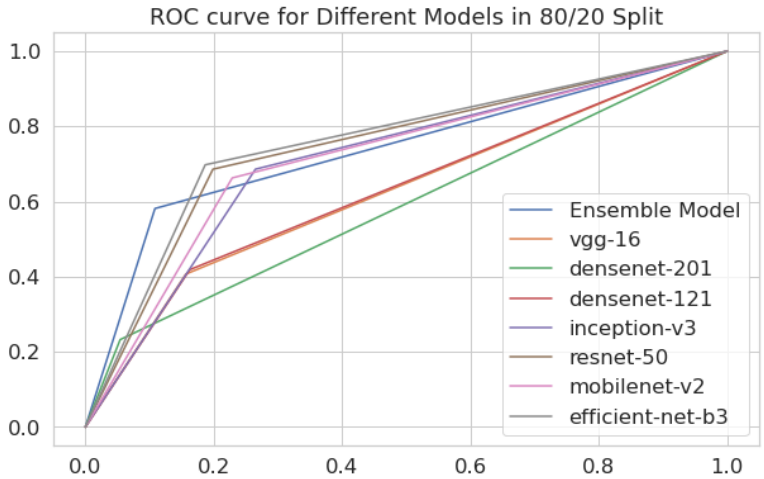

Figure 4.19: The ROC-curve for hard voting strategy in breast pathology diagnosis (binary classifier)

Then the results of the majority voting technique among ensemble learning strategies were compared with another strategy known as soft voting and subsequently summarized in Table 4.7.

Table 4.12: Breast abnormality diagnosis built upon soft voting

Models Combined Constructed	**Soft Voting**		
	Accuracy ($\mu \pm \sigma$)	**F1 score** ($\mu \pm \sigma$)	**AUC**
Ensemble 1 all the models VGG-19,VGG-16, ResNet-152/50, EfficientNet-b0-b3-DenseNet-201-MobileNet-v1-v2-Inception-v3	0.8571 ± 0.0827	0.8421 ± 0.0508	0.8107
Ensemble2 VGG-16, ResNet-50, EfficientNet-b0-b3-DenseNet-201-MobileNet-v1-v2-Inception-v3	**0.8571 ± 0.0827**	0.8277 ± 0.0447	**0.8107**
Ensemble 3 VGG-19,VGG-16, ResNet-152, DenseNet-201, Inception-v3	0.7062 ± 0.1154	0.7239 ± 0.1009	0.7144

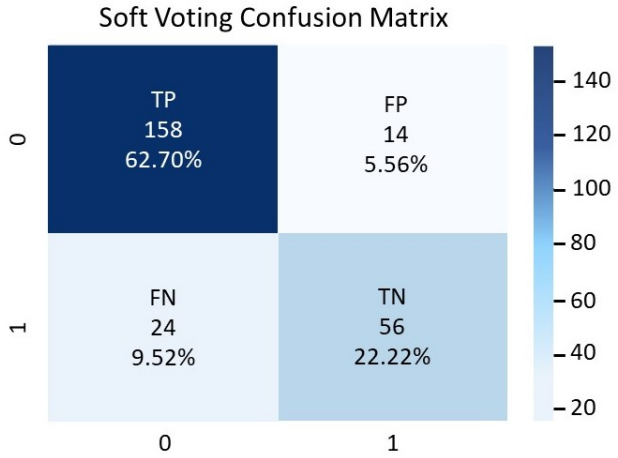

Figure 4.20: The confusion matrix for soft voting strategy in breast pathology diagnosis

By comparing the confusion matrices 4.18 and 4.20, we noticed that not only does the soft voting result in better performance in terms of accuracy, but it also considerably reduces the number of FNs, from 40 to 24 image patches. This is highly important in the medical field since patients whose abnormality is falsely considered benign are more prone to death than those incorrectly told to have malignant pathology.

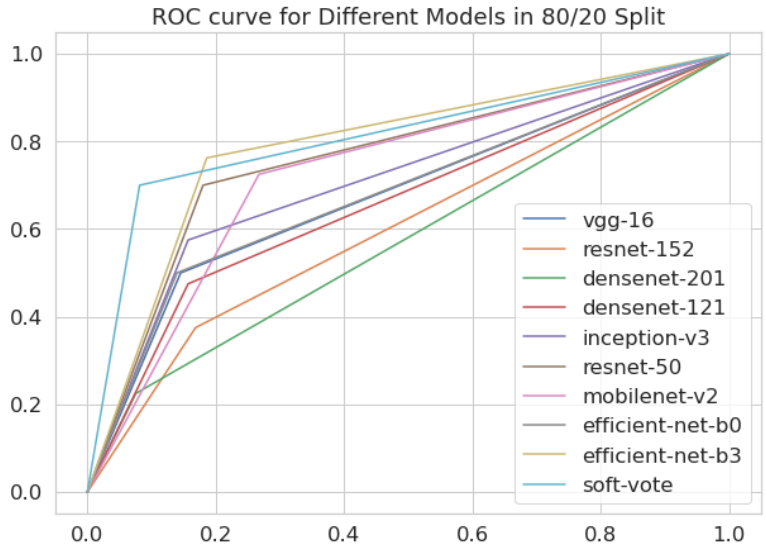

Figure 4.21: The ROC-curve for soft voting strategy in breast pathology diagnosis (binary classifier)

Figure 4.21 shows the ROC curve in breast abnormality diagnosis problem, and it clearly demonstrates that for soft voting the area under this curve has a higher value compared with any individual classifier beside it. Using 10-fold cross-validation technique, the numerical results from implementing the best combinations of ensemble learning are summarized in Table 4.13.

Problem	**Ensemble Learning**	**Accuracy** ($\mu \pm \sigma$)	**AUC-score** ($\mu \pm \sigma$)
Abnormality Classification	Model Averaging	0.9602 ± 0.0377	0.9600
Pathology Diagnosis	(soft voting)	0.8571 ± 0.0827	0.8107

Table 4.13

4.4.4 Comparison with the Previous Results

Our results are compared with the results of previous works that involved CBIS-DDSM classification, in Table 4.14. As this table presents our two-staged classifier, inspired by the three-staged architecture that was proposed by Khan et al. [105] outperforms the other studies in the solving the first problem, abnormality classification. Yet, in the second stage, the pathology diagnosis for the abnormalities, we realized that the results that

Arora et al. [20] surpasses our results, nevertheless, since their study does not include cross-validation, we assume that their results could have been biased. By this we mean, when running an experiment only once the model could be tested on some data that is very similar to the data that our model has seen before.

Table 4.14: Summary of the previous results in related works

Author	**Transfer Learning**	**Ensemble Learning**	**Abnormality Classification**	**Pathology Classification**	**Augmen tation**	**Cross Validation**
L.Falconi et al.	VGG,ResNet, Resnext, Xception	-	-	0.84	yes	no
R.Arora et al.	VGG, ResNet, Inception-v3 AlexNet, InceptionResNet	Stacking	-	0.88	yes	no
H.Khan et al.	VGG,ResNet, Inception-v3	-	92.29 ± 1.15%	77.66± 0.62%	yes	averaging
Our proposed model	VGG, ResNet, inception-v3, DenseNet, MobileNet, EfficienNet	Weighted Averaging	0.96 ± 0.03	0.85 ± 0.08	yes	yes

4.5 Discussion and Limitation

It is an undeniable fact that by replacing the multi-classifier with another stage of binary classification more time would be requires, as this problem was previously proposed by Khan et al. [105] with the title of time-complexity. However, the time required is mostly for training, and in the opinion of these authors, it is completely justified to invest more time in constructing a more complex architecture, to reduce the false negative or false positive rate in the clinical enterprise.

While evaluating the effect of time, two different terms were used, training time and inference time. As the name suggests, training time refers to the time consumed for training purposes. Also, according to a study performed by Canziani et al. [29], VGG variants have the highest inference time when the batch size is unchanged, which makes them a less suitable candidate for real-time applications. The reference [66] conducted a study under constrained time to solve the issue of time complexity and consequently understand the effect of CNN factors such as depth, width, number of filters, filter size, etc.

With regard to the model training stage, the RAM of our GPU was not enough to enable implementing the other variant of EfficientNet family, EffientNet-B7, and even with lowering the batch-size we were not able to implement them. Nevertheless, based on other literatures and the current result we have achieved from EfficientNet-b0 and EfficientNet-b7 we assume that this particular model would not improve our results.

Since this thesis employs a pretty balanced dataset, using an unsupervised method to investigate how these samples can form a different cluster. In other words, merely knowing how the data is distributed and using it as an additional input can improve the results.

Chapter 5

Conclusion and Future Works

The conclusions of our research are summarized using the bullet points below:

- Data Augmentation or increasing the number of training images, was proved to be helpful, even though, it is associated with longer time for training.
- The experimental results suggest that fine-tuning pre-trained on an individual level can lead to promising results, using EfficientNet with 91% and 72% for abnormality and pathology classification, respectively. The fine-tuning applied at this stage did not include significant changes in the architecture, and only the number of the output layers were altered. Additionally, stronger and more robust classifiers can be accomplished using ensemble learning.
- Comparing the training behaviour of different pre-trained models while performing transfer learning, we understood that due to the relatively small size of the CBIS-DDSM dataset, models with a smaller number of parameters (EfficientNet, MobileNet, and ResNet-50) tend to produce better classification accuracy. That is to say, since the mentioned models have a fewer number of trainable parameters, they tend to yield better classification results. Thus their concatenation would provide a better model in solving both problems.
- For both problems, weighted averaging or soft voting tends to be a better strategy because it determines the final vote based on the average of probabilities. rather than just a binary input from each classifier, soft voting takes into account how confident that specific model is. In this work in particular, this ensemble strategy has helped to reduce the number of False Negatives or Type II error The achieved performance,

96% for abnormality type classification, and 85% for abnormality pathology diagnosis are very promising and can be employed in real-life clinical enterprises provide the radiologists with decision support. The results for each and every stage are presented in Table 4.13.

Future Works Owing to the limited nature of time and computational resources, some ideas were not investigated and have been summarized to put a corner stone for future studies:

- The dataset that was employed in this research did not contain normal image patches, and that is the main reason why it is called Breast Abnormality Classifier and Breast Pathology Classifier (CADx). Future version of this study can include normal data so as to develop Breast Cancer Detection (CADe) and Breast Cancer Diagnosis (CADx) systems.

- Some studies have done more modifications in terms of fine-tuning and have re-trained some Convolutional Layers in each pre-trained architectures. Yet, as it was mentioned by Arora et el. [20] we only fine-tuned the last layer and kept the rest of the architecture same to avoid overfitting. Our results were already promising, however, in future versions of this research we can try their approach to check any particular change in the accuracy or the training time would happen. Specifically, for the second problem, abnormality diagnosis.

- We demonstrated that the model averaging or soft voting approach is more effective to boost the classification accuracy. However, the weights by which each probability is summed are equal, which can be replaced more efficiently. In fact, future works can include adding another model or what is known to be a meta-model to best combine the individual classifiers; this method is usually referred to as stacking. Moreover, considering that some models might be superior in terms of TP responses and the others might yield more promising TN responses, a modified stacking. This novel aggregate model can potentially advance the results to a considerable extent, a similar approach to stacking.

References

[1] Breast Cancer Digital Repository, howpublished = https://bcdr.eu/., note = Accessed: 2010-09-30,.

[2] Info @ Www.Radiologyinfo.Org." 2015,, howpublished = http://www.radiologyinfo.org/en/info.cfm?pg=safety-xray., note = Accessed: 2010-09-30.

[3] MS Windows NT kernel description. https://www.acr.org/Clinical-Resources/Reporting-and-Data-Systems/Bi-Rads. Accessed: 2010-09-30.

[4] MS Windows NT kernel description. https://www.docpanel.com/blog/post/practical-guide-understanding-bi-rads. Accessed: 2010-09-30.

[5] MS Windows NT kernel description. http://web.archive.org/web/20080207010024/\http://www.808multimedia.com/winnt/kernel.htm. Accessed: 2010-09-30.

[6] Tomosynthesis Outperforms Digital Mammography in Five-Year Study, howpublished = https://www.rsna.org/news/2020/march/tomosynthesis-versus-digital-mammography, note = Accessed: 2010-09-30.

[7] Transfer Learning. https://www.image-net.org/index.php. Accessed: 2010-09-30.

[8] Types of Mammograms | SSM Health." [Online], howpublished = https://www.ssmhealth.com/womens-health/breast-health/mammogram-types., note = Accessed: 2010-09-30,.

[9] Weight & Biases Tool for Machine Learning. https://docs.wandb.ai/. Accessed: 2010-09-30.

[10] Dina Abdelhafiz, Clifford Yang, Reda Ammar, and S. Nabavi. Deep convolutional neural networks for mammography: advances, challenges and applications. *BMC Bioinformatics*, 20, 2019.

[11] Richa Agarwal, Oliver Diaz, Xavier Lladó, Moi Hoon Yap, and Robert Martí. Automatic mass detection in mammograms using deep convolutional neural networks. *Journal of Medical Imaging*, 6(3):1 – 9, 2019.

[12] Al Hussein Ahmed and Mohammed A.-M. Salem. Mammogram-based cancer detection using deep convolutional neural networks. In *2018 13th International Conference on Computer Engineering and Systems (ICCES)*, pages 694–699, 2018.

[13] Mugahed A Al-Antari, Mohammed A Al-Masni, and Tae-Seong Kim. Deep learning computer-aided diagnosis for breast lesion in digital mammogram. *Deep Learning in Medical Image Analysis*, pages 59–72, 2020.

[14] Mugahed A. Al-antari and Tae-Seong Kim. Evaluation of deep learning detection and classification towards computer-aided diagnosis of breast lesions in digital x-ray mammograms. *Computer Methods and Programs in Biomedicine*, 196:105584, 06 2020.

[15] Mohammed A. Al-masni, Mugahed A. Al-antari, Jeong-Min Park, Geon Gi, Tae-Yeon Kim, Patricio Rivera, Edwin Valarezo, Mun-Taek Choi, Seung-Moo Han, and Tae-Seong Kim. Simultaneous detection and classification of breast masses in digital mammograms via a deep learning yolo-based cad system. *Computer Methods and Programs in Biomedicine*, 157:85–94, 2018.

[16] Abdulkadir Albayrak and Gokhan Bilgin. Mitosis detection using convolutional neural network based features. In *2016 IEEE 17th International Symposium on Computational Intelligence and Informatics (CINTI)*, pages 000335–000340, 2016.

[17] Mohammad Alkhaleefah, Shang-Chih Ma, Yang-Lang Chang, Bormin Huang, Praveen Kumar, and Vishnu Achhannagari. Double-shot transfer learning for breast cancer classification from x-ray images. *Applied Sciences*, 10:3999, 06 2020.

[18] Md. Zahangir Alom, Theus Aspiras, Tarek Taha, and Vijayan Asari. Histopathological image classification with deep convolutional neural networks. page 30, 09 2019.

[19] John Arevalo, Fabio A González, Raúl Ramos-Pollán, Jose L Oliveira, and Miguel Angel Guevara Lopez. Convolutional neural networks for mammography mass lesion

classification. In *2015 37th Annual international conference of the IEEE engineering in medicine and biology society (EMBC)*, pages 797–800. IEEE, 2015.

[20] Ridhi Arora, Prateek Rai, and Balasubramanian Raman. Deep feature–based automatic classification of mammograms. *Medical & Biological Engineering & Computing*, 58, 03 2020.

[21] P. Baldi and Peter Sadowski. Understanding dropout. In *NIPS*, 2013.

[22] Pierre Baldi and Peter J Sadowski. Understanding dropout. In C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 26. Curran Associates, Inc., 2013.

[23] Y. Bengio. Learning deep architectures for ai. *Foundations*, 2:1–55, 01 2009.

[24] Daniel Berrar. *Cross-Validation.* 01 2018.

[25] Azzedine Boukerche and Zhijun Hou. Object detection using deep learning methods in traffic scenarios. *ACM Computing Surveys (CSUR)*, 54(2):1–35, 2021.

[26] Y-Lan Boureau, Jean Ponce, and Yann LeCun. A theoretical analysis of feature pooling in visual recognition. In *Proceedings of the 27th International Conference on International Conference on Machine Learning*, ICML'10, page 111–118, Madison, WI, USA, 2010. Omnipress.

[27] Darren R. Brenner, Hannah K. Weir, Alain A. Demers, Larry F. Ellison, Cheryl Louzado, Amanda Shaw, Donna Turner, Ryan R. Woods, and Leah M. Smith. Projected estimates of cancer in canada in 2020. *CMAJ*, 192(9):E199–E205, 2020.

[28] Tom Brosch, Roger Tam, Alzheimer's Disease Neuroimaging Initiative, et al. Manifold learning of brain mris by deep learning. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 633–640. Springer, 2013.

[29] Alfredo Canziani, Adam Paszke, and Eugenio Culurciello. An Analysis of Deep Neural Network Models for Practical Applications. *arXiv e-prints*, page arXiv:1605.07678, May 2016.

[30] Haichao Cao, Shiliang Pu, Wenming Tan, Junyan Tong, and Di Zhang. Multitasking u-shaped network for benign and malignant classification of breast masses. *IEEE Access*, 8:223396–223404, 2020.

[31] Yuanqin Chen, Qian Zhang, Yaping Wu, Bo Liu, Meiyun Wang, and Yusong Lin. Fine-tuning resnet for breast cancer classification from mammography, 2018.

[32] Travers Ching, Daniel Himmelstein, Brett Beaulieu-Jones, Alexandr Kalinin, T. Do, Gregory Way, Enrico Ferrero, Paul Agapow, Michael Zietz, Michael Hoffman, Wei Xie, Gail Rosen, Benjamin Lengerich, Johnny Israeli, Jack Lanchantin, Stephen Woloszynek, Anne Carpenter, Avanti Shrikumar, Jinbo Xu, and Casey Greene. Opportunities and obstacles for deep learning in biology and medicine. *Journal of The Royal Society Interface*, 15:20170387, 04 2018.

[33] KyungHyun Cho, Tapani Raiko, and Alexander Ilin. Enhanced gradient and adaptive learning rate for training restricted boltzmann machines. In *Proceedings of the 28th International Conference on International Conference on Machine Learning*, ICML'11, page 105–112, Madison, WI, USA, 2011. Omnipress.

[34] Hiba Chougrad, H. Zouaki, and Omar Alheyane. Convolutional neural networks for breast cancer screening: Transfer learning with exponential decay. *ArXiv*, abs/1711.10752, 2017.

[35] Emily F Conant, Samantha P Zuckerman, Elizabeth S McDonald, Susan P Weinstein, Katrina E Korhonen, Julia A Birnbaum, Jennifer D Tobey, Mitchell D Schnall, and Rebecca A Hubbard. Five consecutive years of screening with digital breast tomosynthesis: outcomes by screening year and round. *Radiology*, 295(2):285–293, 2020.

[36] Jack D Cowan. Discussion: Mcculloch-pitts and related neural nets from 1943 to 1989. *Bulletin of mathematical biology*, 52(1-2):73–97, 1990.

[37] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255, 2009.

[38] Li Deng. A tutorial survey of architectures, algorithms, and applications for deep learning. *APSIPA Transactions on Signal and Information Processing*, 3, 2014.

[39] Neeraj Dhungel, Gustavo Carneiro, and Andrew P. Bradley. A deep learning approach for the analysis of masses in mammograms with minimal user intervention. *Medical Image Analysis*, 37:114–128, 2017.

[40] Neeraj Dhungel, Gustavo Carneiro, and Andrew P. Bradley. Fully automated classification of mammograms using deep residual neural networks. In *2017 IEEE 14th International Symposium on Biomedical Imaging (ISBI 2017)*, pages 310–314, 2017.

[41] Thomas G. Dietterich. Ensemble methods in machine learning. In *Multiple Classifier Systems*, pages 1–15, Berlin, Heidelberg, 2000. Springer Berlin Heidelberg.

[42] Andrea Duggento, Marco Aiello, Carlo Cavaliere, Giuseppe L Cascella, Davide Cascella, Giovanni Conte, Maria Guerrisi, and Nicola Toschi. An ad hoc random initialization deep neural network architecture for discriminating malignant breast cancer lesions in mammographic images. *Contrast media & molecular imaging*, 2019, 2019.

[43] Aarthipoornima Elangovan and T. Jeyaseelan. Medical imaging modalities: A survey. In *2016 International Conference on Emerging Trends in Engineering, Technology and Science (ICETETS)*, pages 1–4, 2016.

[44] Mehmet Günhan Ertosun and Daniel L. Rubin. Probabilistic visual search for masses within mammography images using deep learning. In *2015 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pages 1310–1315, 2015.

[45] Mehmet Günhan Ertosun and Daniel L. Rubin. Probabilistic visual search for masses within mammography images using deep learning. In *2015 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pages 1310–1315, 2015.

[46] Andre Esteva, Alexandre Robicquet, Bharath Ramsundar, Volodymyr Kuleshov, Mark DePristo, Katherine Chou, Claire Cui, Greg Corrado, and Jeff Dean. A guide to deep learning in healthcare. *Nature Medicine*, 25, 01 2019.

[47] J.Suckling et al. Mias.

[48] Lenin Falconí, María Pérez, Wilbert Aguilar, and Aura Conci. Transfer learning and fine tuning in mammogram bi-rads classification. In *2020 IEEE 33rd International Symposium on Computer-Based Medical Systems (CBMS)*, pages 475–480, 2020.

[49] Lenin G. Falconí, María Pérez, and Wilbert G. Aguilar. Transfer learning in breast mammogram abnormalities classification with mobilenet and nasnet. In *2019 International Conference on Systems, Signals and Image Processing (IWSSIP)*, pages 109–114, 2019.

[50] Navid Farahani, Anil V Parwani, and Liron Pantanowitz. Whole slide imaging in pathology: advantages, limitations, and emerging perspectives. *Pathology and Laboratory Medicine International*, 7:23–33, 2015.

[51] Rachel Farber, Nehmat Houssami, Sally Wortley, Gemma Jacklyn, Michael L Marinovich, Kevin McGeechan, Alexandra Barratt, and Katy Bell. Impact of full-field digital mammography versus film-screen mammography in population screening: a meta-analysis. *JNCI: Journal of the National Cancer Institute*, 113(1):16–26, 2021.

[52] Rachel Farber, Nehmat Houssami, Sally Wortley, Gemma Jacklyn, Michael L Marinovich, Kevin McGeechan, Alexandra Barratt, and Katy Bell. Impact of full-field digital mammography versus film-screen mammography in population screening: a meta-analysis. *JNCI: Journal of the National Cancer Institute*, 113(1):16–26, 2021.

[53] Y Faridah. Digital versus screen film mammography: a clinical comparison. *Biomedical imaging and intervention journal*, 4(4), 2008.

[54] Wael Fathy and Amr Ghoneim. A deep learning approach for breast cancer mass detection. *International Journal of Advanced Computer Science and Applications*, 10, 01 2019.

[55] Wael E. Fathy and Amr S. Ghoneim. A deep learning approach for breast cancer mass detection. *International Journal of Advanced Computer Science and Applications*, 10(1), 2019.

[56] Paweł Filipczuk, Thomas Fevens, Adam Krzyżak, and Andrzej Obuchowicz. Glcm and glrlm based texture features for computer-aided breast cancer diagnosis. *Journal of Medical Informatics & Technologies*, 19, 2012.

[57] Hiroshi Fujita. Ai-based computer-aided diagnosis (ai-cad): the latest review to read first. *Radiological Physics and Technology*, 13(1):6–19, 2020.

[58] K. Fukushima. Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position. *Biological Cybernetics*, 36:193–202, 2004.

[59] Krzysztof Geras, Stacey Wolfson, S. Kim, Linda Moy, and Kyunghyun Cho. High-resolution breast cancer screening with multi-view deep convolutional neural networks. 03 2017.

[60] Ross B. Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. *CoRR*, abs/1311.2524, 2013.

[61] Xavier Glorot, Antoine Bordes, and Y. Bengio. Deep sparse rectifier neural networks. *Proceedings of the 14th International Conference on Artificial Intelligence and Statisitics (AISTATS) 2011*, 15:315–323, 01 2011.

[62] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep learning*. MIT press, 2016.

[63] Hayit Greenspan, Bram van Ginneken, and Ronald M. Summers. Guest editorial deep learning in medical imaging: Overview and future promise of an exciting new technique. *IEEE Transactions on Medical Imaging*, 35(5):1153–1159, 2016.

[64] Yanming Guo, Yu Liu, Ard Oerlemans, Songyang Lao, Song Wu, and Michael S. Lew. Deep learning for visual understanding: A review. *Neurocomputing*, 187:27–48, 2016. Recent Developments on Deep Big Vision.

[65] Zhongyi Han, Benzheng Wei, Yuanjie Zheng, Yilong Yin, Kejian Li, and Shuo Li. Breast cancer multi-classification from histopathological images with structured deep learning model. *Scientific reports*, 7(1):1–10, 2017.

[66] Kaiming He and Jian Sun. Convolutional neural networks at constrained time cost. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5353–5360, 2015.

[67] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016.

[68] Pınar Uskaner Hepsağ, Selma Ayşe Özel, and Adnan Yazıcı. Using deep learning for mammography classification. In *2017 International Conference on Computer Science and Engineering (UBMK)*, pages 418–423, 2017.

[69] Sylvia H Heywang-Köbrunner, Astrid Hacker, and Stefan Sedlacek. Advantages and disadvantages of mammography screening. *Breast care*, 6(3):199–207, 2011.

[70] Geoffrey E. Hinton. Rectified linear units improve restricted boltzmann machines vinod nair.

[71] Syed A. Hoda and Rana S. Hoda. Rubin's Pathology: Clinicopathologic Foundations of Medicine, 5th Edition. *JAMA*, 298(17):2070–2075, 11 2007.

[72] Andrew G. Howard, Menglong Zhu, Bo Chen, D. Kalenichenko, W. Wang, Tobias Weyand, M. Andreetto, and Hartwig Adam. Mobilenets: Efficient convolutional neural networks for mobile vision applications. *ArXiv*, abs/1704.04861, 2017.

[73] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q. Weinberger. Densely connected convolutional networks. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2261–2269, 2017.

[74] D. H. Hubel and T. N. Wiesel. Receptive fields and functional architecture of monkey striate cortex. *The Journal of Physiology*, 195(1):215–243, 1968.

[75] Benjamin Huynh, Hui Li, and Maryellen Giger. Digital mammographic tumor classification using transfer learning from deep convolutional neural networks. *Journal of medical imaging (Bellingham, Wash.)*, 3:034501, 07 2016.

[76] M Mohsin Jadoon, Qianni Zhang, Ihsan Ul Haq, Sharjeel Butt, and Adeel Jadoon. Three-class mammogram classification based on descriptive cnn features. *BioMed research international*, 2017, 2017.

[77] Afsaneh Jalalian, Syamsiah Mashohor, Rozi Mahmud, Babak Karasfi, M Iqbal B Saripan, and Abdul Rahman B Ramli. Foundation and methodologies in computer-aided diagnosis systems for breast cancer detection. *EXCLI journal*, 16:113, 2017.

[78] Afsaneh Jalalian, Syamsiah BT Mashohor, Hajjah Rozi Mahmud, M Iqbal B Saripan, Abdul Rahman B Ramli, and Babak Karasfi. Computer-aided detection/diagnosis of breast cancer in mammography and ultrasound: a review. *Clinical imaging*, 37(3):420–426, 2013.

[79] Fan Jiang, Hui Liu, Shaode Yu, and Yaoqin Xie. Breast mass lesion classification in mammograms by transfer learning. In *Proceedings of the 5th international conference on bioinformatics and computational biology*, pages 59–62, 2017.

[80] Yun Jiang, Li Chen, Hai Zhang, and Xiao Xiao. Breast cancer histopathological image classification using convolutional neural networks with small se-resnet module. *PLoS ONE*, 14, 03 2019.

[81] Zhicheng Jiao, Xinbo Gao, Ying Wang, and Jie Li. A deep feature based framework for breast masses classification. *Neurocomputing*, 197:221–231, 2016.

[82] Yuliana Jiménez-Gaona, María José Rodríguez-Álvarez, and Vasudevan Lakshminarayanan. Deep-learning-based computer-aided systems for breast cancer imaging: A critical review. *Applied Sciences*, 10(22), 2020.

[83] Mahdi M. Kalayeh, Emrah Basaran, Muhittin Gökmen, Mustafa E. Kamasak, and Mubarak Shah. Human semantic parsing for person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

[84] Harshit Kaushik, Dilbag Singh, Shailendra Tiwari, Manjit Kaur, Chang-Won Jeong, Yunyoung Nam, and Muhammad Attique Khan. Screening of covid-19 patients using deep learning and iot framework. pages 3459–3475, 2021.

[85] Harshit Kaushik, Dilbag Singh, Shailendra Tiwari, Manjit Kaur, Chang-Won Jeong, Yunyoung Nam, and Muhammad Attique Khan. Screening of covid-19 patients using deep learning and iot framework. *Cmc-Computers Materials & Continua*, pages 3459–3475, 2021.

[86] Salman Khan, Hossein Rahmani, Syed Afaq Ali Shah, and Mohammed Bennamoun. A guide to convolutional neural networks for computer vision. *Synthesis Lectures on Computer Vision*, 8(1):1–207, 2018.

[87] Dae Hoe Kim, Seong Tae Kim, and Yong Man Ro. Latent feature representation with 3-d multi-view deep convolutional neural network for bilateral analysis in digital breast tomosynthesis. In *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 927–931, 2016.

[88] Dae Hoe Kim, Seong Tae Kim, and Yong Man Ro. Latent feature representation with 3-d multi-view deep convolutional neural network for bilateral analysis in digital breast tomosynthesis. In *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 927–931, 2016.

[89] Myeong Seong Kim. Investigation of factors affecting body temperature changes during routine clinical head magnetic resonance imaging. *Iranian journal of radiology*, 13(4), 2016.

[90] Alex Krizhevsky, Ilya Sutskever, and Geoffrey Hinton. Imagenet classification with deep convolutional neural networks. *Neural Information Processing Systems*, 25, 01 2012.

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

[92] Rebecca Sawyer Lee, Francisco Gimenez, Assaf Hoogi, Kanae Kawai Miyake, Mia Gorovoy, and Daniel L. Rubin. Data Descriptor: A curated mammography data set

for use in computer-aided detection and diagnosis research. *Scientific Data*, 4:1–9, 2017.

[93] Thomas Lehmann, Mark Güld, Christian Thies, Bartosz Plodowski, Daniel Keysers, Bastian Ott, and Henning Schubert. Irma-content-based image retrieval in medical applications. *Studies in health technology and informatics*, 107:842–6, 02 2004.

[94] Daniel Lévy and Arzav Jain. Breast mass classification from mammograms using deep convolutional neural networks. *CoRR*, abs/1612.00542, 2016.

[95] Zhizhong Li and Derek Hoiem. Learning without forgetting. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(12):2935–2947, 2018.

[96] Geert Litjens, Thijs Kooi, Babak Ehteshami Bejnordi, Arnaud Arindra Adiyoso Setio, Francesco Ciompi, Mohsen Ghafoorian, Jeroen Awm Van Der Laak, Bram Van Ginneken, and Clara I Sánchez. A survey on deep learning in medical image analysis. *Medical image analysis*, 42:60–88, 2017.

[97] Elzbieta Luczyńska, Sylwia Heinze, Agnieszka Adamczyk, Janusz Rys, Jerzy W Mitus, and Edward Hendrick. Comparison of the mammography, contrast-enhanced spectral mammography and ultrasonography in a group of 116 patients. *Anticancer research*, 36(8):4359–4366, 2016.

[98] Andrew L. Maas. Rectifier nonlinearities improve neural network acoustic models. 2013.

[99] Richard Maclin and David W. Opitz. Popular ensemble methods: An empirical study. *CoRR*, abs/1106.0257, 2011.

[100] Tariq Mahmood, Jianqiang Li, Yan Pei, Faheem Akhtar, Azhar Imran, and Khalil Ur Rehman. A brief survey on breast cancer diagnostic with deep learning schemes using multi-image modalities. *IEEE Access*, 8:165779–165809, 2020.

[101] Andreas Maier, Christopher Syben, Tobias Lasser, and Christian Riess. A gentle introduction to deep learning in medical image processing. *Zeitschrift für Medizinische Physik*, 29(2):86–101, 2019.

[102] Edison Moya, Emerson Campoverde, Eduardo Tusa, Ivan Ramirez-Morales, Wilmer Rivas, and Bertha Mazon. Multi-category classification of mammograms by using convolutional neural networks. In *2017 International Conference on Information Systems and Computer Science (INCISCOS)*, pages 133–140, 2017.

[103] Chisako Muramatsu, Mizuho Nishio, Takuma Goto, Mikinao Oiwa, Takako Morita, Masahiro Yakami, Takeshi Kubo, Kaori Togashi, and Hiroshi Fujita. Improving breast mass classification by shared data with domain transformation using a generative adversarial network. *Computers in Biology and Medicine*, 119:103698, 2020.

[104] Ghulam Murtaza, Liyana Shuib, Ainuddin Wahid, Abdul Wahab, Ghulam Mujtaba, Henry Nweke, Mohammed Al-Garadi, Fariha Zulfiqar, Ghulam Raza, and Aniza Azmi. Deep learning-based breast cancer classification through medical imaging modalities: State of the art and research challenges. 05 2019.

[105] Hasan Nasir Khan, Ahmad Raza Shahid, Basit Raza, Amir Hanif Dar, and Hani Alquhayz. Multi-view feature fusion based four views model for mammogram classification using convolutional neural network. *IEEE Access*, 7:165724–165733, 2019.

[106] Maxime Oquab, Léon Bottou, Ivan Laptev, and Josef Sivic. Is object localization for free? - weakly-supervised learning with convolutional neural networks. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 685–694, 2015.

[107] Sinno Jialin Pan and Qiang Yang. A survey on transfer learning. *IEEE Transactions on Knowledge and Data Engineering*, 22(10):1345–1359, 2010.

[108] Babita Pandey, Devendra Kumar Pandey, Brijendra Pratap Mishra, and Wasiur Rhmann. A comprehensive survey of deep learning in the field of medical imaging and medical natural language processing: Challenges and research directions. *Journal of King Saud University-Computer and Information Sciences*, 2021.

[109] Santanu Pattanayak. Pro deep learning with tensorflow: A mathematical approach to advanced artificial intelligence in python. 2017.

[110] Dabal Pedamonti. Comparison of non-linear activation functions for deep neural networks on mnist classification task. 04 2018.

[111] Ana C Perre, Luís A Alexandre, and Luís C Freire. Lesion classification in mammograms using convolutional neural networks and transfer learning. *Computer Methods in Biomechanics and Biomedical Engineering: Imaging & Visualization*, 2018.

[112] Nicholas Petrick, Berkman Sahiner, Samuel G Armato III, Alberto Bert, Loredana Correale, Silvia Delsanto, Matthew T Freedman, David Fryd, David Gur, Lubomir Hadjiiski, et al. Evaluation of computer-aided detection and diagnosis systems a. *Medical physics*, 40(8):087001, 2013.

[113] Sergey M Plis, Devon R Hjelm, Ruslan Salakhutdinov, Elena A Allen, Henry J Bockholt, Jeffrey D Long, Hans J Johnson, Jane S Paulsen, Jessica A Turner, and Vince D Calhoun. Deep learning for neuroimaging: a validation study. *Frontiers in neuroscience*, 8:229, 2014.

[114] Francisco Javier Pérez-Benito, François Signol, Juan-Carlos Perez-Cortes, Alejandro Fuster-Baggetto, Marina Pollan, Beatriz Pérez-Gómez, Dolores Salas-Trejo, Maria Casals, Inmaculada Martínez, and Rafael LLobet. A deep learning system to obtain the optimal parameters for a threshold-based breast and dense tissue segmentation. *Computer Methods and Programs in Biomedicine*, 195:105668, 2020.

[115] Dina A. Ragab, Maha Sharkas, Stephen Marshall, and Jinchang Ren. Breast cancer detection using deep convolutional neural networks and support vector machines. *PeerJ*, 2019(1):1–23, 2019.

[116] Andrik Rampun, Bryan W. Scotney, Philip J. Morrow, and Hui Wang. Breast mass classification in mammograms using ensemble convolutional neural networks. *2018 IEEE 20th International Conference on e-Health Networking, Applications and Services, Healthcom 2018*, 2018.

[117] Joana Reis, Jonas Christoffer Lindstrøm, Joao Boavida, Kjell-Inge Gjesdal, Daehoon Park, Nazli Bahrami, Manouchehr Seyedzadeh, Woldegabriel A Melles, Torill Sauer, Jürgen Geisler, et al. Accuracy of breast mri in patients receiving neoadjuvant endocrine therapy: comprehensive imaging analysis and correlation with clinical and pathological assessments. *Breast cancer research and treatment*, 184(2):407–420, 2020.

[118] Jimmy Ren and Li Xu. On vectorization of deep convolutional neural networks for vision tasks. 01 2015.

[119] Ye Ren, Le Zhang, and Ponnuthurai Suganthan. Ensemble classification and regression-recent developments, applications and future directions [review article]. *IEEE Computational Intelligence Magazine*, 11:41–53, 02 2016.

[120] Chris Rose. Usf digital mammography home page.

[121] Sebastian Ruder, Matthew E Peters, Swabha Swayamdipta, and Thomas Wolf. Transfer learning in natural language processing. pages 15–18, 2019.

[122] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet

large scale visual recognition challenge. *International journal of computer vision*, 115(3):211–252, 2015.

[123] Mehul Sampat, Mia Markey, and Alan Bovik. *Computer-Aided Detection and Diagnosis in Mammography*, pages 1195–1217. 01 2005.

[124] Dominik Scherer, Andreas Müller, and Sven Behnke. Evaluation of pooling operations in convolutional architectures for object recognition. In Konstantinos Diamantaras, Wlodek Duch, and Lazaros S. Iliadis, editors, *Artificial Neural Networks – ICANN 2010*, pages 92–101, Berlin, Heidelberg, 2010. Springer Berlin Heidelberg.

[125] Egemen Sert, Seyda Ertekin, and Ugur Halici. Ensemble of convolutional neural networks for classification of breast microcalcification from mammograms. volume 2017, pages 689–692, 07 2017.

[126] Khalid Shaikh, Sabitha Krishnan, and Rohit Thanki. *Artificial Intelligence in Breast Cancer Early Detection and Diagnosis*. Springer, 2021.

[127] Li Shen, Laurie Margolies, Joseph Rothstein, Eugene Fluder, Russell McBride, and Weiva Sieh. Deep learning to improve breast cancer detection on screening mammography. *Scientific Reports*, 9:1–12, 08 2019.

[128] Jacob Shreffler and Martin R Huecker. Type i and type ii errors and statistical power. 2020.

[129] E A Sickles. Periodic mammographic follow-up of probably benign lesions: results in 3,184 consecutive cases. *Radiology*, 179(2):463–468, 1991. PMID: 2014293.

[130] Rebecca L Siegel, Kimberly D Miller, and Ahmedin Jemal. Cancer statistics, 2015. *CA: a cancer journal for clinicians*, 65(1):5—29, 2015.

[131] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[132] Canadian Cancer Society. Canadian-cancer-society-estimation-2021.

[133] Fabio Alexandre Spanhol, Luiz S. Oliveira, Caroline Petitjean, and Laurent Heutte. Breast cancer histopathological image classification using convolutional neural networks. In *2016 International Joint Conference on Neural Networks (IJCNN)*, pages 2560–2567, 2016.

[134] Yong Joon Suh, Jaewon Jung, and Bum-Joo Cho. Automated breast cancer detection in digital mammograms of various densities via deep learning. *Journal of personalized medicine*, 10(4):211, 2020.

[135] Heung-Il Suk and Dinggang Shen. Deep learning-based feature representation for ad/mci classification. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 583–590. Springer, 2013.

[136] Lilei Sun, Jie Wen, Junqian Wang, Yong Zhao, and Yong Xu. Classification of mammography based on semi-supervised learning. In *2020 IEEE International Conference on Progress in Informatics and Computing (PIC)*, pages 104–111, 2020.

[137] S Suzuki, X Zhang, Noriyasu Homma, Kei Ichiji, Y Kawasumi, T Ishibashi, and Makoto Yoshizawa. We-de-207b-02: Detection of masses on mammograms using deep convolutional neural network: A feasibility study. *Medical Physics*, 43:3817–3817, 06 2016.

[138] T.M. Svahn, N. Houssami, I. Sechopoulos, and S. Mattsson. Review of radiation dose estimates in digital breast tomosynthesis relative to those in two-view full-field digital mammography. *The Breast*, 24(2):93–99, 2015.

[139] Bartosz Swiderski, Jaroslaw Kurek, Stanislaw Osowski, Michal Kruk, and Walid Barhoumi. Deep learning and non-negative matrix factorization in recognition of mammograms. In Yulin Wang, Tuan D. Pham, Vit Vozenilek, David Zhang, and Yi Xie, editors, *Eighth International Conference on Graphic and Image Processing (ICGIP 2016)*, volume 10225, pages 53 – 59. International Society for Optics and Photonics, SPIE, 2017.

[140] Christian Szegedy, Sergey Ioffe, Vincent Vanhoucke, and Alexander A Alemi. Inception-v4, inception-resnet and the impact of residual connections on learning. In *Thirty-first AAAI conference on artificial intelligence*, 2017.

[141] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1–9, 2015.

[142] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2818–2826, 2016.

[143] Mingxing Tan and Quoc V. Le. Efficientnet: Rethinking model scaling for convolutional neural networks. *ArXiv*, abs/1905.11946, 2019.

[144] Luis Torada, Lucrezia Lorenzon, Alice Beddis, Ulas Isildak, Linda Pattini, Sara Mathieson, and Matteo Fumagalli. Imagene: a convolutional neural network to quantify natural selection from genomic data. *BMC Bioinformatics*, 20, 11 2019.

[145] Lazaros Tsochatzidis, Lena Costaridou, and Ioannis Pratikakis. Deep learning for breast cancer diagnosis from mammograms—a comparative study. *Journal of Imaging*, 5(3), 2019.

[146] Aswiga R V, Aishwarya R, and Shanthi A P. Augmenting transfer learning with feature extraction techniques for limited breast imaging datasets. *Journal of Digital Imaging*, 34:618 – 629, 2021.

[147] Stefan Wager, Sida Wang, and Percy Liang. Dropout training as adaptive regularization. *Advances in Neural Information Processing Systems*, 07 2013.

[148] Li Wan, Matthew Zeiler, Sixin Zhang, Yann Le Cun, and Rob Fergus. Regularization of neural networks using dropconnect. In *International conference on machine learning*, pages 1058–1066. PMLR, 2013.

[149] Jun Wang, Qianying Liu, Haotian Xie, Zhaogang Yang, and Hefeng Zhou. Boosted efficientnet: Detection of lymph node metastases in breast cancer using convolutional neural network. *CoRR*, abs/2010.05027, 2020.

[150] Sida Wang and C.D. Manning. Fast dropout training. *30th International Conference on Machine Learning, ICML 2013*, pages 777–785, 01 2013.

[151] David Warde-Farley, Ian Goodfellow, Aaron Courville, and Y. Bengio. An empirical analysis of dropout in piecewise linear networks. 12 2013.

[152] R M L Warren and W Duffy. Comparison of single reading with double reading of mammograms, and change in effectiveness with experience. *The British Journal of Radiology*, 68(813):958–962, 1995. PMID: 7496693.

[153] Pengcheng Xi, Chang Shu, and Rafik Goubran. Abnormality detection in mammography using deep convolutional neural networks. In *2018 IEEE International Symposium on Medical Measurements and Applications (MeMeA)*, pages 1–6, 2018.

[154] Yuanpu Xie, Zizhao Zhang, Manish Sapkota, and Lin Yang. Spatial clockwork recurrent neural network for muscle perimysium segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 185–193. Springer, 2016.

[155] Son Eun Ju Youk Ji Hyun, Gweon Hye Mi. Shear-wave elastography in breast ultrasonography: the state of the art. *Ultrasonography*, 36(4):300–309, 2017.

[156] Mina Yousefi, Adam Krzyżak, and Ching Y Suen. Mass detection in digital breast tomosynthesis data using convolutional neural networks and multiple instance learning. *Computers in biology and medicine*, 96:283–293, 2018.

[157] Yue Zhao, Xuejian Wang, Cheng Cheng, and Xueying Ding. Combining machine learning models and scores using combo library. In *Thirty-Fourth AAAI Conference on Artificial Intelligence*, New York, USA, Feb 2020.