ADDIS ABABA UNIVERSITY
ADDIS ABABA INSTITUTE OF TECHNOLOGY (AAiT)
SCHOOL OF ELECTRICAL AND COMPUTER ENGINEERING

AUTOMATED CONSTRUCTION OF A NEW DATASET FOR HISTOPATHOLOGICAL BREAST CANCER IMAGES

BY
KALKIDAN KEBEDE

ADVISOR
Dr. FITSUM ASSAMNEW

A thesis submitted to the School of Electrical and Computer Engineering in partial fulfillment of the requirements for the Degree of Master of Science in Computer Engineering

JANUARY, 2024
ADDIS ABABA, ETHIOPIA

ADDIS ABABA UNIVERSITY
ADDIS ABABA INSTITUTE OF TECHNOLOGY
SCHOOL OF ELECTRICAL AND COMPUTER ENGINEERING

The undersigned have examined the thesis titled:

AUTOMATED CONSTRICTION OF A NEW DATASET FOR HISTOPATHOLOGICAL BREAST CANCER IMAGES

BY

KALKIDAN KEBEDE

Approval by Boards of Examiners

Dr. Bisrat Derebssa	_____	_____
Dean, SECE, AAiT	Date	Signature
Dr. Fitsum Assamnew	_____	_____
Advisor	Date	Signature
Dr. Menore Tekeba	_____	_____
Internal Examiner	Date	Signature
Dr. Surafel Lemma	_____	_____
External Examiner	Date	Signature

Declaration

I, Kalkidan Kebede Derese, hereby declare that this thesis is entirely my original work. All sources of information used in this study have been appropriately acknowledged and referenced. I confirm that this thesis has not been submitted, either in part or in full, for any other academic requirements or to any other learning institution. This work is solely my own contribution to the field of study.

Student Name: Kalkidan Kebede Derese

Signature: _____

Date: _____

JANUARY, 2024

Acknowledgments

First and foremost, I express my deep gratitude to the almighty GOD for the abundance of blessings, guidance, and protection that have illuminated my journey.

I would like to extend my sincere appreciation to Dr. Firstum Assamnew, whose dedicated mentorship has not only nurtured my academic growth but also inspired me to exceed boundaries. I would also like to extend my sincere thanks to Mr. Kaleab Alemayehu, a Ph.D. candidate, for his valuable support throughout this journey. Additionally, I extend my appreciation to the three pathologists who contributed their expertise in evaluating this project: Dr. Daniel Kassie Molla, MD, pathologist; Dr. Anteneh Belachew Fantaye, Assistant Professor of Anatomical Pathology at Haramaya University; and Dr. Wubshet Assefa, Assistant professor of pathology at the College of Medicine and Health Sciences, Bahir Dar University.

I am grateful to the AAiT ICT department, particularly Mr. Leul K/Mariam, for his constant support. Additionally, I want to thank my friends and fellow Computer Engineering MSc students at AAiT for their provision of resources, unwavering support, and collaborative spirit.

Lastly, I would like to express my deepest gratitude to my family, whose endless love, unwavering support, and encouragement have been my cornerstone through academic and personal challenges.

Abstract

Cancer is a medical condition where cells grow uncontrollably and can spread to other parts of the body, posing a significant global health challenge. Among women worldwide, breast cancer is the most frequently diagnosed cancer and the leading cause of cancer-related deaths. Automated classification of breast cancer has been extensively studied, particularly in differentiating types, subtypes, and stages. However, simultaneous classification of subtypes with stages, such as Lobular Carcinoma In Situ (LCIS) and Invasive Lobular Carcinoma (ILC), remains challenging due to limited data availability.

This research aims to address this gap by generating a new dataset that includes these unclassified subtypes with staging, utilizing existing datasets as primary sources. Labels for ductal and lobular carcinoma from the BreakHis dataset and invasive and in situ carcinoma labels from the Yan et al. dataset are used to train models for generating the new dataset.

To achieve this, two separate ensemble models are trained using distinct datasets. The first ensemble model classifies ductal and lobular carcinoma using the BreakHis dataset. The second ensemble model classifies invasive and in situ carcinoma using the Yan et al. dataset. Both models are then used to extract a new dataset through soft voting techniques. The extracted labels include Ductal Carcinoma In Situ (DCIS), Invasive Ductal Carcinoma (IDC), LCIS, and ILC. This approach aims to provide a more comprehensive classification system by leveraging labels from both datasets.

To validate the newly extracted labels, three pathologists were given randomly extracted images from the Yan et al. dataset test set. The pathologists agreed with the model outputs on 87.5% of the samples. Subsequently, the newly generated dataset was used to classify DCIS, IDC, LCIS, and ILC with an accuracy of 76.06%.

Keywords: Breast cancer, histopathology, DCIS, IDC, LCIS, ILC, BreakHis, Yan et al.

Table of Contents

List of Figures

List of Tables

List of Acronyms

IDC Invasive Ductal Carcinoma

DCIS Ductal Carcinoma In Situ

ILC Invasive Lobular Carcinoma

LCIS Lobular Carcinoma In Situ

MRI Magnetic Resonance Imaging

BC Breast Cancer

CNN Convolutional Neural Network

RGB Red, Green, and Blue

PNG Portable Network Graphics

TIFF Tag Image File Format

SVM Support Vector Machine

KNN k-nearest neighbors

PNN Probabilistic Neural Network

SNR Signal to Noise Ratio

SFS Sequential Forward Selection

PCA Principal Component Analysis

AUC Area Under the Curve

RF Random Forest

LR Logistic Regression

GAP Global Average Pooling

OOD Out-of-Distribution

Chapter 1

Introduction

Cancer is a pervasive and complex disease characterized by the rapid and abnormal growth of cells, leading to the potential spread of these cells to other parts of the body, a process known as metastasis. Globally, cancer remains a leading public health challenge, contributing to nearly 10 million reported deaths in 2020. Breast cancer (2.26 million cases) was one of the most prevalent types of cancer in terms of new cases in 2020, along with lung cancer (2.21 million cases), colon and rectum cancer (1.93 million cases), prostate cancer (1.41 million cases), non-melanoma skin cancer (1.20 million cases), and stomach cancer (1.09 million cases) [1, 2].

Breast cancer, in particular, holds the position of being the most common cancer in the world and the most frequently diagnosed cancer in women. According to recent statistics, approximately 685,000 women lost their lives to breast cancer in 2020 [2, 3, 4]. Furthermore, it is the most frequently diagnosed cancer among women in 140 countries worldwide. Breast cancer also accounts for almost a quarter of all new cancer cases in women [4]. Therefore, accessible breast cancer diagnosis is essential because it facilitates early detection and intervention, ultimately improving outcomes and saving lives.

1.1 Breast Health and Disease Overview

1.1.1 Breast Anatomy

The breast is an organ that produces breast milk and is classified as an exocrine gland in the human body. Mainly, the breast is made up of glandular tissue, fatty tissue, connective tissue, ducts, a nipple and an areola [5, 6]. The glandular tissue consists of lobes, lobules, and milk ducts, which produce and transport milk in nursing women as shown in Figure 1.1. The fatty tissue determines the breast size and fills the spaces between glandular and fibrous tissue. The connective tissue, also known as fibrous or supportive tissue, holds the glandular and fatty tissue in place and includes ligaments that stretch from the skin to the chest wall to support the breast tissue [7, 8].

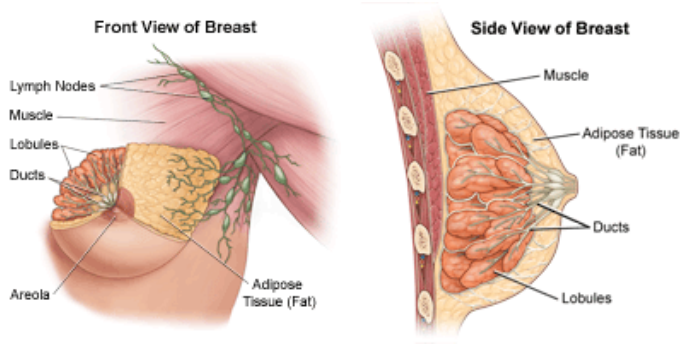

Figure 1.1: Anatomy of the breast

1.1.2 Breast Tumor

A breast tumor is a mass of abnormal tissue that can be either benign or malignant.

(i) Benign (Noncancerous) Tumor: Benign cases are considered noncancerous and thus do not pose any threat to life. However, in some cases, they may become cancerous. Typically, the immune system isolates benign tumors from other cells by surrounding them with a ”sac,” which makes it easy to remove them from the body [9].

(ii) Malignant (Cancerous) Tumor: Malignant tumors are a type of cancer that develops when cells in the breast tissue grow and multiply uncontrollably. If left untreated, these cancerous cells potentially spread to other body parts [10].

1.1.3 Breast Cancer Types

Depending on its ability to spread inside the organism and attack organs and tissue far away from the point of origin, breast cancer can be [11]:

(i) Non-Invasive (or In Situ): It stays localized to the area of the breast where it originates, without spreading through the surrounding breast tissue.

(ii) Invasive: When the neoplasm is able to migrate through the lymphatic system and blood and gradually compromise vital functions.

Regions of the breast may or may not develop into invasive or in situ forms of breast cancer. The type of breast cancer is determined by the specific kind of cells in the breast that are affected. The following list provides the most common types of breast cancer [10]:

- **DCIS:** The most prevalent form of non-invasive breast cancer where abnormal cells have been found in the lining of the breast milk duct.
- **LCIS:** Starts in the breast milk glands(lobules) but does not metastasize to other regions of the body.
- **IDC:** Begins in the milk ducts of the breast and penetrates the wall of the duct, invading the fatty tissue of the breast and possibly other regions of the body.
- **ILC:** Starts in the breast milk glands(lobules) and metastasize to other regions of the body.
- **Medullary Carcinoma:** Is an invasive breast cancer that forms a distinct boundary between tumor tissue and normal tissue.
- **Mutinous Carcinoma:** This is a rare breast cancer formed by the mucus-producing cancer cells.
- **Phyllodes Tumor:** Can be either benign (non-cancerous) or malignant (cancerous) that start in the connective (stromal) tissue of the breast.

1.1.4 Breast Cancer Diagnosis

Breast cancer can be detected using various tests. If a physician finds an area of concern on a screening test, such as a mammogram or if a person experiences symptoms that could indicate breast cancer, further tests will be needed to confirm the diagnosis. Regular breast cancer screening is important to detect breast cancer early before it causes symptoms. Different tests that can be used to diagnose breast cancer include breast exams, mammography, breast ultrasound, breast Magnetic Resonance Imaging (MRI) and biopsy [12].

- **Diagnostic Mammography:** Is a medical tool made specifically to capture an X-ray image of the breast to detect breast cancer. Low-energy X-rays, which is a form of ionizing radiation, are used to photograph the female breast that produce gray scale. These gray scale images are used by radiologists or physicians to identify any lumps or other abnormalities in the breast. Mammography is typically used to examine masses to discover breast cancer early [13].
- **Ultrasound:** Ultrasound is a medical imaging technique that uses sound waves to examine, diagnose, or treat internal organs in the body. These sound waves transform into an ultrasound image that displays the state and limits of the body's interior organs, fluid, and soft tissue [14].
- **Magnetic Resonance Imaging (MRI):** This is a non-invasive imaging technique that uses strong magnetic fields and radio waves to produce detailed images of the breast tissue. It is an ultimate supplementary imaging modality in addition to mammography and ultrasonography in the evaluation of breast disease [15].
- **Biopsy:** A breast biopsy is a procedure where tissue or fluid is taken from the suspected location. The excised cells are studied under a microscope and put through additional testing to see if breast cancer is present. Only a biopsy can properly confirm whether a suspicious spot is malignant in terms of diagnosis. There are different types of breast biopsy, including:
  - **Fine-needle aspiration:** The simplest sort of breast biopsy can be used to examine a lump that is felt during a clinical breast exam. The patient is placed on a table for the treatment, which involves inserting a very thin needle into the mass.

- **Core needle biopsy:** A bigger hollow needle is used in this type of breast biopsy to collect breast alterations that the clinician has felt or that have been seen on an ultrasound, mammography or MRI. The lymph nodes under the arm may also be examined for signs of cancer spread or to evaluate a breast lump that is not visible on imaging.
- **Surgical biopsy:** A larger sample of breast tissue is removed during this sort of biopsy. It is often utilized when a larger tissue sample is required for diagnosis or when other types of biopsy are inconclusive [16].

1.2 Breast Cancer Data Set

Numerous imaging techniques are currently under development to detect breast cancer at its early stages. Histopathology imaging is one such technique that has gained significant attention from researchers. Histopathological images are crucial in medical diagnostics, providing visual representations of biological tissues or cells that undergo histopathological examination. This process involves microscopic study to identify abnormalities or diseases, helping to diagnose various conditions, including cancers. Obtaining a tissue sample through a biopsy, processing and staining the sample, examining it under a microscope, and capturing images are typical steps in this process. These images provide detailed insights into tissue nature, guiding treatment decisions, prognosis, and further research into understanding disease mechanisms, making them crucial in clinical pathology [17]. Several researchers have published different datasets on breast cancer using this technique. To give an idea, below is a Table 1.1 that lists some of these datasets.

Table 1.1: Datasets for breast cancer histopathological imaging modality.

Dataset	Year	Staining	Data size	Magnification
BreakHis [18]	2015	H&E	7909 HI	40×, 100×,200×, 400×
Camelyon16 [19]	2016	H&E	400 lymph node WSIs	40×, 10×, 1×
Camelyon17 [19]	2017	H&E	200 lymph node WSIs	40×
TUPAC [20]	2016	H&E	500 training set and 321 testing set BC-HI WSIs	40×
BACH [21]	2018	H&E	400 microscopy images and 30 WSIs	40×
ICPR [22]	2012	H&E	50 images corresponding to 5 different biopsy slides	40×
IDC [23]	2014	H&E	277,524 patches are from 162 IDC breast cancer histopathological slides	40×
Yan et al. [24]	2019	H&E	249 Initial image and 3771 Extended image all together 4020	100×,200×
Bio-imaging Grand Challenge [25]	2015	H&E	A training set of 249 images, a test set of 20 images and an extended test set of 16 images	200×

- **BreakHis [18] Dataset:** Contains four histological distinct types of benign breast tumors: adenosis (A), fibroadenoma (F), phyllodes tumor (PT), and tubular adenoma (TA); and four malignant tumors (breast cancer): ductal carcinoma (DC), lobular carcinoma (LC), mucinous carcinoma (MC) and papillary carcinoma (PC). The images were saved in 3-channel Red, Green, and Blue (RGB), 8-bit depth in each channel, PNG (Portable Network Graphics) format with no compression, and dimension of 700 × 460 pixels [18].

- **Yan et al. [24] Dataset:** Contains normal, benign, in situ carcinoma and invasive carcinoma regions. Microscopy images are in Tag Image File Format (TIFF) format, with channel RGB and dimension of 2048 x 1536 pixels [24]. The dataset contains a total of 3,771 microscopy images.

- **BACH [21] Dataset:** Contains normal, benign, in situ carcinoma and invasive carcinoma regions. Microscopy images are in TIFF format, with channel RGB and dimension of 2048 x 1536 pixels [26]. The dataset contains a total of 400 microscopy images, distributed as follows:

  - Normal: 100
  - Benign: 100
  - In situ carcinoma: 100
  - Invasive carcinoma: 100

- **IDC Dataset:** Contains the invasive ductal carcinoma breast cancer sub-type. The images are in Portable Network Graphics (PNG) format, with channel RGB and dimensions of 50 x 50 pixels [23].

- **Bio-imaging Grand Challenge:** Contains normal, benign, in situ carcinoma and invasive carcinoma regions [25].

- **Breast Cancer (BC) Histopathological Image Dataset:** Contains normal, benign, in situ carcinoma and invasive carcinoma regions. Microscopy images are in TIFF format, with channel RGB and dimension of 2048 x 1536 [24].

1.3 Problem Statement

Breast cancer is a complex disease that has been classified in various ways by scholars. Some have classified breast cancer as benign and malignant [18, 27, 28, 29, 30], while others have tried to classify benign and malignant with their sub-types [31, 32, 33]. Other classifications include in situ and invasive carcinoma breast cancer staging. In situ carcinoma refers to the first stage of breast cancer when it is still just in the layer of cells where it first emerged, whereas invasive carcinoma refers to the disease spreading outside of the layer of cells where it first appeared. From the malignant subtypes, ductal and lobular have a staging of in situ and invasive.

Many researchers have used the BreakHis [18] dataset that was introduced by Spanhol et al [18] for breast cancer detection and classification purposes. They categorized the sub-types of tumors as benign (adenosis, fibroadenoma, phyllodes tumor, and tubular adenoma) and malignant (ductal carcinoma, lobular carcinoma, mucinous carcinoma, and papillary carcinoma). Also, other researchers used a dataset produced by Yan et al. [24] for the classification of normal, benign, in situ carcinoma and invasive carcinoma. Others attempted classifying breast cancer subtypes with staging such as IDC and DCIS [24, 34, 35] using the IDC Histology Image Dataset [35] and C. Wetstein et al. [34]. However, no research has yet been done on the classification of ILC and LCIS. Moreover, there are no specialized datasets available for these categories. It's noteworthy that ILC accounts for up to 15% of all breast cancer cases, and there's evidence to suggest that LCIS may progress to ILC in some instances [36].

In order to solve the problems stated earlier, this research aim to extract a dataset for ILC and LCIS from two different histopathological breast cancer datasets.

1.4 Objectives

1.4.1 General Objective

In this research, we aim to extract a dataset for invasive and in situ lobular carcinoma from two different histopathological breast cancer datasets.

1.4.2 Specific Objectives

The following specific objectives are carried out in order to achieve the general objective of this study:

- To prepare IDC, DCIS, ILC and LCIS dataset using the BreakHis [18] and Yan et at. [24] datasets.
- To verify the lables in the new dataset by pathologists.
- To build an ensemble of Convolutional Neural Network (CNN) models for classifying the four breast cancer subtypes.
- To evaluate the performance of the proposed model.

1.5 Contribution

There are three significant contributions made in this paper.

- The first contribution of the study is the development of a newly labeled dataset using BreakHis [18] and Yan et al. [24] datasets. Ductal carcinoma and lobular carcinoma labeled images were extracted from the malignant tumor section's BreakHis [18] dataset. From the Yan et al. [24] dataset in situ carcinoma and invasive carcinoma labeled images were taken. New labels DCIS, IDC, ILC, and LCIS were developed from the two distinct labeled datasets.
- We have put forth an approach leveraging an ensemble of models trained with varying magnification factors. This method effectively extracts labels from the Yan et al. [24] dataset, which lacks specific magnification factor information.
- We have developed a classifier for ILC and LCIS breast cancer types using the newly developed dataset.

1.6 Scope and Limitation

In this work, we focused on classifying DCIS, IDC, ILC, and LCIS from the newly constructed data set. Other breast cancer subtypes, including some of the malignant subtypes and the benign subtypes, are not addressed. In addition, we did not focus on the magnification factor, but rather on generating a new labeled dataset.

1.7 Organization of the Study

The remaining sections of this paper are arranged as follows: Chapter Two provides a detailed explanation of the theoretical foundations used for extracting a new dataset for breast cancer. The literature review in Chapter Three discusses the categorization and grading of breast cancer with an emphasis on the algorithm, data sets, and data pre-processing techniques that are employed. The methods used for the research are covered in Chapter Four. Chapter Five presents the outcomes of the analysis and implications, while Chapter Six explains the conclusions and future work.

Chapter 2

Background

This chapter discusses the theoretical foundations of deep learning focusing on CNN. It provides a comprehensive overview of the CNN algorithms relevant to our research goal. Later, we will explain the concept of transfer learning and its application in enhancing model performance. Furthermore, we will explore ensemble models, another critical aspect of our proposed methodology. Finally, we will look into the distribution shift and out-of-distribution detection.

2.1 Deep Learning

Deep Learning is a type of machine learning that draws inspiration from the structure of the human brain. By utilizing a hierarchy of concepts to learn from data, deep learning models solve a wide range of complex problems, such as image recognition, speech recognition, and natural language processing. This approach allows for intricate concepts to be learned by building upon simpler ones, making it possible for deep learning systems to handle enormous amounts of data without becoming overwhelmed [37].

There exist several types of deep learning neural networks, including Artificial Neural Networks (ANNs), Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs), and Generative Adversarial Networks (GANs).

Convolutional Neural Networks (CNNs) are highly effective in image classification tasks [38]. This chapter aims to explain CNN as we will be using this technique.

2.1.1 Convolutional Neural Networks (CNNs)

Convolutional neural networks (CNNs) have emerged as a powerful tool for feature extraction and pattern recognition in a diverse range of domains. Building upon the foundational principles of artificial neural networks (ANNs), CNN excel at processing grid-like structured data, particularly images and time series. Over the past decade, they have revolutionized fields like image processing, computer vision, and natural language processing, achieving groundbreaking results in tasks such as object detection, image segmentation, and speech recognition. CNN consist of three types of layers: convolutional layers, pooling layers, and fully-connected layers [39, 40].

The Convolutional Layer

In a CNN, the convolutional layer is the primary building block and where the majority of computation takes place. It requires three key components: input data, a filter, and a feature map. The convolutional layer performs a dot product between two matrices, where one matrix is the set of learnable parameters (known as a kernel), and the other matrix is the restricted portion of the receptive field. Although the kernel is spatially smaller than an image, it is deeper. For instance, if the image has three RGB channels, the kernel's height and width will be spatially small, but the depth extends up to all three channels.

During the forward pass, the kernel slides across the height and width of the image, producing a two-dimensional representation of the image known as an activation map. This map gives the response of the kernel at each spatial position of the image. The sliding size of the kernel is referred to as a stride. If we have an input of size W x W x D and Dout number of kernels with a spatial size of F, a stride of S, and padding P, we can determine the size of the output volume using the following formula:

$$W_{\text{out}} = \frac{W - F + 2P}{S} + 1 \tag{2.1}$$

Pooling layer

Pooling layers, referred to as downsampling, are responsible for reducing the dimensionality of the input by decreasing the number of parameters. Similar to the convolutional layer, the pooling operation applies a filter to the entire input. However, unlike the convolutional layer, this filter does not have any weights. Instead, the kernel applies an aggregation function to the values within the receptive field, resulting in an output array. Global Pooling, Stochastic Pooling, Max Pooling, Average Pooling, and Lp Pooling are among the various pooling functions available. However, the two most common types of pooling are as follows:

- Max Pooling: As the filter moves across the input, it selects the pixel with the highest value and sends it to the output array. It is worth noting that this method is more frequently utilized than average pooling.
- Average Pooling: As the filter moves across the input, it computes the average value within the receptive field and sends it to the output array.

Fully-connected layer

The fully-connected layer is arranged in a way that is similar to the traditional neural network's neurons. Every node in a fully connected layer is directly connected to every node in both the previous and the next layer.

2.1.1.1 Convolutional Neural Networks (CNNs) Architectures

Due to its ability to automatically learn hierarchical features from data and capture local patterns and relationships, CNN is well-suited for image-related tasks [41]. Therefore, we chose to use CNN in this paper. Over the last ten years, several CNN architectures have been introduced. Among these are LeNet, AlexNet, ResNet, GoogleNet/InceptionNet, Visual Geometry Group (VGG), DenseNet, and Xception. The architectures used in this study are explained below.

Visual Geometry Group (VGG)

The VGG model is a Convolutional Neural Network (CNN) architecture with numerous layers. Its ”deep” nature is attributed to the high number of layers, with VGG-16 and VGG-19 containing 16 and 19 convolutional layers, respectively. This groundbreaking architecture serves as the foundation for cutting-edge object recognition models and is applied in a variety of fields, including computer vision, speech recognition, machine translation, medical imaging, and robotics. The VGG design implements small 3x3 filters, and all hidden layers use the Rectified Linear Unit (ReLU) activation function. The model boasts a 224x224 image input size and features three fully connected layers, with the first two containing 4096 channels each, and the third containing 1000 channels, representing each class. VGG16 and VGG19 are the most widely used models, with VGG19 having three additional convolutional layers than VGG16. VGG16 has 138 million parameters, while VGG19 has 144 million parameters [42]. Figure 2.1 shows the structure of the network.

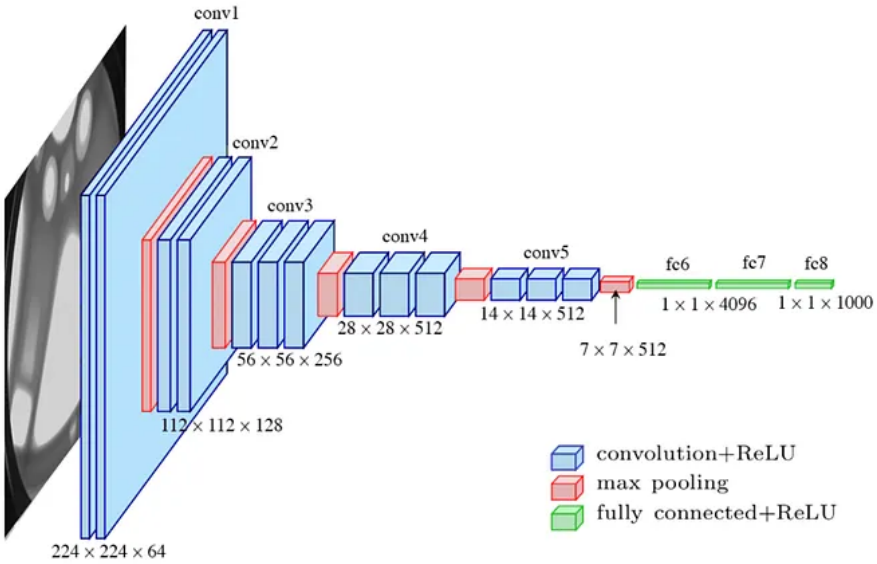

Figure 2.1: The basic structure of VGGNet block

GoogleNet/InceptionNet

In the context of CNN, a novel inception block (module) concept was proposed that combines multiple-scale convolutional transformations through the merge, transform, and split functions for feature extraction. This architecture employs filters of varying sizes (5×5, 3×3, and 1×1) to capture channel and spatial information at different ranges of spatial resolution. The GoogLeNet's traditional convolutional layer is replaced by small blocks, each replaced with a micro-neural network. The concept of merge, transform, and split from GoogLeNet is utilized to address the issue of different learning types of variants existing in a similar class of several images. The purpose of GoogLeNet is to improve the efficiency of CNN parameters and enhance learning capacity while regulating computation by inserting a 1×1 convolutional filter, as a bottleneck layer, before using large-size kernels. GoogleNet utilizes sparse connections to overcome the problem of redundant information and decrease costs. It should be noted that only some input channels connect to some of the output channels. By using a Global Average Pooling (GAP) layer as the end layer, the density of connections is decreased, resulting in a significant decrease in the number of parameters from 40 to 5 million. Additional regularity factors include the use of RmsProp as an optimizer and batch normalization. Furthermore, GoogleNet introduced the idea of auxiliary learners to speed up the convergence rate. However, its main shortcoming is its heterogeneous topology, which requires adaptation from one module to another. Other shortcomings include representation jam, which substantially decreases the feature space in the following layer, resulting in valuable information loss [42]. Figure 2.2 shows the structure of the network.

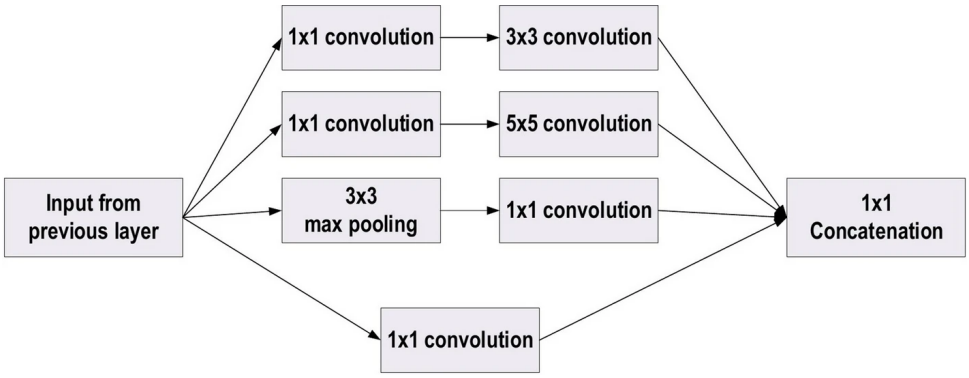

Figure 2.2: The basic structure of InceptionNet block

Xception

Xception is characterized by its extreme inception architecture. Its main idea revolves around depthwise separable convolution. The inception block of Xception is wider and involves exchanging a single dimension (3×3) followed by a 1×1 convolution to reduce computational complexity. By utilizing the decoupling channel and spatial correspondence, the Xception network achieves extra computational effectiveness. It first applies 1×1 convolutions to map the convolved output to the embedding short dimension and then performs k spatial transformations. Here, k represents the width-defining cardinality obtained via the transformation number in Xception. To simplify computations, each channel is distinctly convolved around the spatial axes in Xception. These axes are subsequently used as the 1×1 convolutions (pointwise convolution) to perform cross-channel correspondence. The 1×1 convolution is used to regulate the depth of the channel in Xception. The traditional convolutional operation in Xception involves many transformation segments equivalent to the number of channels. On the other hand, Inception uses three transformation segments, while the traditional CNN architecture only uses a single transformation segment. However, the suggested transformation approach in Xception achieves extra learning efficiency and better performance without minimizing the number of parameters [42]. Figure 2.3 shows the structure of the network.

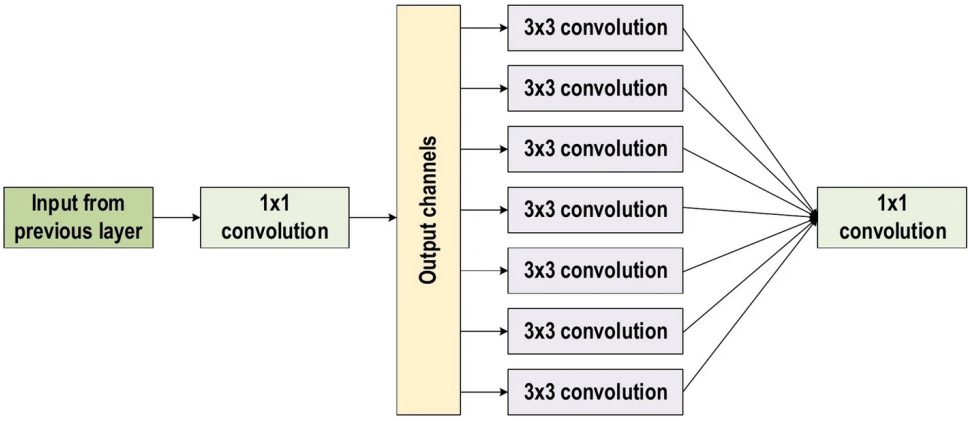

Figure 2.3: The basic structure of Xception block

2.2 Transfer Learning

Transfer learning enhances a learner's performance in one domain by utilizing knowledge from a related domain [43]. Recent research indicates that deep convolutional neural networks (CNNs) are highly effective at solving various classification problems. However, training deep CNN requires a vast amount of data to achieve optimum performance. This can be a significant challenge since collecting and labeling large datasets can be time-consuming, expensive, and sometimes unfeasible.

To address this challenge, researchers have turned to transfer learning. This technique utilizes knowledge from pre-trained models on existing tasks to overcome data scarcity in new tasks. By "fine-tuning" these pre-trained models on smaller, task-specific datasets, transfer learning enables deep CNN to achieve remarkable accuracy even when data is limited. This not only saves time and resources but also unlocks the potential of deep learning for applications where large-scale data collection is impractical. Transfer learning acts as a bridge, allowing deep CNN to excel in data-scarce scenarios and unlock their full potential for a wider range of problems [44, 45].

2.3 Ensemble Model

Machine learning has evolved beyond the use of single models to achieve robust and accurate predictions. Ensemble models combine the strengths of multiple individual models to achieve better performance than their constituent parts. This approach leverages the diversity in the architectures, learning algorithms, or training data to overcome the limitations of single models. The result is a better generalization, reduced variance, and improved accuracy across different tasks [46]. Ensemble techniques can be broadly classified: Bagging, Boosting, Stacking and Voting.

Bagging

Bagging, which is also known as bootstrap aggregating, is a simple and effective method for building a set of independent models. This technique involves training each model on a random sample of instances from the original dataset, with replacement. To ensure that each model has a sufficient number of instances, the sample usually contains the same number of instances as the original dataset. The final prediction for an unseen instance is determined by the majority voting of the predictions made by each model. While some instances may appear multiple times when training the same model, others may not be included at all. Because the models are independently trained, bagging can be implemented in a parallel manner by training each model using different computational units [47].

Boosting

Boosting is a powerful ensemble method that uses the knowledge gained from previous predictor mistakes to improve future predictions. The approach involves sequentially combining multiple weak base learners, with each learner building on the errors of the previous one to create a more accurate predictive model. Ultimately, this results in a single strong learner that significantly enhances the predictability of models.

Stacking

Stacking is a powerful technique that involves combining multiple classification or regression models into a single ensemble. It offers a unique approach to problem-solving, allowing us to explore a wide range of possible models for a given task [48].

In ensemble models, stacking is also known as stacking generalization. It enables us to create a new model that combines predictions from two or more previously trained models. By merging existing sub-model predictions, we can leverage the strengths of each model and create a more accurate overall prediction. We can use a simple linear approach such as simple voting to merge the sub-model predictions or use a weighted sum using linear or logistic regression. Stacking is a powerful tool that can help us build more accurate and robust models for a wide range of applications.

Voting The voting ensemble model is an effective ensemble learning technique that enhances overall performance by combining predictions from multiple individual models. This method involves aggregating predictions from base models through a voting process, ultimately selecting the most probable outcome as the final prediction. There are two commonly used types of voting in ensemble techniques: hard voting and soft voting.

- **Hard Voting:** also known as majority voting, is a classification method that involves combining the predictions made by several base models and selecting the class with the most votes as the final prediction. In hard voting, the final decision is based purely on the majority vote, without considering the confidence or reliability of each base model.

- **Soft Voting:** also known as weighted voting, is a technique used in machine learning to make predictions by considering the probability scores of each base model for each class. The method computes the weighted average of these probabilities to make the final prediction. The main idea behind soft voting is to integrate the predictions from multiple models to produce a more accurate and reliable outcome. To achieve this, soft voting calculates the average probability of each class and then selects the winner by choosing the class with the highest weighted probability. Overall, soft voting is an effective way to enhance the performance of machine learning models by leveraging the collective knowledge of multiple models.

2.4 Out-of-Distribution(OOD) Detection

In machine learning, the process of Out-of-Distribution (OOD) detection is crucial. Out-of-Distribution (OOD) detection, also called out-of-distribution modeling or anomaly detection, involves identifying data points or samples that significantly differ from the training data on which a model was developed.

During the training phase, models are typically trained on a specific set of data that represents the target distribution or class of interest. However, in real-world scenarios or deployment, the model might encounter data that falls outside this distribution or differs significantly from the training data. This can lead to unreliable predictions or erroneous outcomes since the model may not have learned to handle such out-of-distribution data during training.

The primary goal of Out-of-distribution detection is to identify instances where the incoming data does not fit the pattern or characteristics of the data seen during training [49]. Once the system identifies these OOD instances, it can take appropriate actions, such as flagging them for further review, rejecting them to prevent potential errors, or handling them separately to maintain the integrity and performance of the model.

2.5 Distribution Shift

Distribution shift is a term used to describe the situation where the statistical properties of the data that is fed into a machine learning model change between the training and testing stages. When a model is being trained, it learns to recognize patterns and relationships within a particular dataset, so that it can make accurate predictions on unseen data during testing. However, real-world scenarios often involve significant differences between the data seen during training and the data encountered during testing. This shift in data distribution can occur due to various factors such as changes in data collection methods, environmental conditions, demographics, or measurement devices. As a result, the model's performance may decrease as it struggles to adapt to the new or unseen data distribution. This can lead to inaccurate predictions and reduced generalization abilities [50, 51]. There are several types of distribution shift:

1. **Covariate shift:** occurs when there is a change in the distribution of input features (covariates) between the training and testing datasets used to build and deploy a machine learning model. This means that the relevant characteristics of input features for medical diagnosis may differ between the dataset used for training and the real-world data encountered during deployment.

   To illustrate, a machine learning model is trained to diagnose a specific medical condition using patient data from a particular hospital. The training dataset could include patient demographics, medical history, and various diagnostic test results. However, when deploying the model in a different hospital or clinical setting, the distribution of these covariates may vary. Covariate shift could be caused by differences in patient populations, healthcare practices, or available diagnostic tests.

2. **Label shift:** Occurs when the distribution of target variables, or labels, changes between the training and testing datasets. It differs from covariate shift, which pertains to changes in the distribution of input features. Label shift specifically addresses variations in the distribution of the output or target variable.

   In the context of medical diagnosis, label shift can occur when the prevalence or distribution of different medical conditions changes between the dataset used for training a diagnostic model and the real-world data encountered during deployment. For instance, the frequency of rare diseases may differ between the training and testing datasets, or there may be changes in the diagnostic criteria over time.

3. **Concept shift:** The phenomenon of concept shift, also referred to as concept drift, arises when the relationship between the input features and the target variable (concept) changes over time or across different datasets. Unlike covariate shift which involves alterations in the distribution of input features, or label shift which involves changes in the distribution of target labels, concept shift is centered around modifications in the underlying concept or patterns within the data.

   In the context of medical diagnosis, a concept shift may occur when the factors influencing a particular medical condition undergo changes over time or across different patient populations. For instance, advancements in medical research or fluctuations in environmental factors may cause shifts in the diagnostic criteria for a disease. Besides, the prevalence of risk factors or the efficacy of certain diagnostic tests may also evolve.

2.6 Summary

In this chapter, we explore the techniques employed in our research, with a specific focus on those relevant to our field of study. It emphasizes the application of deep learning, specifically convolutional neural networks (CNNs), which serve as the foundation for our proposed ensemble method. Our investigation includes a detailed overview of various CNN algorithms, including transfer learning and ensemble modeling, to offer insights into their potential for our chosen approach.

Chapter 3

Literature Review

In recent years, various machine-learning algorithms have been used to detect and classify breast tumors from histopathology images. This chapter looks into these methods for classifying different types and subtypes of breast tumors, as well as staging. It also discusses the effects of data leakage on breast cancer classification.

3.1 Breast Tumor Classification

In order to differentiate between benign and malignant breast tumors, Osareh et al. [27] investigated the use of three machine learning classifiers (Support Vector Machine (SVM), k-nearest neighbors (KNN), and Probabilistic Neural Network (PNN)) in conjunction with three feature selection techniques (Signal to Noise Ratio (SNR) feature ranking, Sequential Forward Selection (SFS), and Principal Component Analysis (PCA)). They assessed their models using two widely used benchmark datasets for breast cancer research, and the SVM classifier produced an accuracy of 98.80% on the FNAB dataset and 96.33% on the gene microarray dataset for breast cancer detection.

A study by Doyle et al. [28] extracted 3,400 image features from a database of 48 breast biopsy tissue (30 cancerous and 18 benign images). They extracted a set of graph and texture based features to capture the discriminating characteristics of the tissue patterns in each image. Spectral clustering was utilized to decrease the dimensionality of the feature set, and a SVM classifier was employed to differentiate between images with low and high grades of malignancy, as well as benign and malignant images. Using texture-based characteristics, their method obtains a 95.8% accuracy rate in differentiating cancer from non-cancer and a 93.3% accuracy rate in differentiating high from low grades of cancer based on architectural elements.

A dataset called BreakHis [18] with 7,909 breast cancer (BC) histopathology images, separated into benign and malignant tumors, acquired from 82 individuals at varying magnification factors, was presented by A. Spanhol et al. [18]. The dataset comprises four distinct benign breast tumor histological types: adenosis (A), fibroadenoma (F), phyllodes tumor (PT), and tubular adenoma (TA); additionally, there are four malignant tumor (breast cancer) types : ductal carcinoma (DC), lobular carcinoma (LC), mucinous carcinoma (MC), and papillary carcinoma(PC). The authors also present a baseline pattern recognition system that uses Threshold Adjacency Statistics, Local Binary Patterns (LBP), Completed LBP (CLBP), Local Phase Quantization (LPQ), Grey-Level Co-occurrence Matrix (GLCM), and a keypoint descriptor called ORB for feature extraction to distinguish between benign and malignant tumors. Four distinct classifiers were employed for the classification process: Random Forests of Decision Trees, Quadratic Linear Analysis (QDA), SVM, and 1-Nearest Neighbor (1-NN). Their system's accuracy rates range from 80% to 85%, depending on the image magnification factor.

Several other researchers have classified breast cancer as benign and malignant tumors based on histopathological images by employing traditional machine-learning techniques (SVM, KNN, Naive Bayes, Decision Tree) [29, 30, 52, 53, 54].

3.2 Breast Tumor Sub-Type Classification

Various researchers worked on classifying benign and malignant with their subtypes utilizing the BreakHis dataset, which was introduced by A. Spanhol et al. [18]. A study by Han et al. [31] on a BreakHis [18] dataset presents an automated deep-learning model for multi-class classification of breast cancer from histological images. The proposed model dubbed the class structure-based deep convolutional neural network (CSDCNN), achieved an average accuracy of 93.2%, demonstrating the method's strength in providing a helpful tool for breast cancer multi-class classification in clinical settings. The article highlights the challenges in obtaining precise multi-class classification due to the wide range of high-resolution image appearances and the subtle differences in multiple classes, as well as the importance of automated multi-class classification of breast cancer from histopathological images for clinical diagnosis and prognosis. They claimed that the proposed CSDCNN model addresses these issues by making use of feature space distance limitations and hierarchical feature representation.

Research conducted by H.Motlagh et al. [32] centers around using deep learning architectures, such as Inception and ResNet, to differentiate microscopic cancerous imaging, with a specific focus on breast cancer detection and classification of its subtypes. The study utilized tissue micro-arrays (TMAs) as training samples and fine-tuned pre-trained deep neural networks to classify various types of cancer. The ResNet V1 50 model achieved an average accuracy of 99.8% for breast, bladder, lung, and lymphoma cancers. Additionally, the ResNet V1 50 and ResNet V1 152 models achieved accuracies of 94.8% and 96.4%, respectively, in categorizing benign and malignant related sub-types.

Parvin et al. [33] evaluate the performance of five CNN architectures, namely LeNet-5, AlexNet, VGG16, ResNet-50, and Inception-v1, in classifying histological images associated with breast cancer. The BreaKHis [18] dataset, which contains microscopic biopsy images of benign and malignant tumors at different magnification factors, was used to evaluate model performance based on test accuracy, the area under the curve (AUC), precision, recall, and f1-score. The Inception-v1 network outperformed the other designs in the study, achieving the highest test accuracy, Area Under the Curve (AUC), precision, recall, and f1-score at different magnification factors. The study highlights the growing prevalence of breast cancer and emphasizes the need for accurate and efficient prediction systems. It also addresses the importance of early breast cancer detection and the limitations of traditional approaches. It draws attention to the potential of convolutional neural networks in medical image analysis and emphasizes the significance of selecting the best architecture to achieve optimum results in the histopathological image classification of breast cancer.

3.3 Breast Tumor Stage Classification

Breast cancer can be staged according to how easily it can spread across the body and affect distant organs and tissues. A hybrid convolutional and recurrent deep neural network was presented by Yan et al. [24] for classifying breast cancer staging as invasive and in situ carcinoma, as well as benign and normal. The study aimed to develop an automated and accurate histopathological image analysis approach, with a focus on the essential classification task for in-depth research on breast cancer diagnosis. The authors have released a dataset containing 3771 histopathological images for use in scientific research. According to the experimental results, the average accuracy for the four-class classification task is 91.3%. However, the research was limited to identifying breast stages rather than the specific type of breast cancer. To address this gap, some researchers worked on classifying breast cancer type along with its staging. Wetstein et al. [34] developed and evaluated an automated deep-learning system for grading DCIS in breast histopathology images. DCIS is a non-invasive breast cancer that can progress to acIDC. Gupta et al. [35] provide a comprehensive computer-aided diagnosis (CAD) system for classifying IDC using a CNN model they built from scratch (ConvNet-A, ConvNet-B, and ConvNet-C), which they verified against four machine learning models (SVM, KNN, Random Forest (RF), and Logistic Regression (LR)). The paper highlights the need for a computer-aided diagnosis system to support pathologists in detecting breast cancer, aiding in early-stage detection, and improving survival rates.

3.4 Data Leakage in Digital Pathology

The term ”data leakage” describes the usage of non-training datasets for model selection or training. Leakages typically happen when data from test, validation and/or training sets share indirect information, producing unduly optimistic conclusions [55]. Bussola et al. [55] conducted a study on reproducibility issues in bioinformatics. Their focus was on the impact of data leakage in digital pathology and highlighted the prevalence of unrepeatable research papers in bioinformatics due to methodological or clerical errors. The study identified the link between lack of reproducibility and inaccuracies in managing batch effects, small sample sizes, and flaws in experimental design. Specifically, the paper explored the potential impact of data leakage on machine learning algorithms and the effect of data partitioning strategies on the training of backbone architectures in the context of histological data. The study demonstrats that predictive scores can be inflated by up to 41% when tiles from the same subject are used in both training and validation sets. The study provides detailed results on the impact of data partitioning protocols and feature embeddings on image classification in digital pathology based on experiments conducted on three public datasets. Additionally, it compares its findings with those of comparable works.

3.5 Summary

This chapter presents a comprehensive literature review of existing works that focus on classifying breast cancer. We explored the various methods and datasets employed for the classification of breast cancer, while also discussing data leakage and its impact on accuracy. Despite extensive research in the field of breast cancer detection and classification, there is still a gap in classifying subtypes of ILC and LCIS. This issue is further worsened by the lack of dedicated datasets for these subtypes of breast cancer. To the best of our knowledge, no previous research has utilized existing labels as a primary source to generate new labels for breast cancer classification. In this work, we propose to resolve the lack of a dataset by extracting the required dataset from existing breast cancer datasets and utilizing it to classify ILC and LCIS subtypes.

Chapter 4

Methodology

In this section, we propose a methodology for developing new labeled breast cancer datasets and discuss the classification schemes employed on these newly developed datasets.

The proposed method of the architecture is shown in Figure 4.1. The research process starts with two dataset selections, namely the BreakHis [18] and Yan et al. [24]. These datasets are labeled and then divided into training and testing sets to prepare them for machine learning and testing. The data then goes through preprocessing and augmentation to enhance its quality and potentially improve the model's performance. The approach involves using an ensemble model that combines multiple models to achieve the best prediction accuracy. Furthermore, the model's predictions are evaluated on the two datasets to generate a new dataset. Lastly, a new dataset is generated and classified.

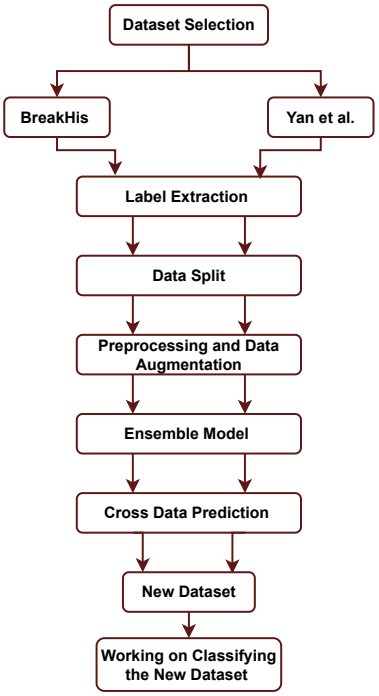

Figure 4.1: The proposed model architecture for the creation of a new dataset and classification system for breast cancer

4.1 Dataset

This work utilized two complementary public datasets i.e. the BreakHis [18] dataset that provided tumor subtypes and the Yan et al. [24] dataset which offered detailed information on tumor stage. The BreakHis [18] dataset comprises microscopic biopsy images of both benign and malignant breast tumors. The dataset consists of 7909 images, collected from 82 anonymous patients of the Pathological Anatomy and Cytopathology (P&D) Lab in Brazil. The images within the datasets are captured in the 3-channel RGB color space (TrueColor with 24-bit depth, 8 bits per channel) and saved in the PNG format. Also, the BreakHis [18] dataset offers four magnification options (40x, 100x, 200x, and 400x) and maintains a consistent dimension of 700 x 460 pixels. The spectrum of breast tumors encompasses four distinct histopathological types for both benign and malignant forms. Benign tumors include adenosis (A), fibroadenoma (F), phyllodes tumor (PT), and tubular adenoma (TA). Malignant types include ductal carcinoma (DC), lobular carcinoma (LC), mucinous carcinoma (MC), and papillary carcinoma (PC). Table 4.1 shows the details of the BreakHis [18] dataset. Structurally, the BreakHis [18] dataset is organized around individual patients and offers images of each patient's data at four magnification factors (40x, 100x, 200x, and 400x) for detailed examination.

Table 4.1: Histopathological image distribution of the BreakHis dataset divided by magnification and class.

Class	**Subclass**	**Magnification factors**				**Total**
		40X	**100X**	**200X**	**400X**	
Benign	A	114	113	111	106	444
	F	253	260	264	237	1014
	TA	109	121	108	115	453
	PT	149	150	140	130	569
Malignant	DC	864	903	896	788	3451
	LC	156	170	163	137	626
	MC	205	222	196	169	792
	PC	145	142	135	138	560
Total		1995	2081	2013	1820	7909

The second dataset was developed by Yan et al. [24] and Peking University International Hospital collaborated on a unique dataset of 3,771 high-resolution (2048x1536 pixels) breast pathology images. Each image, meticulously stained with hematoxylin and eosin (H&E) for detailed analysis, and is fully anonymized for patient privacy. All images maintain consistent capture conditions (100x or 200x magnification) and are categorized according to the diagnosed cancer type: normal, benign, in situ carcinoma, or invasive carcinoma. Additionally, the images are captured in the standard 3-channel RGB color space and saved in the versatile TIF format. Table 4.2 summarizes the image distribution.

Table 4.2: Histopathological image distribution of Yan et al. divided by class.

Class	Normal	Benign	In situ carcinoma	Invasive carcinoma	Total
	299	1106	1066	1300	3771

In our thesis, we utilized the BACH dataset [20], which has a similar data distribution to the Yan et al. dataset and includes regions of normal tissue, benign tumors, in situ carcinoma, and invasive carcinoma. The microscopy images are in TIFF format, with an RGB color channel and dimensions of 2048 x 1536 pixels [26]. Table 4.3 summarizes the image distribution.

Table 4.3: Histopathological image distribution of BACH divided by class.

Class	Normal	Benign	In situ carcinoma	Invasive carcinoma	Total
	100	100	100	100	400

4.1.1 Label Selection

To examine the correlation between tumor type and staging, two specific labeled classes were selected from each dataset:

- **From the BreakHis [18] dataset:** Images labeled as ductal carcinoma and lobular carcinoma were extracted. These two types represent the most common and aggressive forms of invasive breast cancer, making them valuable for targeted analysis.
  - **Ductal Carcinoma:** Primarily exists in both In situ and Invasive forms.
  - **Lobular Carcinoma:** Mostly presents in the In situ stage, though Invasive lobular carcinoma also exists.

- **From the Yan et al. [24] dataset:** Images labeled as In situ carcinoma and Invasive carcinoma were chosen. This selection captures both early and advanced stages of tumor development, providing insights into progression and potential diagnostic markers.

The summary of the selected tumor types and stage information from the two datasets is provided in Table 4.4. The table presents the count of datasets from both sources, thereby offering a detailed overview of the data.

Table 4.4: The extracted labels

Dataset	Labels	Labels Total	Total
BreakHis [18]	Ductal Carcinoma	3451	4077
	Lobular Carcinoma	626	
Yan et al. [24]	In situ carcinoma	1066	2,366
	Invasive carcinoma	1300	

4.2 Data Split

In order to uphold the reliability of our findings and mitigate the risk of any artificial inflation of performance metrics, a stringent data partitioning strategy known as Patient-Wise (PW) splitting has been implemented [55]. The patient-wise split is a well-known data partitioning technique frequently utilized in medical research, particularly for studying data with significant intra-subject variability. This technique treats patients as the main unit of analysis, and their data is kept together throughout all phases of analysis, from training to validation and testing [55]. Ensuring that:

- **No tile from a patient resides in both the training and testing sets:** No tile from a patient resides in both the training and testing sets. This strict separation eliminates the risk of models inadvertently ”memorizing” patient-specific features rather than capturing generalizable patterns, leading to misleading high-performance estimates.
- **To address the class imbalance, we adopt a stratified sampling approach:** When dividing the dataset, it is ensured that each class has the same size of samples in both the training and testing sets. This prevents models from being biased towards majority classes, resulting in a more balanced learning process.

A patient-wise split method was used for the BreakHis [18] dataset to create separate training and test sets of 80% and 20%, respectively, to ensure that the data from individual patients remained consistent throughout the analysis. However, a challenge was encountered with the Yan et al. [24] dataset as there was no clear patient information available. To overcome this, a distribution shift was introduced for the testing dataset, and the entire Yan et al. [24] dataset was used for training. To prevent data leakage, a different dataset, BACH [21], was designated for testing. This maintained the integrity of the evaluation process and ensured that the model's performance was rigorously assessed in a way that was resilient to the unique characteristics of each dataset.

4.3 Preprocessing and Data Augmentation

Deep learning has become an increasingly popular method for image classification, including the identification of breast cancer in medical images. However, it's crucial to understand the importance of preprocessing in this context. Preprocessing involves preparing the images to be fed into the model by resizing them to a uniform size, adjusting their lighting, and ensuring that the pixel values are within the appropriate range. This process helps the computer model to learn and recognize patterns more effectively [56, 57]. In this study all images were resized to a unified dimension of 224x224 pixels and image pixel values were normalized to the range [0, 1] using a scaling factor of 1/255.

Another important aspect of preprocessing is data augmentation, which entails applying various changes to the images to expand the dataset and improve the model's robustness. This can include techniques like rotating, shifting, zooming, flipping, and adjusting brightness [58]. By incorporating data augmentation, the model becomes better equipped to handle variations in real-world scenarios, ultimately enhancing its accuracy and generalizability. Specific techniques employed include:

- **Random Rotations:** Images were randomly rotated up to 10 degrees, simulating slight camera angle variations during image acquisition.
- **Horizontal Shifts:** Images were randomly shifted horizontally by up to 20% of their width, mimicking potential misalignments during tissue preparation or scanning.
- **Vertical Shifts:** Similar to horizontal shifts, images were also shifted vertically by up to 20% of their height to account for potential tissue positioning variations.

- **Zooming:** Random zooming (in or out) by up to 20% was implemented to simulate magnification changes during microscopy or digital imaging.
- **Horizontal Flips:** Images were randomly flipped horizontally to augment the training data and improve the model's ability to learn features regardless of tissue orientation.
- **Brightness Adjustments:** To account for potential lighting variations or inconsistencies during image capture, brightness was randomly adjusted within a range of 20% darker to 20% brighter.

Table 4.5: Parameters of data augmentation.

Parameters of Image Augmentation	Values
Rotations Range	-10,10
Height Shifts Range	-0.2,0.2
Width Shifts Range	-0.2,0.2
Zooming Range	0.8,1.2
Random Horizontal Flips	True
Brightness Adjustments Range	0.2-1.2

To ensure the reliability of the deep learning pipeline for breast cancer classification, a technique called stratified k-Fold cross-validation is used. This method helps to address issues related to class imbalances, which is a common concern in imbalanced datasets.

The stratified k-fold cross-validation technique builds upon the traditional k-fold cross-validation approach, where data is partitioned into k subsets (or folds), and the model is trained and evaluated k times, with each fold acting as the test set once.

However, standard k-fold cross-validation may not be ideal for imbalanced datasets, where some classes have considerably fewer instances than others. The reason is that the class distribution in each fold may not be representative. Stratified k-fold cross-validation resolves this concern by ensuring that each fold preserves the same class distribution as the original dataset.

By utilizing stratified k-fold cross-validation, deep learning models can be effectively used to classify breast cancer images accurately. This can assist in the diagnosis and treatment of this disease.

4.4 Ensemble Model

Research studies [59, 60, 61] have shown that incorporating ensemble learning in deep learning systems can enhance their generalization and enable more accurate predictions. Ensemble learning integrates the outputs of multiple classifiers, either at the final or intermediate stages, to overcome the limitations of a single classifier and thereby enhance the resilience of the classification framework. Based on these research, the Ensemble model was incorporated to improve the prediction performance of the model on the cross-data.

This section provides a detailed elaboration on the ensemble model developed, as shown in Figure 4.2 which have three distinct phases. Phase one highlights the specific deep learning models utilized in the initial experimentation. Phase two showcases the selection of the best performing models identified through the performance analysis conducted in phase one. Finally, phase three illustrates the implementation of a soft voting ensemble model, which leverages the strengths of the individual models to achieve enhanced predictive capabilities. Soft voting is a technique where each model in an ensemble generates probability predictions for every class, which are then averaged across all models to arrive at the final prediction. By considering the confidence levels of each model in its predictions, this approach offers a more subtle and probabilistic decision-making process. Ultimately, the ensemble's final predicted class is the one with the highest average probability.

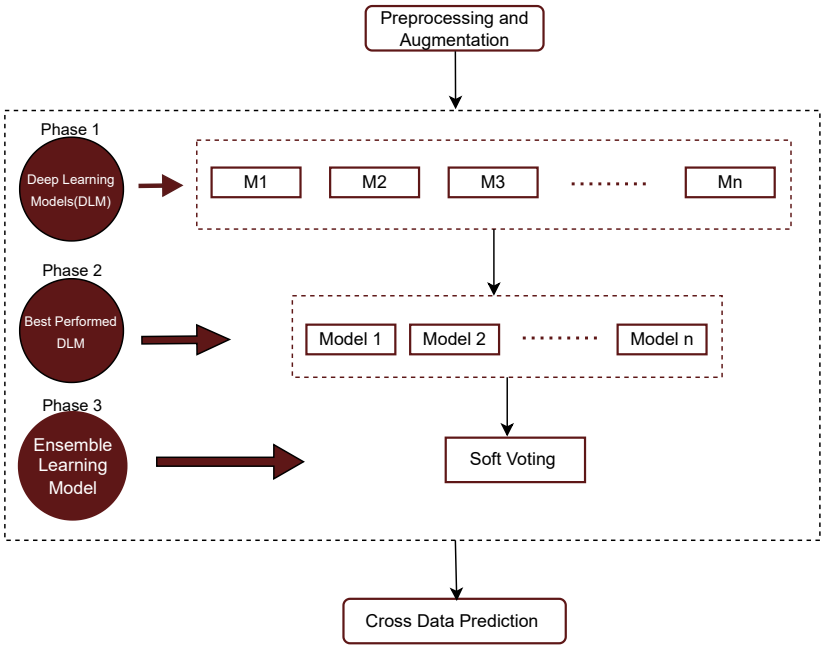

Figure 4.2: Architecture of the proposed ensemble model

4.4.1 Deep Learning Models Selection

The process of selecting the most suitable Deep Learning Model (DLM) for breast cancer classification was dedicated to the initial stage. This critical step involved evaluating four different models - VGG16, VGG19, Inception, and Xception - based on their performance on the BreakHis [18] dataset. The dataset was divided into a training set (80%) and a test set (20%) using a patient-wise split to prevent data leakage. The models were designed to incorporate a 5-fold stratified cross-validation, which proved to be highly effective during experiments on the 80% training set. This approach ensured that both ductal and lobular carcinoma cases were evenly represented in the training process. The goal was to create a fair training process for both classes. This method led to more dependable model evaluations, especially when dealing with variations in data size between the two cancer types. Data augmentation techniques described in Table 4.5 were applied during training. Each model was trained with 40x, 100x, 200x, and 400x magnification factors separately for 50 epochs, saving its weights. Finally, the unseen 20% test set was used to generate final predictions for each trained model. The model with the highest accuracy on the test set was selected based on its ability to generalize to unseen data.

4.4.1.1 Transfer Learning

The breast cancer classification method uses a transfer learning technique to improve its performance. The method utilizes four pre-trained models trained on the ImageNet dataset, which can leverage the valuable features already learned from millions of natural images. This approach helps extract essential information from medical images despite having limited data and provides a strong foundation for the task. After that, the final layers of the model are fine-tuned using the breast cancer dataset, which saves time and allows the model to learn crucial cancer-specific features despite the limited dataset size. In conclusion, employing transfer learning from ImageNet combined with a fine-tuning process significantly enhances the model's ability to accurately classify breast cancer images [62, 63].

- **ImageNet:** Is a massive library of visual knowledge designed to drive advancements in image recognition [45]. It contains over 14 million images that have all been meticulously annotated to identify the objects depicted in them. The database is organized into over 20,000 categories, with each category typically containing multiple images.

4.4.1.2 Hyperparameter Tuning

The hyperparameter tuning process is essential for achieving optimum performance and adaptability in convolutional neural networks (CNNs). It involves carefully adjusting various control knobs, including model architecture, learning process, and regularization, to ensure alignment with the specific task at hand [59]. Some of the crucial hyperparameters that require meticulous fine-tuning in CNN:

- **Learning rate:** This hyperparameter plays a significant role in model convergence by determining the step size during gradient descent. It directly impacts the speed and accuracy of the model's learning process.
- **Number of filters:** Filters are the building blocks of feature extraction, and their precise calibration is crucial for capturing rich visual patterns. The number of filters determines the model's ability to identify intricate details.
- **Kernel size:** The receptive fields of filters, known as kernel size, determine the extent of spatial information they encompass. Careful tuning is necessary to match the scale of relevant features in the images.
- **Number of layers:** The depth of a CNN, determined by the number of layers, affects its ability to learn hierarchical representations. Adjusting this hyperparameter thoughtfully is essential for achieving optimum complexity.
- **Batch size:** This hyperparameter determines the volume of data processed in each training iteration, balancing computational efficiency with learning speed and stability.
- **Regularization techniques:** Techniques such as dropout and L1/L2 regularization are crucial for preventing overfitting, ensuring that the model doesn't memorize training data at the expense of generalization.

The values of the hyperparameters with which models performed well are provided in Table 4.6.

Table 4.6: Hyperparameters used in the models for the BreakHis dataset.

Hyperparameters	With Data Augmentation			
	VGG16	**VGG19**	**Inception**	**Xception**
Train approach	5-fold stratified cross validation	5-fold stratified cross validation	5-fold stratified cross validation	5-fold stratified cross validation
Optimizer	SGD/ Adam	SGD	SGD	SGD/ Adam
Loss function	binary crossentropy	binary crossentropy	binary crossentropy	binary crossentropy
Batch size	16	16	16	16
Activation function	softmax	softmax	softmax	softmax
Dropout	0.3	0.3	0.3	0.3
Epoch	50	50	50	50
Number of nodes in output layer	2	2	2	2
Learning Rate	0.0001	0.0001	0.0001	0.0001

4.4.2 Best Performing DLM Selection

In the initial testing phase, we evaluated the performance of different models using the BreakHis [18] dataset. This involved a thorough assessment of how well each model performed across various magnification levels within the dataset. Based on the results of these evaluations, this phase identifies and selects the top-performing models for further analysis and development. Our selection criteria prioritized models that showed high accuracy and robustness in handling the diverse magnification levels present in the BreakHis [18] dataset.

4.4.3 Ensemble Learning Models

On the last stage of the analysis, various ensemble techniques were carefully evaluated, and the soft voting method was ultimately decided upon for implementation. This technique was chosen as the most efficient approach for combining the predictions of the best-performing models, and it contributed to the improvement of the overall accuracy and reliability of the final results. The evaluation process involved the careful testing and comparison of different ensemble methods, and the soft voting approach was determined to be the best choice for our specific needs.

4.5 Cross Data Prediction

Figure 4.3 outlines a methodology to generate a new dataset by utilizing the BreakHis [18] and Yan et al. [24] datasets through a two-phase cross-data prediction approach. In the first phase, the ensembled model is trained on the BreakHis [18] dataset, which is intended to classify the Ductal and Lobular breast cancer subtypes. This ensemble model is then used to predict outcomes on the Yan et al. [24] dataset, which contains Insitu and Invasive breast cancer staging. The aim of this phase is to transfer the knowledge gained from the BreakHis [18] dataset to make accurate predictions on the Yan et al. [24] dataset. By combining the knowledge gained from the Yan et al. [24] labels and the predictions, labels for IDC, DCIS, ILC, and LCIS are extracted.

In the second phase, the process is reversed. The ensemble model is trained on the Yan et al. [24] dataset, which includes Insitu and Invasive breast cancer staging. This trained model is then used to predict outcomes on the BreakHis [18] dataset containing the Ductal and Lobular breast cancer subtypes. This reverse prediction process aims to apply the insights and patterns learned from the Yan et al. [24] dataset to the BreakHis [18] dataset. By doing so, IDC, DCIS, ILC, and LCIS labels are extracted from the BreakHis [18] dataset. This methodology ensures that the knowledge from each dataset is cross-applied, allowing for an understanding and utilization of both datasets' unique characteristics.

Finally, the predictions from both phases are merged to create a new, enriched dataset extracted from the BreakHis [18] and Yan et al. [24] datasets, containing the IDC, DCIS, ILC, and LCIS labels. This new dataset benefits from the strengths and information of both original datasets, providing a more comprehensive resource for future analysis and research.

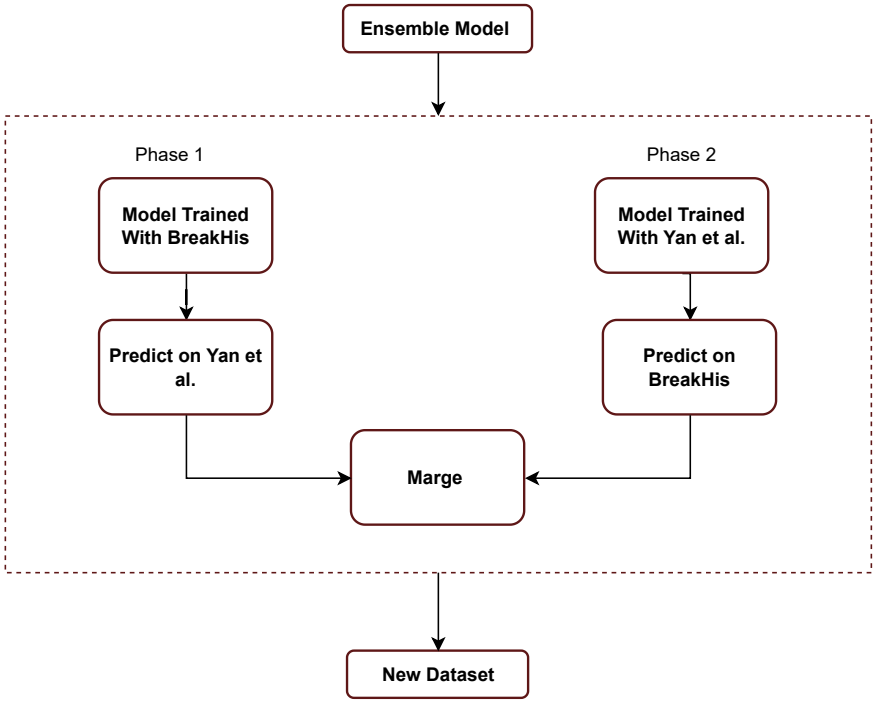

Figure 4.3: Diagram for the proposed cross-data prediction step

4.6 Classification on The New Dataset

This section outlines the method used to classify IDC, DCIS, ILC, and LCIS from the newly developed dataset using the methodology described earlier. The proposed method for classifying breast cancer using a new dataset is described in Figure 4.4. The classification process begins by splitting the dataset into training and testing sets using the early patient-wise splitting techniques. This ensures that the model is trained and tested on different patient data to avoid data leakage and improve the model's generalization performance.

After splitting the dataset, the preprocessing and augmentation techniques that were utilized on the BreakHis [18] and Yan et al. [24] datasets are applied to it. This step is critical as it improves the quality of the data and potentially the performance of the model.

The ensemble method, which was initially employed for training on the BreakHis [18] and Yan et al. [24] datasets, was subsequently utilized for classifying the data based on new labels.

Finally, the predicted results for the IDC, DCIS, ILC, and LCIS were obtained using the unseen test dataset. The results of this study are promising and indicate that the proposed method can classify breast cancer using the new dataset.

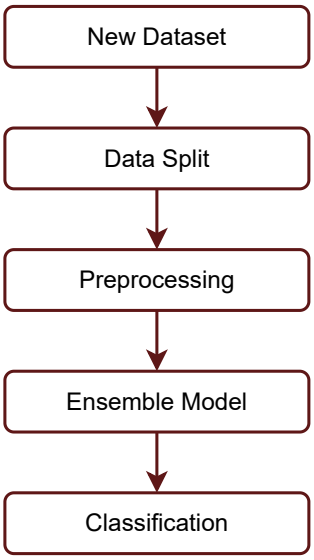

Figure 4.4: The proposed architecture for classification on the new dataset

4.7 Evaluation Metrics

The evaluation metrics employed in Deep Learning tasks play a critical role in achieving an optimized classifier. They are utilized in the standard data classification process, encompassing two main stages of training and testing. During the training stage, these metrics are employed to optimize the classification algorithm. In the context of classifying breast cancer subtypes or any other classification task, several commonly used evaluation metrics help to measure the effectiveness of the model. Here are detailed explanations of some of the key evaluation metrics:

Accuracy: Calculates the percentage of correctly predicted instances in a dataset by dividing the number of correct predictions by the total number of instances.

$$\text{Accuracy} = \frac{\text{number of correct classified} * 100}{\text{Total numbers of input samples}} = \frac{TP + TN}{TP + TN + FP + FN} \tag{4.1}$$

In a classification scenario, True Positives (TP) denote instances that were accurately predicted as positive, True Negatives (TN) represent instances that were correctly predicted as negative, False Positives (FP) indicate instances that were erroneously predicted as positive, and False Negatives (FN) signify instances that were mistakenly predicted as negative.

Precision: Measures the accuracy of positive predictions. It is the ratio of correctly predicted positives to total predicted positives.

$$\text{Precision } = \frac{TP}{TP + FP} \tag{4.2}$$

Recall (Sensitivity): Measures how well the model identifies positive instances. It is the ratio of true positives to the total number of actual positives in the dataset.

$$\text{Sensitivity } = \frac{TP}{TP + FN} \tag{4.3}$$

F1 Score: Measures the harmonic average between precision and recall. 4.4

$$F1_{\text{score}} = 2 \times \frac{\text{Precision } \times \text{ Recall}}{\text{Precision } + \text{ Recall}} \tag{4.4}$$

4.8 Summary

In this chapter, we have discussed the working principles and methodology used to create a new dataset for breast cancer from the existing dataset. We have also explained the evaluation strategy utilized to assess the new dataset, as well as our methodology for classifying IDC, DCIS, ILC, and LCIS breast cancer subtypes. In the next section, we will discuss the results obtained from our experiment.

Chapter 5

Result and Discussion

In this chapter, we will discuss the experiments carried out and the results obtained. The result of utilizing the ensemble model is presented. The comparison of the model results with those of pathology diagnoses is also discussed. Furthermore, we provide the classification results using the newly developed breast cancer dataset.

5.1 Experimentation Setup

All experiments were conducted using Python 3.9.16, TensorFlow 2.12.0 and Keras 2.12.0, installed on a Dell Precision 7920 Tower server. The server is equipped with an Intel(R) Xeon(R) Gold 6230R CPU, 64GB of RAM, NVIDIA RTX A4000 GPU and is running Ubuntu 22.04.2 LTS Server operating system.

5.2 Deep Learning Models Selection Result

5.2.1 The Models Performance When Trained with BreakHis Dataset

In Chapter 4, a brief explanation of the proposed methodology for extracting a new labeled dataset for breast cancer and the classification process undergone by the new dataset is provided. An experiment using four different Convolutional Neural Network (CNN) architectures, namely VGG16, VGG19, Inception, and Xception were conducted. The experiment was carried out on the BreakHis [18] dataset, which consists of instances at four different magnification factors—40x, 100x, 200x, and 400x.

In order to achieve optimum performance in the classification task, each model's performance was evaluated individually at every magnification factor on the BreakHis [18]. By having the models' performance assessed at different levels of magnification, the most effective models were able to be chosen, each of which was tailored to excel at a specific magnification factor.

A hyperparameter sweep was made for each CNN model, and the best parameters were found as has been given in Table 4.6. A consistent approach has been taken in our experiment by using the same hyperparameters for all models that are being considered. This choice has been deliberately made for several reasons. Firstly, it ensures a fair and unbiased comparison between the models can be made, as each one is trained under the same conditions with the same optimization settings. By maintaining uniformity in hyperparameters, any differences in performance that are observed can be more confidently attributed to the inherent differences in the model architectures rather than variations in training conditions. Additionally, this standardized approach simplifies the analysis of model performance, allowing the impact of the model itself on the outcomes to be discerned and evaluated. Table 5.1 provides detailed performance metrics results obtained from the CNN models when applied to the BreakHis [18] dataset at various magnification levels.

Table 5.1: Models result on the BreakHis dataset

Architecture	**Magnification Factors**															
	40X				**100X**				**200X**				**400X**			
	Acc.%	Pre.%	Rc.%	F1%	Acc.%	Pre.%	Rc.%	F1%	Acc.%	Pre.%	Rc.%	F1%	Acc.%	Pre.%	Rc.%	F1 %
VGG 16-SGD	**84.21**	**84.35**	**84.2**	**84.27**	85.71	86.86	85.08	85.96	**90.29**	**90.23**	**89.96**	**90.10**	84.70	89.80	81.21	85.28
VGG 16-Adam	75.78	75.63	78.24	73.19	81.63	81.29	81.33	81.31	76.69	76.99	76.99	76.99	85.88	86.79	85.48	86.13
VGG 19	76.84	76.03	75.58	75.80	91.84	91.79	91.75	91.77	85.44	85.93	85.28	85.61	88.23	91.50	85.95	88.63
Inception V3	75.78	75.63	70.91	73.19	81.63	81.29	81.33	81.31	76.69	76.99	76.99	76.99	85.88	86.79	85.49	86.13
Xception-SGD	69.47	67.63	58.61	62.79	**95.91**	**95.71**	**94.13**	**94.91**	**90.29**	**90.44**	**89.53**	**89.98**	**90.58**	**91.99**	**89.68**	**90.82**
Xception-Adam	**84.21**	**85.52**	**84.55**	**83.32**	89.79	89.71	89.71	89.71	87.37	87.43	86.90	86.37	83.52	88.05	84.14	80.56

Table 5.1 shows four convolutional neural network models, namely VGG16, VGG19, Inception V3, and Xception, across four magnification levels: 40x, 100x, 200x, and 400x. The table presents the performance metrics of these models, which include accuracy (Acc.%), precision (Pre.%), F1 score (F1%), and recall (Rc.%). When comparing models at each magnification factor, the red highlighting indicates the highest values in each metric. To better understand this information, we will analyze and discuss below.

- **Architectures performance at 40X Magnification**
  - VGG16-SGD shows the highest accuracy (84.21%), precision (84.35%), recall (84.2%), and F1-score (84.27%) at this magnification. These results indicate that it effectively processes low-magnification images.
  - Xception-Adam also performs well, with metrics similar to VGG16-SGD, making it an alternative.

- **Architectures performance at 100X Magnification**
  - Xception-SGD significantly outperforms other models with an accuracy of 95.91%, precision of 95.71%, recall of 94.13%, and F1-score of 94.01%.
- **Architectures performance at 200X Magnification**
  - Xception-SGD shows an accuracy of 90.29%, precision of 90.44%, recall of 89.53%, and F1-score of 89.98%. This consistency across magnifications shows its adaptability.
  - VGG16-SGD also performs well with a 90.29% accuracy and 90.10% F1-score, comparable to Xception-SGD,
- **Architectures performance at 400X Magnification**
  - Xception-SGD shows the highest metrics: accuracy (91.99%), precision (90.68%), recall (91.82%), and F1-score (90.82%), highlighting its overall reliability across various image resolutions.

5.2.1.1 Model Performance Discussion

The Xception architecture consistently outperformed all architectures across all magnification levels. This highlights the significance of choosing architectures like Xception for pathology image based breast cancer classification tasks. Its capability to effectively process complex image data makes it a valuable tool for improving diagnostic accuracy in breast cancer imaging.

The selection of the optimizer has a significant impact on model performance. Models that use the Adam optimizer tend to show lower performance compared to those utilizing SGD, indicating that SGD is more effective for this particular classification task. This finding emphasizes the importance of carefully selecting optimization algorithms to maximize the effectiveness of models in breast cancer classification.

It's important to note that at lower magnifications (40X), VGG16-SGD demonstrates superior performance, whereas at higher magnifications (200X and 400X), Xception-SGD excels. This variance suggests that different models are more suitable for different levels of image detail, highlighting the importance of tailoring model selection to the specific characteristics of the imaging data. Furthermore, the consistently strong performance of Xception-SGD across all magnifications indicates its robust generalization capabilities.

Based on the findings from Table 5.1, both the VGG16 and Xception models were selected for ensemble model development. The best models selected are:

- **VGG16 and Xception with Adam optimizer for 40x magnification factor**
- **Xception with SGD optimizer for 100x magnification factor**
- **Xception with SGD optimizer and VGG16 for 200x magnification factor**
- **Xception with SGD optimizer for 400x magnification factor**

5.2.2 The Models Performance When Trained with Yan et al. Dataset

As mentioned in the methodology, the magnification factors for the Yan et al. [24] dataset are not specified. However, a general description was provided, stating that the dataset contained images captured at 100x and 200x magnifications. Based on this information, the best-performing models from the BreakHis [18] dataset were selected and trained with the Yan et al. [24] dataset. The results of this experiment are presented in Table 5.2. Based on our results, all three models were selected for ensemble model development on the Yan et al. [24] dataset.

Table 5.2: VGG16 and Xception Models result on the Yan et al. dataset

Architecture	**Evaluation matrix**			
	Accuracy%	**Precision%**	**Recall%**	**F1 score%**
VGG16 with SGD optimizer	81	81.61	80.21	80.90
Xception with SGD optimizer	85.5	85.53	85.46	85.44
Xception with Adam optimizer	86.5	86.53	86.46	86.49

Table 5.2 compares performance metrics for two deep learning models, VGG16 and Xception, using different optimization techniques: Stochastic Gradient Descent (SGD) and Adam. The metrics are standard evaluation criteria used in classification tasks, ensuring credible and reliable results.

Here's an analysis of the data:

1. **VGG16 with SGD Optimizer:** when used with the SGD optimizer, has shown an accuracy of 81%. Additionally, it has a precision of 81.61%, an F1 score of 80.90%, and a recall of 80.21%. These values indicate that the model's performance is balanced across all metrics and thus, it is reasonably effective in its classifications.

2. **Xception with SGD Optimizer:** It has been observed that when trained with the same SGD optimizer, the Xception model outperforms the VGG16 model in all metrics. The Xception model achieved an accuracy of 85.5%, a precision of 85.53%, an F1 score of 85.44%, and a recall of 85.46%. This suggests that the more advanced architecture of Xception benefits from the capabilities of the SGD optimizer, leading to better overall performance compared to VGG16.

3. **Xception with Adam Optimizer:** Adam optimizer improves the performance of the Xception model, achieving the highest scores among the three configurations with an accuracy of 86.5%, precision of 86.53%, recall of 86.46%, and F1 score of 86.49%.

5.3 New Dataset Extraction

Sections 5.2.1 and 5.2.2 of the document present the results of the best-performing models for the BreakHis and Yan et al. datasets, respectively. These models were further refined using an ensemble method applied separately to each dataset. The ensemble method employed soft voting, allowing the models to leverage the strengths of multiple classifiers to improve overall prediction accuracy.

1. **Extracting Ductal and Lobular Labels on Yan et al. Dataset Using the Ensemble Model Trained on BreakHis**

   An ensemble model was created by combining the best-performing models previously identified using the BreakHis [18] dataset, focusing on the classification of ductal and lobular breast cancer labels. This ensemble model employs a soft voting technique, which involves aggregating predictions from multiple individual models and determining the final prediction based on a weighted average. The Yan et al. [24] dataset, which includes labels for invasive and in situ breast cancer staging, was used to test this ensemble model. The resulting predictions were stored in a designated IDC, DCIS, ILC, and LCIS folder, enabling the identification and classification of ductal and lobular breast cancer labels within the Yan et al. [24] dataset.

2. **Extracting Invasive and In situ Labels on BreakHis Dataset Using the Ensemble Model Trained by Yan et al.**

An ensemble model was created by combining three models trained using the Yan et al. [24] dataset, which classifies invasive and in situ breast cancer staging labels. This ensemble model also uses the soft voting technique. The BreakHis [18] dataset, which contains labels for ductal and lobular breast cancer subtypes, was used to predict this ensemble model. The resulting predictions were stored in a designated IDC, DCIS, ILC, and LCIS folder, allowing for the identification and classification of invasive and in situ breast cancer staging labels within the BreakHis [18] dataset.

Throughout the extraction process of the datasets, a binary classification strategy was utilized to assign labels of either 0 or 1. Images that produced prediction scores between 0 and 0.4 were identified as 0, whereas those with prediction scores ranging from 0.6 to 1 were labeled as 1. Predictions that fell within the 0.4 to 0.6 range were not included in the classification process due to uncertainty of the model's prediction probability, as they did not clearly indicate either class.

Using the first two steps mentioned in Section 5.3, we created a labeled dataset. Figure 5.1 displays dataset details.

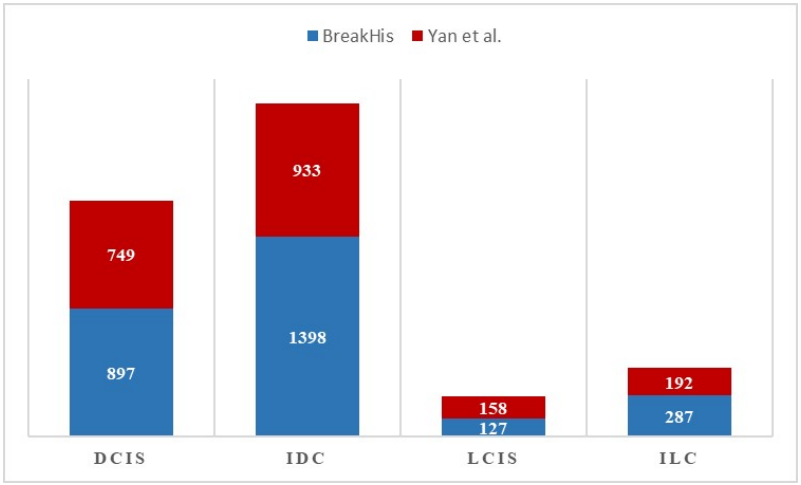

Figure 5.1: The new dataset

The bar chart displays the distribution of breast cancer sub-types extracted from two different datasets - one from Yan et al. [24] and another from BreakHis [18]. The sub-types of breast cancer presented in the chart include IDC, DCIS, ILC, and LCIS.

Each bar on the chart has three sections:

1. **Yan et al. [24] Dataset (Red Section):** This shows the number of datasets for each sub-type within the dataset extracted from Yan et al. [24].

2. **BreakHis [18] Dataset (Blue Section):** This shows the number of datasets for each sub-type extracted from the BreakHis [18] dataset.

5.4 Pathologists Evaluation

In order to assess the validity of the newly generated dataset, three expertise pathologists were enlisted. A total of 80 medical images were selected for evaluation, with a careful distribution across four subtypes of breast cancer: DCIS, IDC, LCIS and ILC. Each sub-type was equally represented by 20 images, ensuring a balanced and diverse set for each pathologist to review. Each pathologist received an identical dataset, which included a range of cases from each subtype. A variety of cases from each subtype were presented, aiming to provide a more reliable assessment of the dataset's performance in capturing the complexities of breast cancer pathology.

Table 5.3: Pathologists evaluation

Expertise	**DCIS (20)**		**IDC (20)**		**LCIS (20)**		**ILC (20)**	
	agree	**disagree**	**agree**	**disagree**	**agree**	**disagree**	**agree**	**disagree**
Pathologist 1	20	0	19	1	5	15	10	10
Pathologist 2	15	5	17	3	17	3	11	9
Pathologist 3	20	0	16	4	20	0	20	0
2 or 3 agreement	20	0	18	2	17	3	15	5

Table 5.3 presents detailed results from the three pathologists based on the four classes. The table lists the agreement and disagreement of each pathologist with the classification made by our model. An attempt was made to combine the three pathologists' agreements and present the resulting data in the table where any two or all three pathologists agreed. Through this process, it was found that, out of 80 images, the expertise all agreed on 70 of them and had different opinions on the other 10 images. This translates to an agreement percentage of 87.5% with the machine learning algorithm, providing valuable insight into the level of alignment between the doctors' opinions and our model's decisions.

Based on expert results, a 12.5% error rate was indicated on the new dataset. Though such errors are significant in the medical field, the ability of the approach to generate a newly labeled dataset using existing labels and to simplify the data collation process by effectively utilizing an underused dataset has been demonstrated.

5.5 Breast Cancer Classification with the New Dataset

The new dataset was utilized to categorize various breast cancer subtypes, such as DCIS, LCIS, IDC, and ILC, using the methodology shown in Chapter 4. The classification process involved a stratified cross-validation technique with five folds, allocating four folds for training and one for validation. Furthermore, unseen test data was separated based on patient-wise split.

The VGG16 and Xception models were trained for multi-class classification tasks using categorical cross-entropy as the loss function and SGD optimizer. Notably, hyperparameters such as the learning rate (0.0001), batch size (16), and dropout rate (0.3) were fine-tuned. The training process spanned 50 epochs to ensure comprehensive learning. After the training process, a soft voting ensemble technique was used to predict the unseen dataset.

The results of this approach and configuration are presented in Table 5.4, which summarizes the model's effectiveness in classifying breast cancer subtypes based on the provided dataset.

Table 5.4: The new dataset

Evaluation Matrix	Classification (%)
Accuracy	76.06
Precision	71.65
Recall	73.29
F1 Score	72.46

According to the expertise results of the analysis, the dataset has an error rate of 12.5%. It is important to take this into account when interpreting the results, as the presence of erroneous data can have a significant impact. The mislabeled data in the dataset suggests that the metrics used underestimate the model's actual performance.

5.6 Summary

In this chapter, we have presented and analyzed the results of our experiment aimed at generating a new labeled dataset. Our methodology, as described in Section 4, consisted of a series of structured steps to extract data from the BreakHis [18] and Yan et al. [24] histopathological breast cancer datasets for labeling the DCIS, IDC, LCIS, and ILC breast cancer subtypes using an automatic dataset labeling method. Our findings show that a newly labeled dataset can indeed be successfully generated, and we were able to utilize this dataset to classify breast cancer subtypes, including DCIS, IDC, LCIS, and ILC. To further validate our model, we randomly selected 80 images from the Yan et al. [24] dataset and had them evaluated by pathologists. The resulting assessment accuracy was 87.5%, indicating that the newly generated dataset achieved a high level of accuracy according to expert evaluation.

Chapter 6

Conclusion and Future Work

6.1 Conclusion

The task of classifying breast cancer presents a significant challenge due to the limited availability of datasets, especially for specific types of breast cancer. Despite the presence of labeled datasets in various research domains, the scarcity of datasets tailored to distinct breast cancer types remains a persistent obstacle. This scarcity hinders the effectiveness of classification efforts for these specific types of breast cancer.

To address this limitation, this research introduces a new approach - creating a new labeled dataset by using existing datasets dedicated to LCIS, and ILC breast cancer types. We utilized the BreakHis [18] dataset, which includes labels for ductal and lobular carcinoma, and the Yan et al. [24] dataset, labeled with invasive and in situ carcinoma, to establish labels for DCIS, IDC, LCIS, and ILC.

The newly generated dataset was subjected to careful evaluation by three pathologists, revealing an average accuracy of 87.5%. This newly generated dataset was then employed for the classification of DCIS, IDC, LCIS, and ILC. The innovative dataset creation process contributes to overcoming the challenges associated with limited datasets in breast cancer research, offering a more comprehensive resource for the classification of breast cancer types.

6.2 Future Work

We were able to generate a new dataset on breast cancer by using the existing dataset. However, upon expert analysis, it was confirmed that this generated dataset had an error rate of 12.5%. We recommend that this error rate be reduced by using Vision Transformer, as it shows promising results in medical applications [64], in the future.

References

[1] World Health Organisation. Cancer, 2017. https://www.who.int/news-room/fact-sheets/detail/cancer, Last accessed on 2023-8-18.

[2] World Health Organisation. Breast cancer, Jul 2023. https://www.who.int/news-room/fact-sheets/detail/breast-cancer, Last accessed on 2023-8-18.

[3] World Cancer Research Fund International. Breast cancer statistics, Apr 2022. https://www.wcrf.org/cancer-trends/breast-cancer-statistics/, Last accessed on 2023-8-18.

[4] Breast Cancer Reseach Foundation . Breast cancer statistics and resources, Aug 2021. https://www.bcrf.org/breast-cancer-statistics-and-resources/, Last accessed on 2023-8-18.

[5] National Cancer Institute. Breast anatomy. https://training.seer.cancer.gov/breast/anatomy/, Last accessed on 2023-8-18.

[6] Canadian Cancer Society. The breasts. https://cancer.ca/en/cancer-information/cancer-types/breast/what-is-breast-cancer/the-breasts, Last accessed on 2023-8-18.

[7] Memorial Sloan Kettering Cancer Center . Anatomy of the breast, 2023. https://www.mskcc.org/cancer-care/types/breast/anatomy-breast, Last accessed on 2023-8-18.

[8] Amy Kryzak. Breast anatomy. https://www.nationalbreastcancer.org/breast-anatomy/, Aug 2019. Last accessed on 2023-8-18.

[9] Abdullah-Al Nahid, Yinan Kong, et al. Involvement of machine learning for breast cancer image classification: a survey. *Computational and mathematical methods in medicine*, 2017, 2017.

[10] Ganesh N Sharma, Rahul Dave, Jyotsana Sanadya, Piush Sharma, and KK22247839 Sharma. Various types and management of breast cancer: an overview. *Journal of advanced pharmaceutical technology & research*, 1(2):109, 2010.

[11] Breast cancer: types, stages of the disease, prevention, May 2023.

[12] Lulu Wang. Early diagnosis of breast cancer. *Sensors*, 17(7):1572, 2017.

[13] Peter C Gøtzsche and Karsten Juhl Jørgensen. Screening for breast cancer with mammography. *Cochrane database of systematic reviews*, (6), 2013.

[14] Karen. What is an ultrasound machine and how does it work? - ultrasound solutions corp., Apr 2022.

[15] Gisela LG Menezes, Floor M Knuttel, Bertine L Stehouwer, Ruud M Pijnappel, and Maurice AAJ van den Bosch. Magnetic resonance imaging in breast cancer: a literature review and future perspectives. *World journal of clinical oncology*, 5(2):61, 2014.

[16] Amy Kryzak. Breast biopsy: Procedure types, what to expect results guide. `https://www.nationalbreastcancer.org/breast-cancer-biopsy/`, Aug 2019. Last accessed on 2023-8-18.

[17] Metin N Gurcan, Laura E Boucheron, Ali Can, Anant Madabhushi, Nasir M Rajpoot, and Bulent Yener. Histopathological image analysis: A review. *IEEE reviews in biomedical engineering*, 2:147–171, 2009.

[18] Fabio A Spanhol, Luiz S Oliveira, Caroline Petitjean, and Laurent Heutte. A dataset for breast cancer histopathological image classification. *Ieee transactions on biomedical engineering*, 63(7):1455–1462, 2015.

[19] Geert Litjens, Peter Bandi, Babak Ehteshami Bejnordi, Oscar Geessink, Maschenka Balkenhol, Peter Bult, Altuna Halilovic, Meyke Hermsen, Rob Van de Loo, Rob Vogels, et al. 1399 h&e-stained sentinel lymph node sections of breast cancer patients: the camelyon dataset. *GigaScience*, 7(6):giy065, 2018.

[20] Mitko Veta, Yujing J Heng, Nikolas Stathonikos, Babak Ehteshami Bejnordi, Francisco Beca, Thomas Wollmann, Karl Rohr, Manan A Shah, Dayong Wang, Mikael Rousson, et al. Predicting breast tumor proliferation from whole-slide images: the tupac16 challenge. *Medical image analysis*, 54:111–121, 2019.

[21] Guilherme Aresta, Teresa Araújo, Scotty Kwok, Sai Saketh Chennamsetty, Mohammed Safwan, Varghese Alex, Bahram Marami, Marcel Prastawa, Monica Chan, Michael Donovan, et al. Bach: Grand challenge on breast cancer histology images. *Medical image analysis*, 56:122–139, 2019.

[22] Roux Ludovic, Racoceanu Daniel, Loménie Nicolas, Kulikova Maria, Irshad Humayun, Klossa Jacques, Capron Frédérique, Genestie Catherine, et al. Mitosis detection in breast cancer histological images an icpr 2012 contest. *Journal of pathology informatics*, 4(1):8, 2013.

[23] paultimothymooney. Predict idc in breast cancer histology images. `https://www.kaggle.com/code/paultimothymooney/predict-idc-in-breast-cancer-histology-images/notebook`, Mar 2018. Last accessed on 2023-9-4.

[24] Rui Yan, Fei Ren, Zihao Wang, Lihua Wang, Tong Zhang, Yudong Liu, Xiaosong Rao, Chunhou Zheng, and Fa Zhang. Breast cancer histopathological image classification using a hybrid deep neural network. *Methods*, 173:52–60, 2020.

[25] `https://rdm.inesctec.pt/dataset/nis-2017-003`. Last accessed on 2023-9-4.

[26] `https://iciar2018-challenge.grand-challenge.org/Dataset/`, 2018. Last accessed on 2023-9-4.

[27] Alireza Osareh and Bita Shadgar. Machine learning techniques to diagnose breast cancer. In *2010 5th international symposium on health informatics and bioinformatics*, pages 114–120. IEEE, 2010.

[28] Scott Doyle, Shannon Agner, Anant Madabhushi, Michael Feldman, and John Tomaszewski. Automated grading of breast cancer histopathology using spectral clustering with textural and architectural image features. In *2008 5th IEEE international symposium on biomedical imaging: from nano to macro*, pages 496–499. IEEE, 2008.

[29] Bailing Zhang. Breast cancer diagnosis from biopsy images by serial fusion of random subspace ensembles. In *2011 4th International Conference on Biomedical Engineering and Informatics (BMEI)*, volume 1, pages 180–186. IEEE, 2011.

[30] Yungang Zhang, Bailing Zhang, Frans Coenen, and Wenjin Lu. Breast cancer diagnosis from biopsy images with highly reliable random subspace classifier ensembles. *Machine vision and applications*, 24(7):1405–1420, 2013.

[31] 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):4172, 2017.

[32] Mehdi Habibzadeh Motlagh, Mahboobeh Jannesari, HamidReza Aboulkheyr, Pegah Khosravi, Olivier Elemento, Mehdi Totonchi, and Iman Hajirasouliha. Breast cancer histopathological image classification: A deep learning approach. *BioRxiv*, page 242818, 2018.

[33] Farjana Parvin and Md Al Mehedi Hasan. A comparative study of different types of convolutional neural networks for breast cancer histopathological image classification. In *2020 IEEE Region 10 Symposium (TENSYMP)*, pages 945–948. IEEE, 2020.

[34] Suzanne C Wetstein, Nikolas Stathonikos, Josien PW Pluim, Yujing J Heng, Natalie D Ter Hoeve, Celien PH Vreuls, Paul J van Diest, and Mitko Veta. Deep learning-based grading of ductal carcinoma in situ in breast histopathology images. *Laboratory Investigation*, 101(4):525–533, 2021.

[35] Isha Gupta, Soumya Ranjan Nayak, Sheifali Gupta, Swati Singh, KD Verma, Abhishek Gupta, and Deo Prakash. A deep learning based approach to detect idc in histopathology images. *Multimedia Tools and Applications*, 81(25):36309–36330, 2022.

[36] Amy E McCart Reed, Lauren Kalinowski, Peter T Simpson, and Sunil R Lakhani. Invasive lobular carcinoma of the breast: the increasing importance of this special subtype. *Breast Cancer Research*, 23:1–16, 2021.

[37] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning*. MIT Press, 2016. `http://www.deeplearningbook.org`.

[38] Neha Sharma, Vibhor Jain, and Anju Mishra. An analysis of convolutional neural networks for image classification. *Procedia computer science*, 132:377–384, 2018.

[39] Keiron O'Shea and Ryan Nash. An introduction to convolutional neural networks. *arXiv preprint arXiv:1511.08458*, 2015.

[40] Saad Albawi, Tareq Abed Mohammed, and Saad Al-Zawi. Understanding of a convolutional neural network. In *2017 international conference on engineering and technology (ICET)*, pages 1–6. Ieee, 2017.

[41] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

[42] Laith Alzubaidi, Jinglan Zhang, Amjad J Humaidi, Ayad Al-Dujaili, Ye Duan, Omran Al-Shamma, José Santamaría, Mohammed A Fadhel, Muthana Al-Amidie, and Laith Farhan. Review of deep learning: Concepts, cnn architectures, challenges, applications, future directions. *Journal of big Data*, 8:1–74, 2021.

[43] Karl Weiss, Taghi M Khoshgoftaar, and DingDing Wang. A survey of transfer learning. *Journal of Big data*, 3:1–40, 2016.

[44] Musa Adamu Wakili, Harisu Abdullahi Shehu, Md Haidar Sharif, Md Haris Uddin Sharif, Abubakar Umar, Huseyin Kusetogullari, Ibrahim Furkan Ince, Sahin Uyaver, et al. Classification of breast cancer histopathological images using densenet and transfer learning. *Computational Intelligence and Neuroscience*, 2022, 2022.

[45] Gelan Ayana, Kokeb Dese, and Se-woon Choe. Transfer learning in breast cancer diagnoses via ultrasound imaging. *Cancers*, 13(4):738, 2021.

[46] Rich Caruana and Alexandru Niculescu-Mizil. An empirical evaluation of supervised learning for roc area. In *ROCAI*, pages 1–8. Citeseer, 2004.

[47] Omer Sagi and Lior Rokach. Ensemble learning: A survey. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 8(4):e1249, 2018.

[48] Divya Khyani, Soumya Jakkula, S Gowda, Anusha KJ, and Swetha KR. An interpretation of stacking and blending approach in machine learning. *International Research Journal of Engineering and Technology (IRJET)*, 8(07), 2021.

[49] Jiashuo Liu, Zheyan Shen, Yue He, Xingxuan Zhang, Renzhe Xu, Han Yu, and Peng Cui. Towards out-of-distribution generalization: A survey. *arXiv preprint arXiv:2108.13624*, 2021.

[50] Meelis Kull and Peter Flach. Patterns of dataset shift. In *First international workshop on learning over multiple contexts (LMCE) at ECML-PKDD*, volume 5, 2014.

[51] Joaquin Quinonero-Candela, Masashi Sugiyama, Anton Schwaighofer, and Neil D Lawrence. *Dataset shift in machine learning*. Mit Press, 2008.

[52] Marek Kowal, Paweł Filipczuk, Andrzej Obuchowicz, Józef Korbicz, and Roman Monczak. Computer-aided diagnosis of breast cancer based on fine needle biopsy microscopic images. *Computers in biology and medicine*, 43(10):1563–1572, 2013.

[53] Paweł Filipczuk, Thomas Fevens, Adam Krzyżak, and Roman Monczak. Computer-aided breast cancer diagnosis based on the analysis of cytological images of fine needle biopsies. *IEEE transactions on medical imaging*, 32(12):2169–2178, 2013.

[54] Yasmeen Mourice George, Hala Helmy Zayed, Mohamed Ismail Roushdy, and Bassant Mohamed Elbagoury. Remote computer-aided breast cancer detection and diagnosis system based on cytological images. *IEEE Systems Journal*, 8(3):949–964, 2013.

[55] Nicole Bussola, Alessia Marcolini, Valerio Maggio, Giuseppe Jurman, and Cesare Furlanello. Ai slipping on tiles: Data leakage in digital pathology. In *Pattern Recognition. ICPR International Workshops and Challenges: Virtual Event, January 10–15, 2021, Proceedings, Part I*, pages 167–182. Springer, 2021.

[56] M Thilagaraj, N Arunkumar, Petchinathan Govindan, et al. Classification of breast cancer images by implementing improved dcnn with artificial fish school model. *Computational Intelligence and Neuroscience*, 2022, 2022.

[57] Zeinab Sajjadnia, Raof Khayami, and Mohammad Reza Moosavi. Preprocessing breast cancer data to improve the data quality, diagnosis procedure, and medical care services. *Cancer Informatics*, 19:1176935120917955, 2020.

[58] Inês Domingues, Pedro H Abreu, and João Santos. Bi-rads classification of breast cancer: a new pre-processing pipeline for deep models training. In *2018 25th IEEE international conference on image processing (ICIP)*, pages 1378–1382. IEEE, 2018.

[59] Zabit Hameed, Sofia Zahia, Begonya Garcia-Zapirain, Jose Javier Aguirre, and Ana Maria Vanegas. Breast cancer histopathology image classification using an ensemble of deep learning models. *Sensors*, 20(16):4373, 2020.

[60] Shubham Mittal. Ensemble of transfer learnt classifiers for recognition of cardiovascular tissues from histological images. *Physical and Engineering Sciences in Medicine*, 44(3):655–665, 2021.

[61] Emanuela Paladini, Edoardo Vantaggiato, Fares Bougourzi, Cosimo Distante, Abdenour Hadid, and Abdelmalik Taleb-Ahmed. Two ensemble-cnn approaches for colorectal cancer tissue type classification. *Journal of Imaging*, 7(3):51, 2021.

[62] Nima Tajbakhsh, Jae Y Shin, Suryakanth R Gurudu, R Todd Hurst, Christopher B Kendall, Michael B Gotway, and Jianming Liang. Convolutional neural networks for medical image analysis: Full training or fine tuning? *IEEE transactions on medical imaging*, 35(5):1299–1312, 2016.

[63] Simon Kornblith, Jonathon Shlens, and Quoc V Le. Do better imagenet models transfer better? In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2661–2671, 2019.

[64] Fahad Shamshad, Salman Khan, Syed Waqas Zamir, Muhammad Haris Khan, Munawar Hayat, Fahad Shahbaz Khan, and Huazhu Fu. Transformers in medical imaging: A survey. *Medical Image Analysis*, 88:102802, 2023.