BREAST TUMOR SEGMENTATION AND CLASSIFICATION ON HISTOPATHOLOGICAL IMAGES USING MACHINE LEARNING

2023
MASTER THESIS
COMPUTER ENGINEERING

Zeyad Abdalkareem KHALAF

Thesis Advisor
Assist. Prof. Dr. Nehad T.A RAMAHA

BREAST TUMOR SEGMENTATION AND CLASSIFICATION ON HISTOPATHOLOGICAL IMAGES USING MACHINE LEARNING

Zeyad Abdalkareem KHALAF

Thesis Advisor

Assist. Prof. Dr. Nehad T.A RAMAHA

T.C.

Karabuk University

Institute of Graduate Programs

Department of Computer Engineering

Prepared as

Master Thesis

KARABUK

June 2023

I certify that in my opinion the thesis submitted by Zeyad Abdalkareem KHALAF titled "BREAST TUMOR SEGMENTATION AND CLASSIFICATION ON HISTOPATHOLOGICAL IMAGES USING MACHINE LEARNING" is fully adequate in scope and in quality as a thesis for the degree of Master of Science.

Assist. Prof. Dr. Nehad T.A RAMAHA .....

Thesis Advisor, Department of Computer Engineering

This thesis is accepted by the examining committee with a unanimous vote in the Department of Computer Engineering as a Master of Science thesis. June 15, 2023

Examining Committee Members (Institutions)	Signature
Chairman : Assist. Prof. Dr. Nehad T.A RAMAHA (KBU)	.....
Member : Assoc. Prof. Dr. Adib HABBAL (KBU)	.....
Member : Assist. Prof. Dr. Ali HAMİTOĞLU (İU)	.....

The degree of Master of Science by the thesis submitted is approved by the Administrative Board of the Institute of Graduate Programs, Karabuk University.

Assoc. Prof. Dr. Zeynep ÖZCAN .....

Director of the Institute of Graduate Programs

“I declare that all the information within this thesis has been gathered and presented in accordance with academic regulations and ethical principles and I have according to the requirements of these regulations and principles cited all those which do not originate in this work as well.”

Zeyad Abdalkareem KHALAF

ABSTRACT

M. Sc. Thesis

BREAST TUMOR SEGMENTATION AND CLASSIFICATION ON HISTOPATHOLOGICAL IMAGES USING MACHINE LEARNING

Zeyad Abdalkareem KHALAF

Karabük University
Institute of Graduate Programs
The Department of Computer Engineering

Thesis Advisor:
Assist. Prof. Dr. Nehad T.A RAMAHA
June 2023, 72 pages

Breast cancer detection using software has gained significant attention and development due to its potential to improve early detection, accuracy, and efficiency in identifying breast cancer. Early detection and classification is crucial for successful breast cancer treatment. Software-based detection algorithms can analyze medical images, such as mammograms, and identify potential abnormalities at an early stage, increasing the chances of successful treatment and reducing mortality rates. Therefore, the distinction between normal and malignant breast tissue can be used in diagnostic procedures as well as preoperative and postoperative assessments. Radiologists will be able to identify malignancies without making incisions in patients thanks to the growth of machine learning models and other technologies. Therefore, this study proposed a model consisting of two steps: first, characteristics from CNN are extracted, and then machine learning (SVM) is used to recognize and categorize breast

tumors (benign or malignant). Due to the extensive amount of training pictures used, CNN-based SVM eventually becomes overfitted. We now have an SVM that employs transfer learning and is based on CNN. Tumors in brain histopathological pictures are categorized using CNN-based Relu architecture and SVM with fused retrieved features via CNN. Precision, recall, F-measure, and accuracy have all been used to gauge the techniques' effectiveness. According to the findings, SVM-based CNN has a 97% success rate on BreakHis.

Key Words : Machine Learning, Breast Cancer, Classification, Feature Extraction, Measurements, Histopathological Image.

Science Code : 92402

ÖZET

Yüksek Lisans Tezi

MAKİNE ÖĞRENMEYİ KULLANARAK HİSTOPATOLOJİK GÖRÜNTÜLER ÜZERİNDEN MEME TÜMÖRÜNÜN BÖLÜMLENDİRİLMESİ VE SINIFLANDIRILMASI

Zeyad Abdalkareem KHALAF

Karabük Üniversitesi
Lisansüstü Eğitim Enstitüsü
Bilgisayar Mühendisliği Anabilim Dalı

Tez Danışmanı:
Dr. Öğr. Üyesi Nehad T.A RAMAHA
Haziran 2023, 72 sayfa

Meme tümörleri tedaviden önce histopatolojik görüntülerini kullanılarak sınıflandırılabilir; çünkü bu tümörler bir grup anormal dokudan oluşturulmuştur. Beyin histopatolojik resimlerinden tümör segmentasyonu ve sınıflandırması karmaşık ama çok önemli olduğunu biliyoruz. Ancak, bu segmentasyon ve kategorizasyon postoperatif değerlendirmelere, preoperatif planlamaya ve teşhise uygulanabilir. Bu nedenle, normal ve habis meme dokusu arasındaki ayrım, teşhis prosedürlerinde olduğu kadar ameliyat öncesi ve ameliyat sonrası değerlendirmelerde de kullanılabilir. Uzmanlar, makine öğrenimi modellerinin ve diğer teknolojilerin büyümesi sayesinde hastalarda kesi yapmadan maligniteleri belirleyebilecekler. Bu çalışmada iki adımdan oluşan bir model önerdik: ilk olarak, CNN'den özellikler çıkarıldı ve ardından tümörleri (iyi veya kötü) tanımak ve kategorize etmek için makine öğrenimi (SVM) kullanıldı. Kullanılan çok sayıda eğitim görüntüleri nedeniyle, CNN tabanlı SVM

sonunda gereğinden fazla takılır. Artık transfer öğrenmeyi kullanan ve CNN'ye dayalı bir SVM'ye sahibiz. Histopatolojik resimlerindeki tümörler, CNN tabanlı Relu mimarisi ve CNN aracılığıyla birleştirilmiş özellikleri içeren SVM kullanılarak kategorize edilir. Kesinlik, hatırlama, F-ölçüsü ve doğruluğun tümü, tekniklerin etkinliğini ölçmek için kullanılmıştır. Bulgulara göre SVM tabanlı CNN, BreakHis üzerinde %97 başarı oranına sahip model elde ettik.

Anahtar Kelimeler : Makine Öğrenimi, Tümör Segmentasyonu, Sınıflandırma, Özellik Çıkarma, Ölçümler, MRI Görüntüsü.

Bilim Kodu **:** 92402

ACKNOWLEDGMENT

I would like to give thanks to my advisor, Assist. Prof. Dr. Nehad RAMAHA, for his great interest and assistance in preparation of this thesis.

CONTENTS

Page

Page

Page

LIST OF FIGURES

Page

LIST OF TABLES

Page

SYMBOLS AND ABBREVITIONS INDEX

ABBREVITIONS

MRI	:	Magnetic Resonance Imaging
DSA	:	Domain Specific Architectures
PCA	:	Principal Component Analysis
FNN	:	Feed-Forward Neural Network
ACS	:	American Cancer Society
KNN	:	K-Nearest Neighbor
FCT	:	Fuzzy Cluster Techniques
WSI	:	Whole Slide Images
WDMRG	:	Wavelet Decomposition And Multiscale Region Growth
CSS	:	Curvature Scale Space
CAGA	:	String-like Factor Genetic Algorithm
SVM	:	Support Vector Machine
BCD	:	Breast Cancer Diagnosis
ROI	:	Region Of Interest
CAD	:	Computer Aided Diagnostics (CAD)
DOPS	:	Darwinian Optimization of Particle Swarm
X-Ray	:	X-Ray Imaging
CT	:	Computed Tomography
PET	:	Positron Emission Tomography
TE	:	Transmission Interval
BCDR	:	Breast Cancer Digital Repository
MAS	:	Mammography Analysis Society
BSO	:	The Brain-Storming Optimization
PSNLM	:	Pre-Smooth Non-Local Means Filter
SMO	:	Sequential Minimal Optimization
DPOS	:	Darwinian Optimization Of Particle Swarm

ML	:	Machine Learning
HOG	:	Histogram Of Oriented Gradients
LBP	:	Local Binary Pattern

PART 1

INTRODUCTION

Breast tumor classification is crucial for guiding treatment decisions, predicting outcomes, and ensuring the best possible care for patients with breast cancer. It is a fundamental aspect of breast cancer diagnosis and management, impacting both patients' lives and the overall healthcare system. In cases where breast cancer treatment is ongoing, tumor classification helps monitor the response to therapy and enables adjustments in the treatment plan if needed. Nowadays, breast cancer is a widespread disease in women. Various Computer Aided Diagnostics (CAD) models have been utilized based on Machine Learning (ML) algorithms. Computed tomography is a screening procedure for the early identification of breast tumour symptoms such as masses, calcifications, bilateral asymmetry, and architectural deformation. With the advent of Domain Specific Architectures (DSA) and the automatic features extraction capability of Convolutional Neural Networks (CNN), deep learning has become very popular in healthcare. For Breast Cancer Diagnosis (BCD), researchers are utilizing CNN-based architectures to enhance decision consistency and error reduction. This study critically analyzed the state-of-the-art CNN-based breast cancer diagnosis architectures. Moreover, it introduced a CNN model with Transfer Learning (TL) to design an efficient and accurate BCD pipeline. The proposed pipeline efficiently classifies histopathological images for BCD. A standard publicly available benchmark (BreakHis) [1-3] was used to validate the performance of the proposed pipeline.

Cancerous cells multiply quickly; if not identified at an early stage, these cancer cells can be fatal to the victim. Histopathology is the gold standard for detecting defects in tissues [3]. Breast Cancer (BC) is a disease in which the breast cells grow abnormally. It is one of the top reasons for deaths in females. For BC diagnosis, several Machine Learning (ML) based approaches are being used [4]. Nowadays, CNN-based schemes are commonly used for diagnosis because of their capacity to extract features

automatically. Histopathology is the examination of tissues under a microscope. Molecular Information, Cell Morphology, and Organization of tissues are studied under Histopathology, which studies disease symptoms under the microscope [5].

1.1. RESEARCH MOTIVATION

According to surveys, BC is the second most deadly common cancer after lung cancer [5]. According to the American Cancer Society (ACS), one out of every eight women in the United States is diagnosed with BC [6]. While other regions, namely Western Europe and Eastern Africa, have been reported to have cases of BC of 89.7 women and 19.3 women per 100,000, respectively [7-11]. 2030 cancer is projected to grow in upcoming years, and the number of cancer cases is expected to reach 27 million. If detected and diagnosed early, chances of successful treatment increase, and the suffering period of victims can also be avoided [12].

For BCD, traditional approaches are costly, time-consuming, and cause fatigue [13]. Moreover, a highly specialized pathologist is required to make a conclusive decision [14]. Even then, diverse inferences of pathologists on the same parameters cause a change in diagnostic Errors and confusion. Sometimes, these non-invasive techniques will not give a detailed, correct diagnosis [15-17]. So, this research proposed a pipeline to develop a CAD-based model for automatically classifying histopathological images for BCD.

1.2. PROBLEM STATEMENT

The problem is that the Speed and Efficiency without using software techniques. Classifying breast tumors using machine learning techniques offers several advantages and benefits over traditional methods. Machine learning algorithms have the potential to enhance the accuracy of breast tumor classification. By analyzing large amounts of data, including patient demographics, clinical features, and imaging characteristics, machine learning models can identify patterns and relationships that may not be apparent to human observers. This can lead to more precise and reliable tumor classification results. Machine learning algorithms can process and analyze large

volumes of data quickly, providing rapid tumor classification results. This speed and efficiency can support timely decision-making in clinical settings, enabling healthcare professionals to initiate appropriate treatment plans promptly. However, constructing CNN architecture capable of handling large input sizes is a difficult task. Regardless of how, downscaling the entire histopathology image to the size of CNN is not a favorable task, which may lead to a loss of information. To solve this problem, different image-level classification and patch-based techniques have been used to leverage CNN activation features [17]. Training deep learning models on many patches is a difficult task that takes a huge amount of time. A successful discriminatory patches selection procedure is crucial that utilizes geometric and static features [18, 19].

1.3. RESEARCH OBJECTIVE

The primary objective of breast cancer classification is to accurately categorize breast tumors or lesions into distinct subtypes based on their characteristics. This classification helps in understanding the nature of the tumor, guiding treatment decisions, and predicting patient outcomes. Here are some specific objectives of breast cancer classification:

Diagnosis: Classification systems aim to differentiate between benign (non-cancerous) and malignant (cancerous) breast tumors. Accurate diagnosis is crucial for initiating appropriate treatment and determining prognosis.

Prognosis: Breast cancer classification provides valuable information about the potential aggressiveness and likelihood of recurrence of the tumor. Subtyping tumors based on their molecular characteristics or histopathological features helps predict patient outcomes and guide treatment planning.

Treatment selection: Different subtypes of breast cancer respond differently to various treatment modalities. Classification allows for personalized medicine by identifying the most effective treatment options for a specific tumor subtype. To

improve the accuracy of tumor segmentation and classification (as benign or malignant), as well as reduce the number of false positives.
Prognostic and predictive biomarkers: Classification systems help identify specific molecular markers or genetic alterations associated with different breast cancer subtypes. These biomarkers serve as important tools for prognosis prediction, treatment response monitoring, and developing targeted therapies.

Disease monitoring and surveillance: Breast cancer classification aids in monitoring disease progression and evaluating treatment response over time. It allows healthcare professionals to track changes in tumor characteristics, detect recurrence, and modify treatment strategies accordingly.

By achieving these objectives, breast cancer classification plays a crucial role in improving patient care, optimizing treatment outcomes, and advancing our understanding of the disease.

1.4. RESEARCH QUESTIONS

Based on the research objectives, we have the following questions to find the optimal solution:

- What are the limitations of state-of-the-art BC classification architectures?
- How to deal with the problem of limited training data?
- What is the optimal technique for selecting discriminative patches, and how information from patches can be retrieved?

1.5. BENCHMARK SELECTION

The effectiveness of the suggested architecture is tested using the publicly available BreakHis dataset. The eosin-stained and hematoxylin slides from 82 female patients make up 7909 of the samples in this BCD dataset. There are 5429 photos of cancerous tissue and 2480 photographs of normal tissue on slides, all of which range in

magnification (i.e., 40x, 100x, 200x). Figures 1.2 and 1.3 show some example photos from the BreakHis dataset.

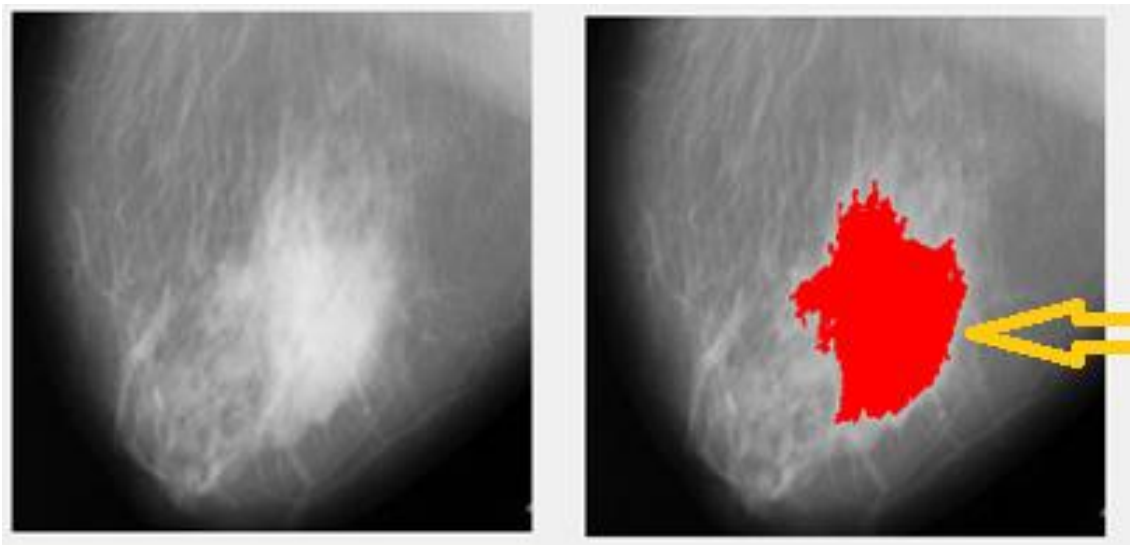

Figure 1.1. CT image of the breast, (a) is an original image, (b) is the output image (Class of Tumor is Benign, Size of Tumor is 15 %).

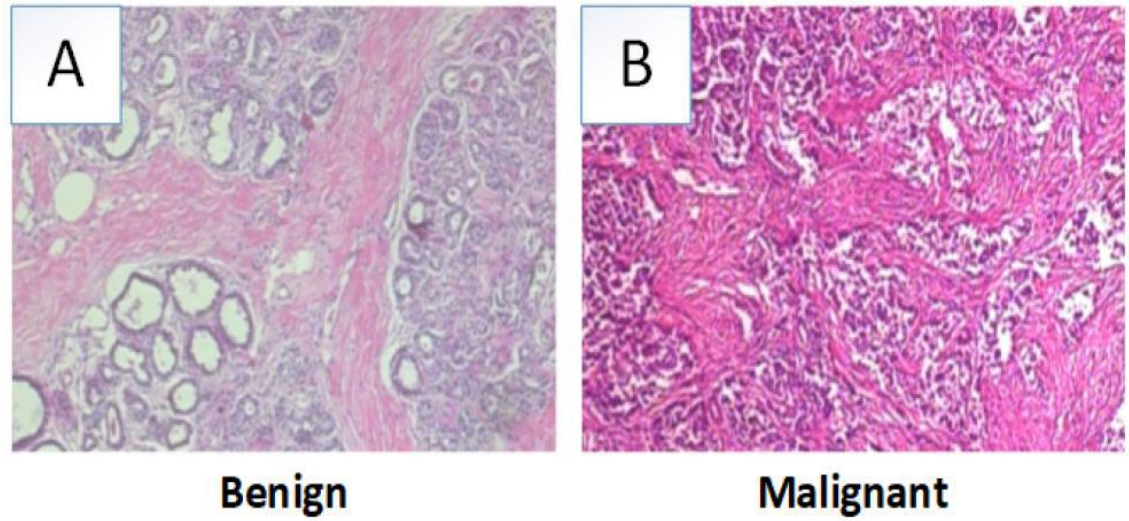

Figure 1.2. BreakHis dataset sample images.

To overcome the limitation of existing BCD models, this research utilizes a CNN architecture based SVM with transfer learning to efficiently classify Histopathological images. The Efficient-Net architecture will be used to learn image features from different databases. This research will introduce a discriminative patches-based features extraction scheme to enhance state-of-the-art Efficient-Net architecture.

Moreover, to handle data limitation problem transfer learning will be used for data augmentation.

1.6. ORGANIZATION OF THESIS

In this thesis, the neural networks used for breast tumor classification and extraction are reviewed and analyzed, as indicated before. In addition, the effectiveness of the approaches is assessed and contrasted with the dataset. In Chapter 2, we look at the most cutting-edge neural networks currently being used in the area of segmentation and classification. Chapter 3, this study explains both the research questions that were asked and the research procedures that were used. In addition to that, a few studies and pieces of previous research are discussed. Experiments are performed, and their outcomes are discussed in Chapter 4. After that, in Chapter 5, some conclusions and some suggestions for the future are formed.

PART 2

LITERATURE REVIEW

The most frequent kind of cancer among females is breast cancer (BC). Imaging techniques including mammography, MRI, and ultrasound are often used for diagnosis and screening. Although mammography and ultrasound imaging have improved greatly over the years, they still have their limitations, particularly in the presence of thick breast parenchyma, when it comes to detecting tumors and distinguishing between malignant and benign ones. When compared to other methods for detecting and diagnosing lesions, MRI has the highest specificity and sensitivity because of the superior picture quality it provides. However, even MRI has limitations, notably for diagnosis, that are only partly alleviated by combining it with mammography. Because of the limitations of these imaging methods, patients sometimes have to undergo painful and expensive optics operations only to be sure of a diagnosis. Numerous computational strategies have been developed to improve the sensitivity of BC diagnosis and screening without compromising specificity. Radionics, in particular, is gaining interest in oncology as a method of improving all three of these facets of cancer treatment. The term "radiomics" refers to the practice of extracting various quantitative aspects from single or several medical imaging modalities, therefore revealing qualities of pictures that are not immediately evident and substantially increasing the diagnostic and prognostic utility of medical imaging. A summary of current radiographic investigations concerning breast cancer is the focus of this research. Radiology can enhance breast cancer diagnosis, subtype, and grade categorization, treatment results, and recurrence prediction, according to the great majority of data. Radiology has the potential in the era of personalized medicine to enhance several facets of breast cancer care, including diagnosis, prognosis, surveillance, image-based intervention, and evaluation of response to therapy.

The widespread applicability of DL and NN has resulted in their widespread adoption and popularization. NN-based AI algorithms have been shown to be much more effective than conventional AI techniques for jobs that demand the aggregation of large quantities of data and the making of difficult decisions. To create desired (often predictive) outcomes given particular inputs, deep learning algorithms (a subtype of machine learning algorithms) utilize learning by example without requiring human participation in the selection of input characteristics (a large family of AI). Given enough training data and experience, machine learning can surpass human specialists in many cognitive tasks. Autonomous vehicles [20], poker [21], images [22], voices [23], translation [24, 25], and synthesis [26, 27] are only a few of the fields that have benefited greatly from deep learning. These architectures have been heavily trained on data from the internet since the early 2000s, with deep learning methods proving especially beneficial in image-based applications (such as image identification, segmentation, and classification). Because of the tools' generalizability and adaptability, it's a breeze to apply them to photos from different domains, so long as a substantial amount of training data is available from the desired area. In several areas, including drug discovery [26], genomics [27], and image processing [28], NN-based techniques are now considered the gold standard.

In the area of complicated slide imaging, a pathologist typically must sift through a large number of pictures in order to establish or confirm a diagnosis. Because certified whole-slide scanners and WSI digital infrastructure are so widely used, automated AI-based solutions [29] may now be successfully applied in the area of (digital) pathology (comparable to classical microscopy [30]).]). Another factor contributing to physicians' skepticism and unease is the fast development of AI algorithms for assessing and categorizing medical pictures. What role AI may play in defining the future of professional practice [31] is an important question, as is how much faith can be placed in a diagnostic system whose inner workings (with available architecture) are little known. Nevertheless, a vast majority of individuals (90%) [32-35] think that AI's benefits exceed its drawbacks. Artificial intelligence (AI) is positioned to provide professionals with value in their workloads, such as reducing busy schedules and freeing resources that could be devoted to important areas such as interactions between

professionals, patient-doctor relationships, and possibly playing a more satisfying role in improving patient safety.

There have been recent innovations in Desert Locust management strategies [36] that make better use of complicated and massive amounts of medical data in the decision-making process. Following this introduction to the framework of the deep learning methodology, we will compare and contrast it with more conventional machine learning techniques. We will discuss a variety of DL methods that have proven effective for AI breast cancer detection, as well as numerous important public data sources, such as hand-annotated breast cancer photos that are often used to train DL models. Here, we talk about the results, obstacles, and problems of the rising body of research in these domains, and the question of WSI specifically.

2.1. RELATED WORKS IN MEDICAL IMAGING

Radiology, pathology, and dermatology are just a few of the many medical specialties thamedica on imaging to diagnose a wide range of disorders. The use of machine learning in medical image analysis, object recognition, segmentation, and registration has been the subject of a number of recent review publications [37 – 41]. In [42], the most popular convolutional neural networks (CNNs) utilized in medical imaging, including ResNet and GoogleNet, are discussed. Data limits are only one of the obstacles to the expansion of ML applications in medical pictures, which are discussed by Varoquaux and Cheplygina in 2022. In addition, they provide suggestions for improving ML application research, such as implementing more stringent benchmarks for assessment [42]. Brief but useful, [42] guides you through the process of creating your first DL application in the field of medical imaging. [43] provides another summary of both cutting-edge and mainstream medical imaging techniques. See Figure 2.1 to observe how visually similar benign and cancerous pictures may be. This problem, brought on by sloppy picture tagging, is studied in [44]. For the PCAM dataset, the authors developed a new method based on co-training with global and local interpretations.

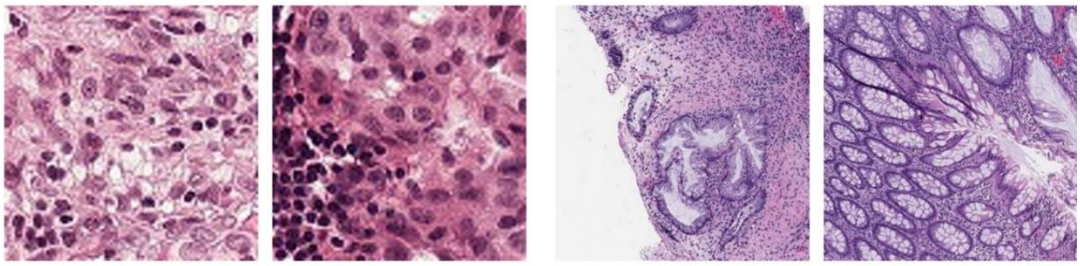

Figure 2.1. One example picture from each dataset. Evidence from a PCAM database and images from the MHIST.

Class imbalance is a typical issue in medical datasets, and the study examines this issue and potential remedies including assessing the dataset using various loss functions. The human eye has difficulty detecting subtle differences in patterns, which is an inherent issue in most medical imaging.

2.2. WORKS ASSOCIATED WITH MATHEMATICAL PATHOLOGY

In this study, we concentrate on histopathology photographs, which are histology images used by pathologists for diagnosis, since they are a good example of the kind of pathology images studied by computational pathologists. Developments in image processing, particularly with the use of CNNs, have spurred the emergence of the field of computational pathology in recent years. Whole Slide Images (WSI) are the primary data source in computational pathology. These are very high-resolution pictures, with resolutions of up to 10 k x 10 k pixels. Numerous computational pathology facets have been investigated by researchers [45 –50]. Due to the difficulties inherent in processing WSI directly, a unique and compressed form of WSI is presented in [51]. This representation, built on top of the cellular data in the WSI, has the potential to boost prediction accuracy by as much as 26%. According to [52], the PCAM dataset is the best for histopathology image classification, and they advise employing color modification methods. Utilizing a collection of CNN models, they determined that the accuracy results produced using a mixture of color changes were superior to those using the original RGB color.

2.3. THE PROCESSING AND ANALYZING METHODS

Using methods like fuzzy logic, evolutionary computing, and artificial neural networks, scientists have attempted several times to improve the diagnostic accuracy of breast cancer detection. Most clinical trials now include digital tools for confirmation and emphasis on decision-making. Medical professionals often make use of a patient's medical picture to help them diagnose the issue. Medical professionals formulate a strategy for patient care based on what they learn from a picture (particularly the limits of an item). However, inaccurate diagnoses arise because of insufficient imaging data (because of ineffective processing).

In order to arrive at an accurate diagnosis, more data must be gathered via the use of computational tools. Edge detection and pattern recognition are two key methods that are used in the area of image processing. These techniques are particularly useful for locating items of interest and conducting searches for them. The primary objective of these approaches is to identify image stabilizers that are used in digital photography [53].

Extraction of white breast tissue and identification of single cells [54], as well as topological imaging of human breast proliferation using magnetic resonance imaging [55], are just two examples of the numerous research that have been conducted on this subject. There have been many more studies as well. When identifying and interpreting medical imaging data, segmentation is a process that is known for being infamously difficult but is still essential.

You will need to split the MRI pictures in half if you wish to get rid of anything particular (like a breast tumor, for example) [56]. A local research focused on breast cancer cells was conducted, with normal breast cells serving as the study's control group [57]. Image segmentation was discussed in reference (58), where it was suggested that fuzzy cluster techniques (FCM) be used. It has been suggested [59] that breast cancer be split into two distinct groups by using a combination of different approaches.

An essential component of our study is the evaluation of the individual gray and white matter in MRI images based on the degree of similarity between the two types of tissue. Methods such as and may be used to distribute an object's edges, bounds, and characteristics either linearly or non-linearly, respectively, depending on the technique used.

There is a wide variety of transformations available to handle issues that arise in the course of picture analysis (such as Canny, Pruitt, Logue, Roberts and Pruitt [60, 61] and transformation transformation [62]).

These days, there are two categories that may be used to classify the methods that are used to identify and evaluate tumors: spatial and contour. The spatial approaches [63–66] identify groupings of pixel clusters that have a visually comparable appearance. These techniques streamline some aspects of the low-level processing, which ultimately makes the radiologist's work simpler (eg, histogram analysis, classification, and thresholding).

It is common knowledge that the dangers connected with neoplasia may be mitigated by performing an early diagnosis and then receiving therapy for the condition [67]. The fact that this tumor is cancer assists medical professionals in determining the best course of treatment for it, which in turn assists people in regaining their health [66, 67].

According to the findings of studies [68], the categorization of tumors might be difficult. It is difficult to distinguish solitary ducts of uncertain size, some aggregated symptoms of tumors, and other conditions like this because of inadequate contrast and opacity of the tumor boundaries on ultrasound pictures. This is mostly the cause of the problem.

[69] presents a novel algorithmic foundation for object recognition and segmentation in ultrasonic scanning, specifically for breast cancer. [70] suggested a method for identifying a local component (breast tumor) by segmenting mammographic pictures

(making use of fundamental image processing methods). This method is able to offer trustworthy findings in real time.

These techniques (wavelet transform and K-combining) are used in the clinical setting for the purpose of mammographic tumor segmentation [70]. After that, a K-means clustering algorithm is used to the contrast picture, and a threshold is used to locate the location of the tumor.

In [71], an effort was made to enhance mammography segmentation by using the double slash technique. A box was superimposed over mammograms of various quality in the final segmented picture so that it would be easier to diagnose breast cancer in the patients.

It's possible that using a deep neural network to identify breast cancers on mammograms would help medical professionals make more precise diagnoses. Deep learning techniques (such as neural networks, for example) may be used to address the diagnostic issue by classifying objects of interest as abnormal or normal [72, 73].

2.4. BREAST HISTOLOGY REQUIREMENTS

Integrating data from several sources is essential for drawing conclusions and making diagnoses in the study of medical imagery. Several methods for the automated evaluation of photographic targets are discussed in the dissertation (breast cancer).

Automatic structure recognition and segmentation in digital histopathology pictures was the focus of the work presented in [74]. Three distinct spatial scales are combined in the suggested approach of picture analysis:

First-level information gleaned from pixel values: In this stage, we utilize a Bayesian classifier to calculate the probability that each pixel represents a target feature.

Object identification using a higher level of related pixel data: Two techniques exist for compiling such summaries:

a. Employ a level set method; in probabilistic scenes created by a Bayesian classifier, the contour develops with time, making it possible to identify object boundaries. b. Low-probability identification of nuclei and glands by the use of a pattern-matching technique that employs shape models. Third, the connections between histological structures as a wellspring of specialized knowledge: In order to verify that the discovered components are constituents of the structures of interest, we utilize our domain knowledge to place structural restrictions on the elements.

Using a nuclear segmentation algorithm for precise gland extraction has recently been demonstrated to be useful for automatically classifying breast and prostate cancers and for differentiating normal from malignant breast tissue in a tissue sample.

A novel strategy for breast ROI extraction was presented in [75]. This approach uses a neural network and information from nearby pixels in an effort to lower the number of false positives (FP). As with other models, this one has two phases:

A trained model is created during the training phase by extracting several packets from the ROI and background.

Second, in order to isolate the target area from the surrounding noise, a fixed-size testing window must be applied to the picture.

After identifying the ROI, the model does a distance transformation to filter out extraneous information. The authors employed a total of 250 ultrasound pictures in their experiments (100 malignant and 150 benign).

Mammography mass detection and segmentation computational approaches in terms of MLO and CC were reported in [76]. To get started, we put into action the algorithm for disabling the tools. Thus, the wavelet transform and the Wiener filter constitute the backbone of the technique used to reduce noise in images and boost their grayscale. Researchers employed a genetic algorithm, a wavelet transform, and repeated thresholding to identify and isolate mammographic lesions indicative of breast cancer.

Figure 2.2 demonstrates how the Digital Database for Screening Mammography (DDSM) is used to randomly select mammograms for screening.

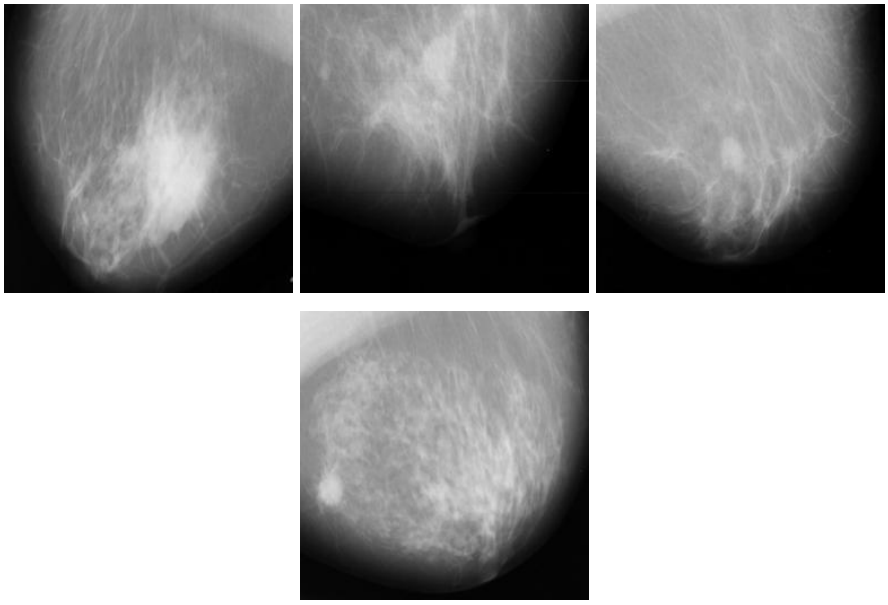

Figure 2.2. An example of Mammography images of breast cancer (benign and malignant).

In [77], a straightforward technique is provided for identifying cancerous tissue in mammograms. After the tumor has been outlined, the affected areas are segmented so that mammography may be performed. Raw image processing techniques like averaging and thresholding form its foundation. The tumor detection industry has inspired the creation of novel statistical techniques such as the Max-Mean and Least-Variance approaches.

Breast mass retrieval parameters, including circumference, were used in [78]. (As established by means of computational segmentation). Mammographic breast masses may be diagnosed and classified using these methods (malignant or benign). There were two methods used to determine the bounds of a cluster. Both dynamic programming and enlarging a bounded region are included in the first approach. In addition, miniature replicas of the lines were used to extract a set of 6 attributes that describe the block's borders (similar to ellipses). The first is a quantitative assessment of an educated estimate (informed by knowledge of edge signatures), while the second

is a piece of data that has been collected and analyzed to provide more insight. Estimated Quantity (using relative gradient direction as standard).

Three widely-used classifiers employed the aforementioned features and their many permutations to generate predictions in 349 instances of group diagnosis. Classifiers include support vector machines, Fisher's linear features, and Bayes classifiers. When everything was said and done, the systems were compared to see how effectively breast cancers were identified.

The authors of [79] developed a deep learning-based computer-aided method for detecting, classifying, and segmenting the malignant area in mammograms. To get rid of the muscular area, artifacts, and noise, a preprocessing method is recommended. You can't trust the sums, since keeping muscle tissue intact often yields erroneous results. To further improve system performance and deal with the heavy load placed on the system's resources, the pre-processed picture is transformed to 512 x 512 patches. There are now two freely accessible breast cancer databases online. The Society for Breast Image Analysis Digital Mammography Data Set and the Digital Mammography Screening Database (CBIS-DDSM). In addition, DeepLab and Mask RCNN are two cutting-edge deep learning-based instance segmentation frameworks.
The authors of referenced paper [80] suggested a technique for quantitative automated picture analysis (BCH images). We can increase the picture quality by isolating the cores from their backgrounds using the top-lower hat transform. The area of interest (ROI) and its precise location are then identified by wavelet decomposition and multiscale region growth (WDMR). Following this, overlapping cells are segmented using a dual strategy partition model (DSSM) for improved resilience and accuracy. Angle detection in CSS (Curvature Scale Space) and adaptive mathematical modeling are also components of DSSM. We use 138 color and shape compositional cues to categorize cell nuclei. Then, a better feature set is obtained by combining the String-like Factor Genetic Algorithm (CAGA) with the Support Vector Machine (SVM). Finally, 68 BCH pictures were used to assess the motion, which included around 3,600 cells.

It is suggested that FCNNs (fast convolutional neural networks) be employed for segmenting pixel regions [81]. FCNN removes the extra calculations included in the CNN standard without sacrificing performance. According to our tests, it takes just 2.3 seconds to segment a 1000 by 1000 pixel picture.

The research suggests an improved threshold-based approach for producing ROI, as well as a trainable segmentation technique. A novel hybrid technique was used to segment the pectoral muscle and the breast border area. The company's founding principles were based on ML and thresholding techniques. After the tape was taken out of the wavelet transform, the breast region could be estimated. The original breast contours were determined using a novel thresholding technique. Removing insignificant items fixed the upper bound on estimates. This method was implemented using a combination of morphological and masking techniques. The segmentation process in medical imaging has come a long way. The main focus of recent advancements has been on increasing the speed and precision of machine learning methods. There has been a lot of talk in the academic literature about how crucial ML approaches are to the future of effective and precise segmentation algorithms. To pinpoint the pectoral muscles and their respective ROI, the research used an ML technique. This technique integrated directed gradient histogram (HOG) features with neural network classifiers [82]. The segmentation strategy was validated across three independent datasets consisting of 100, 200, and 100 mammograms, respectively. Mammography Analysis Society (MAS) and Breast Cancer Digital Repository (BCDR) pictures were used in this study (mini-MIAS).

Last but not least, researchers from [83] suggest an automated system that is based on the expansion of the ideal area to diagnose breast masses. In addition to this, a mechanism known as swarm optimization is applied in order to give ideal seed sites and thresholds (Dragon Fly Optimization, or DFO). In order to extract the texture qualities from images, GLCM and GLRLM are used to segment the images. After that, the textural characteristics are sent into a Feed Forward Neural Network (FFNN) classifier that has been trained via backpropagation. Keep in mind that the FFNN has the ability to determine whether or not a picture reveals malignancy. In the last step of this investigation, the efficacy of the suggested detection approach was assessed with

the use of photos taken from the DDSM database [83, 84]. Techniques that are based on machine learning (ML) and deep learning (DL) are the two categories that make up the current BC categorization. This investigation will exclusively concentrate on methods that are based on deep learning and are utilized for the detection of cancer. The most cutting-edge CNN-based strategies were dissected and evaluated in detail. The critical evaluation of CNN-based techniques for BCD is shown in Table 2.1.

Table 2.1. State of arts in machine learning on breast cancer diagnosis.

Author	Year	Techniques	Dataset	Limitations
Nahid et al.	2018	CNN, LSTM, and a combination of CNN and LSTM	BreakHis	Overfitting, gradient vanishing, and exploding problems.
Ahmad et al.	2022	CNN with multiple Instance Learning (MIP)	BreakHis	Overfitting, and depth-wise feature scaling.
Vo et al.	2019	Resnet-50, VGG-16 and VGG-19 models	BreakHis	Gradient vanishing, feature map redundancy, and Single dimension scaling.
Gour et al.	2021	Transfer Learning, Alexnet, VGG-16 and SVM	BreakHis	Gradient vanishing, kernel function dependency.
Sheela et al.	2022	Deep Learning and Excited ResNet	BreakHis	Feature map redundancy, single-dimension scaling, and network saturation.

2.5. BACKGROUND

In the subject of Artificial Intelligence known as ML, statistical methods are used to give computers the capacity to "learn" and develop autonomously over time with no human instruction. To this end, researchers in the field of ML [56] investigate and build algorithms that can learn from and make predictions on data and datasets using a variety of different pedagogical approaches. Arthur Samuel first used the phrase "ML" [20] in 1959. The following are descriptions of the three main types of ML:

- To learn the mapping function from the input to the output, a computer may utilize supervised learning, in which part of the training data is already tagged with the right answers. The aim is to get a close enough approximation of the mapping function that you can anticipate the values of the output variables (Y) given a fresh set of input data (x).

- The second kind of ML is called "unsupervised learning," and it involves teaching a system to function using data that has not been sorted or tagged. Because of this, the algorithm may independently carry out computations. The computer is tasked with discovering the underlying structure by categorizing the data into groups based on similarities and differences.
- Third, we have reinforcement learning, in which a machine or agent "learns" how to behave in a given environment by carrying out actions and observing the subsequent consequences; this allows the machine or agent to choose the most effective course of action to take under certain conditions.

In today's world, it's just not possible for a single person to constantly monitor and evaluate all of the data that's being generated. The field of ML, an area of study within computer science and a significant part of AI, is largely concerned with the creation of algorithms to solve this issue. Recent progress in this area has led to almost endless potential uses in disciplines as diverse as finance, data security, and medicine. However, there is a lot of room for improvement when it comes to employing ML in social media services, illness prediction and identification, virtual assistants, search engine refinement, fraud detection, manufacturing, etc. To make our lives simpler and more convenient, it only gets better and more integrated in the future.

2.5.1. Supervised Deep Learning

Efforts to predict medical visual classifications began with the advent of AI. Using a standard Naive Bayes classifier [25] automated the process of diagnosing breast cancer. As it turns out, AI is doable for this diagnosis, as even basic classifiers perform adequately.

Preprocessing was the initial stage of their thesis. There were numerous noisy pixels in their initial data because of this. The picture was blurred using a Gaussian filter, which also helped to minimize noise. The histogram was then extended to increase clarity. Since determining a tumor's categorization relies on locating its nuclei, the second stage is nucleus segmentation.

They used four different algorithms for grouping data: a competitive neural network, a fuzzy C-means method, a K-means strategy, and a Gaussian mixture model. Then, we took each segment and retrieved 42 characteristics from it. In-depth human pathologists were responsible for selecting the characteristics. The characteristics were then used as inputs for classifiers. They used estimated kernel densities to train a Naive Bayes classifier. The training dataset consisted of 500 actual medical photos from 50 individuals.

In order to evaluate the results, they used the n-fold cross-validation technique. The results showed a high degree of accuracy (96%) and showed great potential for the use of AI in the creation of breast cancer diagnosis. Their preparation and data collection methods were shown to be effective, resulting in a reliable and impartial dataset.

Supervised learning is regarded as one of the cornerstones of machine learning. The machine learning technique in this scenario is taught using labeled data. Despite the fact that supervised learning requires correctly labeled data in order to function, it is very powerful when used in the right situations [13].

2.5.2. Machine Learning Without Human Supervision

One advantage of unsupervised machine learning is that it can work with data that has not been labeled. This eliminates the need for any kind of manual labor to render the dataset machine-readable, enabling the application to scale to far bigger datasets. Without the constraint of labels, unsupervised learning is able to uncover previously unseen patterns. There is no need for human intervention since the computer can infer abstract associations between data points on its own [8].

In Semi-Supervised Machine Learning

The semi-supervised learning paradigm combines aspects of both supervised and unsupervised learning. During training, a semi-supervised learning model uses a combination of a small quantity of labeled data and a big amount of unlabeled data.

For prediction purposes, this learning model utilizes just a small sample of labeled data.

2.5.3. Algorithms for Supervised Machine Learning

In this section, we will explain the Supervised machine learning approaches that were used in this thesis.

In supervised learning, the method that is used the most often is known as the Random Forest Classification System. The idea of ensemble learning is at the heart of this methodology. This idea is broken down even further into separate classifiers, which are then merged to provide accurate predictions. A number of different classifiers are used in order to improve the performance of the model and address increasingly complicated problems. Using a Random Forest, which is made of several Decision trees that are subsets of the dataset and uses an average of the subsets, is one way to improve the accuracy of the dataset that has been given. This method utilizes an average of the subsets.

A method for supervised learning on a computer, known as decision tree learning, decision trees may be used for both classification and regression analysis. The root node of the Decision Tree is meant to stand in for the whole collection of examples, while the branches are meant to represent the rules themselves. The leaf nodes are what determine the properties of the dataset, and as a result, those qualities are mirrored in the output.

Support Vector Machine

With the assistance of a Support Vector Machine, one is able to solve problems relating to both classification and regression. On the other hand, it is often used in situations when classification problems arise. A single coordinate in an n-dimensional space is used to represent each individual data point when using the Support Vector Machine (SVM) approach (where n is the number of features). After that, we give the data their proper classification by locating the hyper-plane that most effectively divides the two

categories. A support vector, in its most basic definition, is an array of coordinates that stands for a single observation. It is not necessary to go any farther than the Support Vector Machine classifier if you are looking for a boundary that can consistently split hyperplanes and lines.

Logistic Regression

Logical regression is one of the Machine Learning algorithms that is used the most, and it is a component of the Supervised Learning technique. It is put to use in the process of forecasting the value of a categorical dependent variable with the help of a group of independent variables. The value of a categorical dependent variable may be predicted with the use of a technique called logistic regression. Logistic regression is a useful approach for machine learning because it can be used for both continuous and discrete datasets to calculate probabilities and classify new data. This flexibility makes logistic regression one of the most often used techniques for ML. Using logistic regression, you may classify observations based on many types of data and rapidly determine which characteristics are the most important.

2.6. SUMMARY

The emerging field of research called "radiomics" [85][86] focuses on the extraction of meaning qualities from clinical images. For example, radionics is being used alongside other imaging modalities, diagnostic features, and machine-learning approaches in breast cancer research to predict not only the presence and location of malignant lesions but also prognostic factors like the response to NAC treatment and the risk of tumor progression. The main drawback of radionics is the significant training in computer methods required before they can be applied in a clinical environment. The time required for model construction, training, and execution is influenced by several factors, including but not limited to the size and structure of the dataset, the available computational resources, and the complexity of the model. Processes in traditional radionics that rely on manually constructed qualities need less computing power than DCNN-based radiomics [87]. In addition, the computing requirements of Neural Network-based deep learning systems are higher than those of

more traditional machine learning architectures like KNN and SVM. However, the accuracy of tissue classifications achieved using KNN or SVM is lower than that of classifications obtained using NN [88]. As the size of a dataset grows, so does the amount of time needed to calculate anything. So, various techniques use picture reduction or preemptive resolution degradation to focus on specific areas of interest (ROIs) in the input pictures [89]. For medical image classification, option (a) is preferred because it enables the training of algorithms to recognize lesions in specific areas of clinical interest without having to throw away potentially relevant data (regions outside of the targeted regions) [90]. However, decreased sensitivity may result from lower picture resolutions in the second option [91]. It's possible that the time needed to evaluate a picture may be considerably reduced if researchers in the fields of radiomics and deep learning adopted the practice of using parallel computing approaches on Graphics Processing Units (GPU) [92].

A summary of machine-learning approaches for breast cancer segmentation and classification is provided in Table 2.2.

Table 2.2. Describes the methodology and performance indicators used in research that used machine-learning approaches for the detection, segmentation, and classification of breast cancer.

Year	Used Images	Methods	Classification	Accuracy	Limitation
[2] 2018	Histology Images	Convolutional Neural Networks	binary classification	96.15 %	Overfitting, gradient vanishing CNN and LSTM and exploding problems.
[3] 2020	Histology Images	MIL Convolutional neural network	Multi classification	89.52%	Overfitting, and depth wise Learning (MIP) feature scaling.
[6] 2020	Histology Images	Residu learning-based CNN	binary classification	90.49%	Gradient vanishing, feature map models redundancy, and Single dimension scaling.
[18] 2019	Histology Images	incremental boosting CNN	Multi classification	96.4 %	Gradient vanishing, kernel 12, and CNN function dependency.
[68] 2023	Histology Images	LBP + SVM	binary classification	96.91%	The limitation of LBP and SVM are two separate components in the pipeline, and their performance is dependent on how well

					the features are extracted and the classifier is tuned.
[76] 2022	Histology Images	Feed Forward Neural Network (FFNN)	binary classification	97.8%	Feature map redundancy, ssingle-dimension scaling, and network saturation.
[93] 2019	Histology Images	CNN and LSTM	binary classification	91.00%	Overfitting, gradient vanishing CNN and LSTM and exploding problems.
[94] 2020	Histology Images	CNN, SE-ResNet	binary classification	98.87%	Gradient vanishing, feature map models redundancy, and Single dimension scaling
			Multi classification	90.66%	
[95] 2018	Histology Images	(CNN), ResNet-50 and DenseNet-161	binary classification	97.89%	Feature map redundancy, single dimension scaling, and network saturation.
[96] 2023	Histology Images	FabNet model	binary classification	97.10%	Training data limitations: The model's performance heavily relies on the quality, size, and representativeness of the training dataset. Insufficient or unrepresentative data may hinder its effectiveness.
2022 [115]	Histology Images	deep CNN models	binary classification	95.85	Overfitting, and depth wise Learning (MIP) feature scaling.
2022 [116]	Histology Images	CNN Models	binary classification	96.38	Gradient vanishing, feature map models redundancy, and Single dimension scaling.
2022 [117]	Histology Images	DCNN with fficientNetV2-B0	binary classification	97.82	Gradient vanishing, kernel 12, and CNN function dependency.
2022 [118]	Histology Images	6B-Net deep CNN	binary classification	94.20	Overfitting, gradient vanishing ML and Region Growing and exploding problems.

PART 3

METHODOLOGY

3.1 INTRODUCTION

Cancer, which is caused by cellular development in an uncontrolled way, is a major health risk for people. According to data, millions of individuals throughout the globe are afflicted with cancer. These numbers show both the new cases of cancer diagnosed in Australia in 2017 and the total number of deaths from cancer that year. These results also show that both the incidence of and mortality from breast cancer is higher among women than among men. It's clear from this that females are more likely to be diagnosed with breast cancer (BC) than men. While these numbers are just for Australia, they may be reflective of a global trend.

Thousands of women's lives may be saved every year by accurate BC diagnosis; however, this relies heavily on accurate identification of the malignancy. Capturing a snapshot of the cancer-affected region is crucial for locating BC since it provides information about the disease's present status. Many noninvasive biomedical imaging modalities have been used. These include ultrasound, x-ray, and computerized tomography (CAT) scans. Histopathological Images are one kind of imaging technology; however, they are intrusive. Histopathological imaging, in particular, presents a number of difficulties that make investigation difficult. It's not easy to do analysis on histopathological images, and research of this nature is a complex endeavor. The photographs, there is certain to be debate among doctors. Inasmuch as they are human, physicians and other medical professionals are also subject to error. Computer-aided diagnosis (CAD) information, such as illness categorization, is very helpful to doctors and other medical personnel. Numerous research institutions are investigating how to make CAD programs more effective. The breast image classifier, for instance, is the result of applying cutting-edge engineering methods to a standard

image classifier. The BC image classifier employs cutting-edge DNN algorithms to reliably classify patient images for usage in clinical settings.

Simply said, a DNN's primary functionality is built on a neural network (NN). Rosenblatt developed the concept of NN in 1957 [97], which bases its decisions on a threshold. The ignorance model, first proposed by Fukushima [98], is a lightweight convolutional neural network (CNN) model with a sophisticated design. The primary objective of this research is to establish routines of motivation that are remarkably resilient in the face of moderate modifications to the initial conditions. This "Negotron" model was CNN's initial attempt at interpreting biological signals [99]. For mammography classification, in particular, Wu et al. [100] were the first to design and evaluate a CNN-based model. Because of its high computational cost, the CNN model wasn't widely used for breast image categorization until the authors of [101] created their model dubbed AlexNet. The AlexNet model is a major step forward in the field of image analysis and classification. This model has been used as a basis for the development of several models for biomedical image classification, including ResNet [102], Inception, Inception-V4, and Inception-ResNet [103]. AJ's CNN model classified a mammography dataset with 93.35 percent accuracy and 93.00 percent AUC [104]. (MIAS-mini, DDSM). To classify mammograms using CNN, Keo et al. [105] were the pioneers. To do this, they employed Cards 2, 5, and 10 and had a 71.40 percent success rate overall. Using a CNN technique, Ertosun and Rubin [106] can automatically identify lesions and categorize breast pictures with an accuracy of 85.0%.

Qui et al. [107] used 560 ROIs to classify a series of mammograms into benign and malignant categories. The study's author [108] successfully distinguished between benign and malignant mammography pictures with a rate of success of 96.70 percent. Sahiner et al. [109] analyzed a collection of mammograms and found a ROC of 0.87. M.M. Jadoon et al. implemented the CNN model. Mammograms were sorted into three categories: benign, suspicious, and malignant. The histopathological mammograms were assessed in the same manner as the mammograms. Recently, [110] used point detection to categorize a collection of histological pictures into benign and malignant classes by locating nuclei in each image. Normal tissue, benign tissue, localized

carcinoma, and invasive carcinoma are the four categories used by Araujo et al. (CNN) to classify histopathological pictures, whereas normal tissue, benign epithelial tissue, localized carcinoma, and malignant (cancerous) tumor are the two categories used (malignant and non-malignant). accuracy of 77.80% for four-class classification and accuracy of 83.3% for two-class classification. All the pictures are based on a greater magnification. At 40x magnification, the best pictures were recognized with an accuracy of 85.64.88%. The study's General Image Classification Model, seen in Figure 3.1.

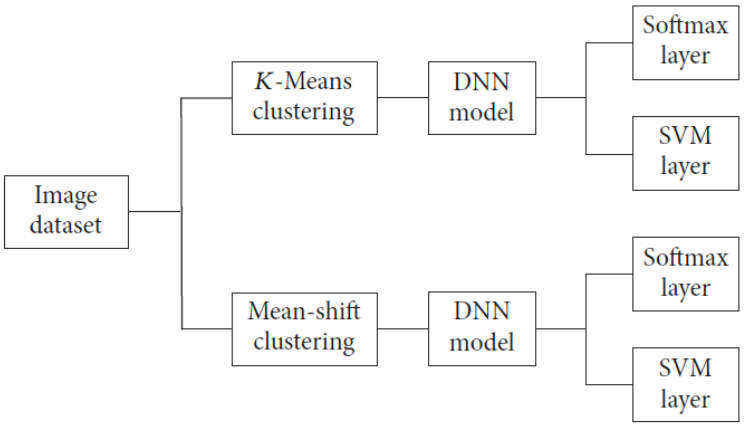

Figure 3.1. The image classification model (benign and malignant) can distinguish between healthy and cancerous tissue.

Images often keep both a local pattern and a hidden pattern that both signal the same sort of data. This is because images are stored in RAM. Histopathology slides are visual representations of the numerous different outcomes that might occur because of a biopsy. Photos in a biopsy collection that have the same taxonomy will often have similar forms of information saved in them. Unsupervised learning makes it possible to uncover patterns like these that could otherwise go unnoticed. The development of an innovative deep neural network model that is competent for breast cancer picture classification within the framework of a biological context is the primary contribution that this thesis makes. An unsupervised clustering algorithm serves as the engine that drives this model. In this piece, we discuss three innovative architectures for deep neural networks (DNNs): a convolutional neural network (CNN), a long short-term memory, and a combination of the two. The classification of the photographs is determined, in the end, by the classifier layer of the DNN model. This occurs after the

model has extracted both the local and global attributes from the photos. In this specific study endeavor, the Softmax function and the Support Vector Machine were used in their respective capacities as the classification layers (SVM).

3.2. FEATURE PARTITIONING

The images themselves include a significant amount of information, whether it be statistical or engineering-related. The process of modeling this kind of structural learning is an essential stage in the process of many various types of data analysis, one of which is picture categorization. The process of identifying structured information may make use of a number of different methods, one of which is the aggregation of unmoderated data. Because it enables you to split a vector of the same type into an area, grouping is an extremely useful operation. The method of clustering collects data sets that are of a similar kind and have comparable features and organizes them in a manner that emphasizes the differences between each set of data. There is a wide variety of options available to choose from when it comes to collecting methods. In this thesis, we employ the K-Means and Mean-Shift fuzzy sets techniques to analyze the data in order to uncover the underlying structure that is concealed inside it. The graphic presents illustrations of both benign and malignant tumors, in addition to images of the clusters that each kind of tumor produces. 3.2

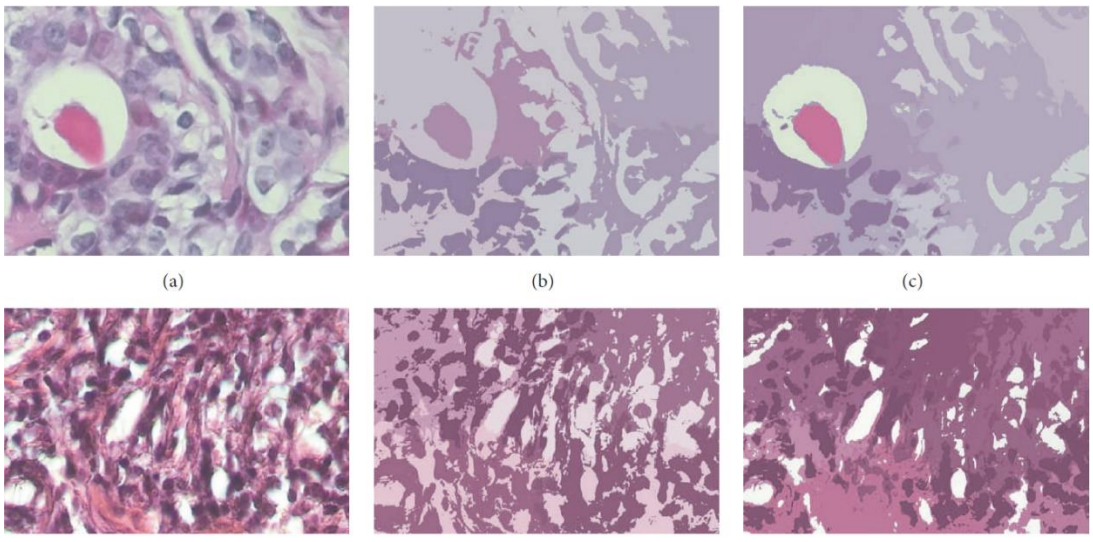

Figure 3.2. (a) A benign source scan, the KM loose collection image (b), the MS loose collection image (c), Malignant primary images (d), KM cluster-transformed images (e), and MS loose collection images (f).

3.2.1. Feature Scaling

Attributes in the dataset have a broad variety of values, thus feature scaling is employed to normalize them. In other words, data is made uniform through this process. Data should be normalized before conducting principal component analysis (PCA) since the algorithm's treatment of high- and low-variance variables is different. This article utilizes the sklearn. Preprocessing module from the sci-kit-learn library. The Python library Standard Scaler is used for standardization. In order to get the mean of a given sample, m, we use the formula:

$$z = \frac{m - n}{s} \tag{3.1}$$

3.2.2. Principal Component Analysis

Principal Component Analysis (PCA) is a statistical method that uses fewer variables to describe the same amount of information as the original set of variables. Principle component analysis (PCA) is a statistical technique that applies a discrete cosine transform to reduce a large number of associated variables to a smaller set of statistically independent linear variables called principal components [55]. After data normalization, PCA was employed by Random Forest, Support Vector Machine, K-Nearest Neighbors, Logistic Regression, and Gaussian Naive Bayes classifiers, but not ANN or CNN. Without the constraints of a linear model, you can learn to approximate any non-linear arrangement. The number of parameters used to define each note is decreased from 30 to 8 when principal component analysis is performed on a dataset.

3.3. VISUALIZATION OF DATA

3.3.1. Histogram

A histogram is a bar chart in which the bars have varied heights to indicate different ranges of data. Higher bars represent greater concentrations of data. Histograms are useful for depicting the distribution and shape of continuous sample data. Malignant

(M) and benign (B) tumors are shown in Figure 3.2 according to their histological classification. The prediction class includes 212 malignant tumors (about 38%) and 357 benign tumors (62%). A scatter plot of the characteristics of the nucleus compared to the diagnosis is displayed in the histological classification. Based on this, we can draw the conclusion that the median values of cell radius, perimeter, area, confluence, concavity, and points of concavity can be used to classify cancer. The presence of increased levels of these indicators is often linked to the development of cancer. It is unclear which diagnosis is chosen over another based on mean values for texture, smoothness, symmetry, or fractal dimension. This is because it is unclear which diagnostic best describes the patient's symptoms. There are no obviously huge outliers in any frequency distribution that require further cleaning.

3.3.2. Heatmap

For visualizing both basic and complicated data, a heatmap uses color to create a two-dimensional depiction. A heatmap is an effective tool for determining whether nodes within a set of values contain a disproportionately large amount of data compared to the rest. The use a heatmap to depict a correlation matrix. It is meant to illustrate the interconnectedness of this dataset's 30 characteristics.

3.4. DATA PRE-PROCESSING

3.4.1. Categorical Variable Conversion

The data set included both numerical and categorized information. The 'Diagnosis' column included a categorical indicator of whether the cancer was M = malignant or B = benign. The remaining characteristics are numeric in nature. Practically all algorithms function more effectively when dealing with numeric variables. Attribute values cannot be directly fitted into a regression model and the python module "sklearn" needs features in mathematical arrays. To do this, we utilized Label Encoder to convert our textual labels into numerical ones.

3.4.2. Data Reshaping

For a CNN to function properly, additional processing in the form of data reshaping is required at its input. The data collection was originally a 2-dimensional shape, (569, and 30). Specifically, the NumPy module is used to transform the data into a three-dimensional shape (569, 10, 3) suitable for CNN.

3.4.3. Train-Test Split

Most of the time, material is split into two parts: training data and test data (and sometimes into three sets: training, validation, and test). The term "training dataset" refers to the real data used to educate the model (weights and biases in the case of a neural network). Model judgments are influenced by this information. To objectively evaluate how well a final model fits the training dataset, it is useful to first create a subset of the dataset known as a test dataset. When tweaking model parameters, it is usual practice to first assess how well the model fits the validation dataset before moving on to the training dataset. The split ratio depends on both the overall number of samples and the actual training model. In order to divide the training data from the test data, we used the scikit-model-group-training package's training and testing separation class with a 70:30 split. Seventy percent of the information was put into the training set, while the remaining 30 percent was used as the test data. This proportion is ideal [111]. The sklearn intercept in the test data. The RF, SVM, KNN, LR, and GNB accuracy, correctness, recall, and F1 scores were analyzed using the cross-validation package. Ten individuals served as assessment subjects.

3.4. DEEP NEURAL NETWORK

Today, deep neural networks are widely used in state-of-the-art methods for data categorization and analysis. Convolutional neural networks (CNNs) and recurrent neural networks (RNNs) are the most typical varieties of DNNs (RNNs). They have helped pave the way for many future developments in data analysis. In the following paragraphs, we'll discuss the inner workings of CNNs and RNNs, with a special emphasis on long short-term memory (LSTM) algorithms. Because of their emphasis

on the binary categorization of the grouped data, these approaches were selected for the model. Since there is no one-size-fits-all solution, however, deep learning models have also been tried out. These include artificial neural networks and convolutional neural networks. We decided that Random Forest would be the best approach for this ML problem (RF). Machine learning techniques include support vector machines (SVMs), support vector networks (SNNs), k-nearest neighbors (KNNs), logistic regression (LR), and gaussian naive Bayes. We compared each algorithm's output to identify the best classifier for this particular issue.

3.4.1. Convolutional Neural Network

The CNN model enhances the decision-making capabilities of a regular neural network by extracting local and global characteristics from the input data through a linear transformation. CNN features multiple intermediary layers in addition to the convolutional layer for fine-grained control. The details are expanded upon below. Some of the most common uses for the CNN deep learning model are in image and video recognition, recommendation systems, image classification, medical image analysis, and natural language processing. Compared to other image classification methods, CNNs need very minimal pre-processing before they can be put to use. While there are many varieties of neural networks, CNNs have several unique characteristics. To begin, the layers exist not only in the plane but also in the vertical and horizontal dimensions. Furthermore, interlayer connectivity between neurons is weak. Common examples of such layers are the fully connected layer, the activation layer (also known as ReLU), and the max-pooling, all of which modify inputs to better understand the data. The algorithm's output will be the likelihood that a picture feature corresponds to the prediction. In this research, I modeled a convolutional neural network using data that was first converted to three dimensions. Despite skepticism that emerged from the very small sample size, the findings proved otherwise. The CNN diagram is shown in Figure 3.3.

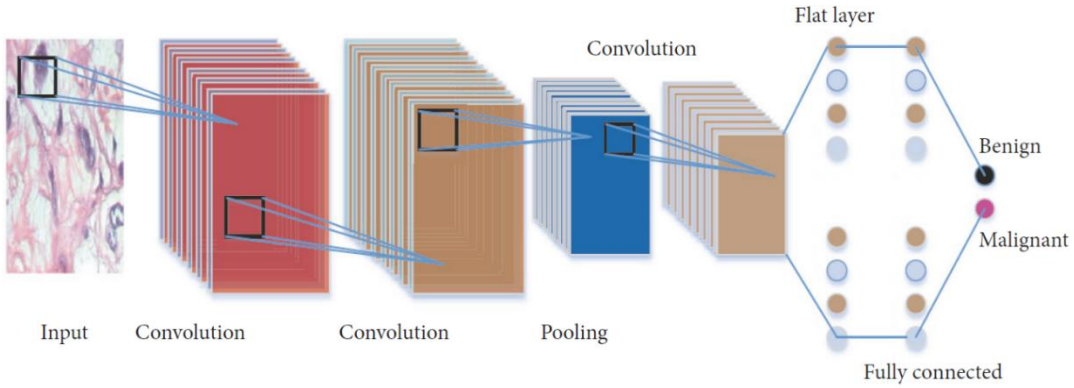

Figure 3.3. Workflow of a Convolutional Neural Network [86].

3.4.2. Evolutional Layer

This CNN model has relied heavily on A, which has been seen as its primary advantage. The feature map is the result of convolving the values at each position $(w1.w2)$ of the input data $(w1.w2)$ with the kernel Kk1k2. For a specific point $(w1.w2)$ in the input data Im1m2, the convolutional output of layer l and feature t can be expressed as:

$$I_{w1.w2} * K_{k1*k2} = \sum_{i=(-k_1+1)/2}^{(k_1-1)/2} \sum_{j=(-k_2+1)/2}^{(k_2-1)/2} I_{w1-i.w2-j} * K_{ixj} \quad (3.2)$$

After adding the bias term $\beta^{(l.t)}$ the previous equation will be

$$F^{(l.t)} = (I_{w1.w2} * K_{k1*k2}) + \beta^{(l.t)} \quad (3.3)$$

Each neuron gives forth a linear voltage signal. If you take one neuron's output and feed it into another, you'll get another linear result. Linear regression that are not linear, such:

- Sigmoid
- Tan-H

- ReLU
- Leaky-ReLU has been introduced.

Figure 3.4 (1) represents the Sigmoid function characteristic which follows the equation:

$$\sigma(s) = \frac{1}{1 - e^{-s}} \tag{3.4}$$

This intelligence challenge has a lot of computing costs and may have issues with vanishing gradients. Tan-H is a non - linear activation product that is just a scaled version of the (s) operator, and it may be used in the same as:

$$Tan - H\ (s) = 2\ \times\ \sigma\ (s) - 1. \tag{3.5}$$

This, as seen in Figure 3.4, has the ability to circumvent the vanishing-gradient issue, which is displayed (2). Rectified Linear Unit (ReLU), shown in Figure 3.4(3), is the most often used nonlinear operator due to its ability to remove unwanted details.

$$ReLU\ (s) = maximum\ (0.s) \tag{3.6}$$

Figure 3(d) shows the Leaky-ReLU rectifier's characteristics, which is a modification of ReLU:

$$Leaky - ReLU\ (s) =\ \sigma(s) +\ \beta\ \times ReLU\ (s). \tag{3.7}$$

where β is a parameter that has already been determined. In the convolutional layer, the kernel is the most important part of the layer. It is the job of this component to examine all of the incoming data and make an effort to extrapolate global features from those analyses. The term "stride" refers to the overall number of steps that a kernel works through throughout each iteration of the process. It is likely that the border row and column positions will not be convolved in the right manner in the case that we pick faulty stride steps and size. The practice of adding extra rows and columns that

are completely filled with zeroes is referred known as "zero padding." This is done in order to ensure that the convolution operation can be performed accurately at the boundary of the image.

The multilayer model is responsible for the generation of a significant amount of information on the features. As the model structure becomes larger, the amount of feature information likewise grows larger, which in turn enhances the level of computational complexity of the model and makes it much more sensitive. Organization for Research in Biomedicine To get around problems of this kind, we have devised a system of random sampling, which is as follows:

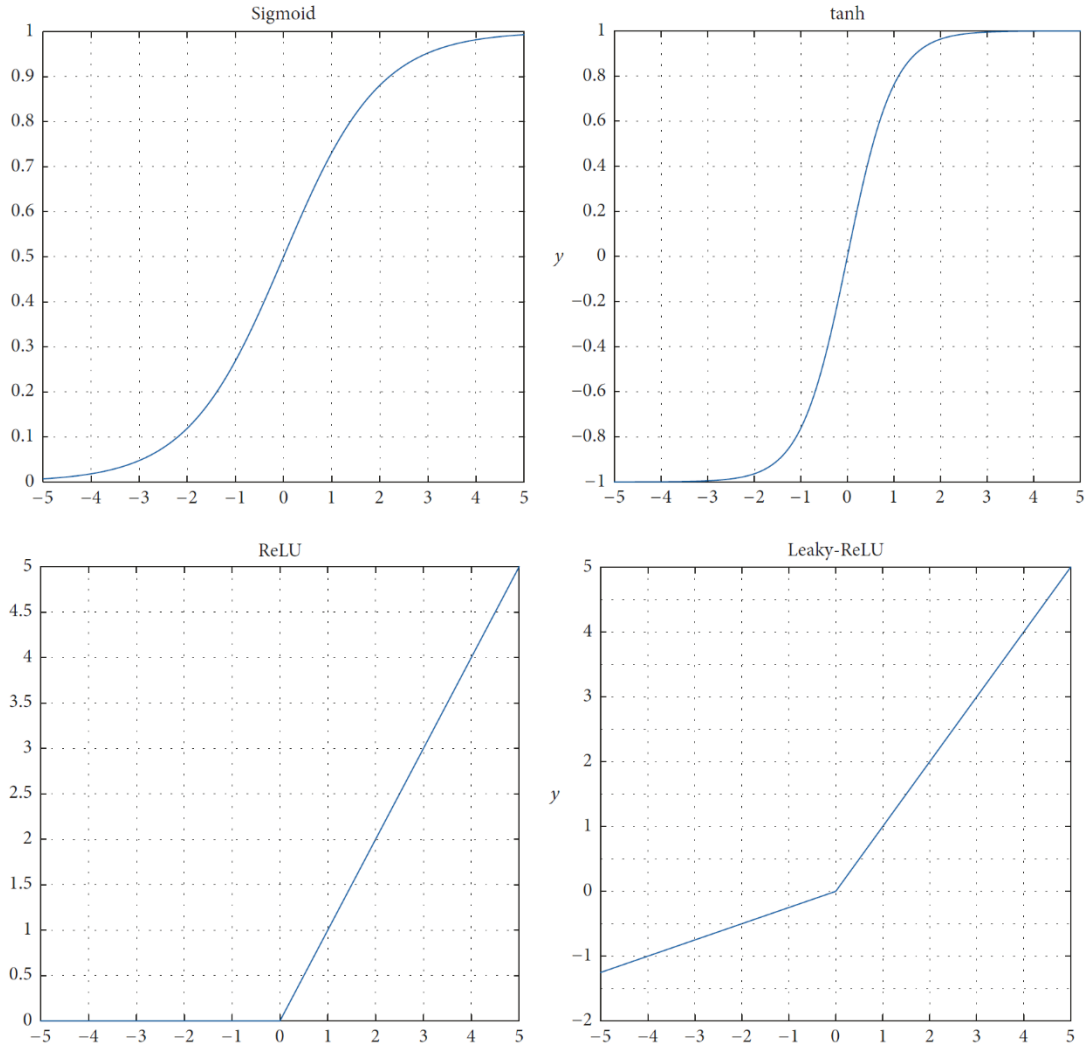

Figure 3.4. Sigmoid, TanH, ReLU, and Leaky-ReLU.

Subsampling: Down-sampling, also called pooling or dimension reduction, is a process that may be used to decrease the dimensionality of the features. In the end, it brings the entire multiplicity and complexity down to a lower level. An example of a generalized pooling technique for a CNN model is shown in Figure 3.5.

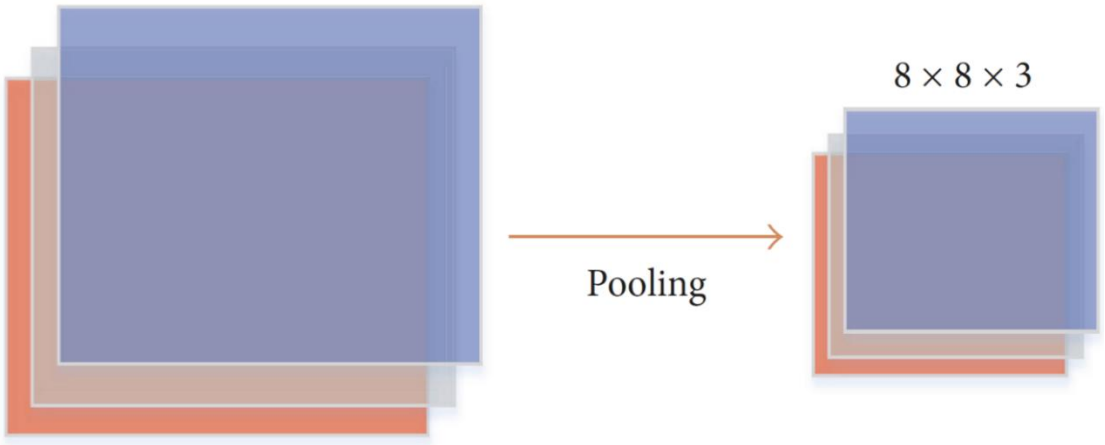

Figure 3.5. Pooling operation performed by 2 × 2 kernels.

There are four different kinds of pooling operations available:

- Max-Pooling
- Average Pooling
- Mixed max-average pooling
- Gated max-average pooling.

With so many neurons to manage, a DNN may steer in a path that takes into account many different predictions. Results are excellent on the training set but suffer on the test set when this occurs. One may speak of an overfitting issue when this occurs. The dropout process was established to address this specific issue. Figure 3.6 shows more information about the dropout.

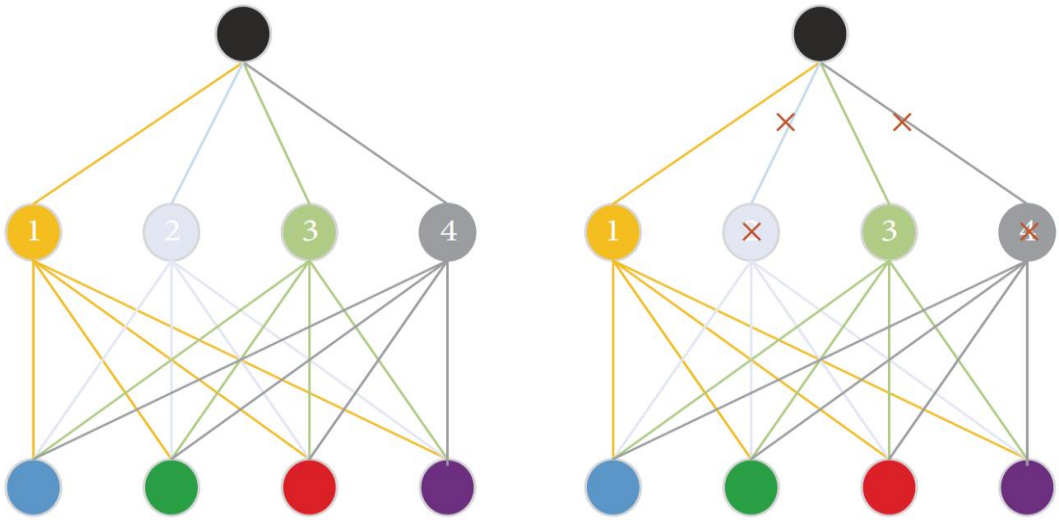

Figure 3.6. Diagram of Dropout.

Dropout: As a solution to the overfitting issue, certain neurons are selectively destroyed at random. In this process, some of the neurons are removed at random (with a certain preset probability) so that the network may learn more stable characteristics. A basic version of a dropout mechanism is seen in Figure 3.5. Four hidden neurons numbered 1 through 4, are visible in the right-hand picture, whereas neurons 2 and 4 have been eliminated in the left-hand image so that they do not contribute to the network's final judgment. Eventually, the neurons in a network will all be laid out in a flat, uniform fashion. All of the flat layer's neurons communicate with each other and the layer above them, just as they would in a regular neural network. It is common practice to implement many completely linked layers. If you think of the very final layer as the "end" layer, then you need at least one flat layer or completely linked layer before the very last one. If this is the case, the last layer's function may be written as:

$$F_k^{end} = \sum_{j=1}^{end-1} w_{k.j}^{end} F_g^{end-1} + \beta_k^{end-1} \tag{3.8}$$

In Figure 3.3, we see an example of a generic convolutional neural network (CNN) model used for image categorization. The terminal layer represents the ultimate point of decision-making.

Decision Layer: Both the Softmax-Regression method and the Support Vector Machine method are used in the determination layer. Calculations of cross-entropy losses, including:

$$L_k = -\ln(\overline{y_k}). \tag{3.9}$$

Where the $(\overline{y_k})$ can be written in equation 3.10

$$(\overline{y_k}) = \frac{\exp(F_k^{end})}{\sum_{k=1}^{2} \exp(F_k^{end})} \tag{3.10}$$

Here $k = \{1, 2\}$ where 1 is for benign and 2 is for the malignant case. The value of Lk provides the final decision such as if $L1 > L2$ the network will produce malignant output.

3.4.3. Support Vector Machine

Support Vector Machine (SVM) is a technique for supervised machine learning that has found widespread use in the discovery and categorization of patterns, particularly in circumstances in which a dataset consists of just two unique classes. SVMs are put to use in the process of locating the appropriate hyperlevel for dividing the classes. The input pattern, which is also referred to as a feature vector, is evaluated by the classifier in order to establish the appropriate category for the pattern to be placed in. This strategy is only useful for the classification of data that can be linearly separated; in actuality, however, feature vectors cannot be linearly separated. The kernel technique is an approach that may be used to resolve this issue [105]. The Support Vector Machine (SVM) is a kind of machine learning technology that uses kernel techniques to map input data to multidimensional data. In addition to that, it takes the form of rapid learning. Its applications include, among other things, pattern categorization and regression analysis. It is essential to keep in mind that selecting the appropriate kernel function may have a significant impact on the SVM classifier's overall performance. Depending on the categorization problem at hand, the appropriate kernel function will be selected. In order to implement SVM for this project, the SVC

class found in Scikit-learn was utilized. On the other hand, SVMs might need a lot of memory and can be difficult to comprehend and adjust. An SVM [107] layer may be used in place of a Softmax layer, with the inclusion of the following requirements. Let $x = x1, x2, \ldots, xn$ represent the training data, and let $y = y1, y2, \ldots, yn$ represent the appropriate label for a generalized case. If we assume that the data can be divided into linearly distinct parts, then the optimization constraint will be $y_i w^T w_{x_i.}$ less than 0. On the other hand, there are situations when the data cannot be separated linearly. In these instances, soft thresholding has been implemented, and the constraint has been redefined as $y_i w^T w_{x_i.}1$ I where I equal 0. The optimization issue has been rethought as of now. As

$$minimum_{w.\xi i}\ \frac{1}{2}\ w^T w + B \sum \xi_i\ \ s.t\ \ \xi_i \geq 1 -\ y_i w^T w_{x_i.}\ \xi\ \geq \forall_i \tag{3.11}$$

3.4.4. CNN-LSTM

One of the strengths of CNN is that it compiles news and information from a variety of sources from across the globe. On the other hand, the long-term memory model, often known as LSTM, is able to make use of the long-term dependencies that are present in data patterns. In order to improve classification, CNN and LSTM models were merged together and used jointly [112–115]. In order to make advantage of these two functionalities, this step has to be taken first. It is difficult to generate an undirected graph from the output of the CNN model in order to convert the information into a time series format, which is necessary for the network to extract the relationships that exist within the data. This is because the network needs to convert the information into a time series format. In order to do this, we start by decreasing the dimensionality of the data coming from the convolutional output, which was just two-dimensional to begin with. Figures 3.7 and 3.8 below illustrate the fundamental architecture of the LSTM and CNN models, respectively.

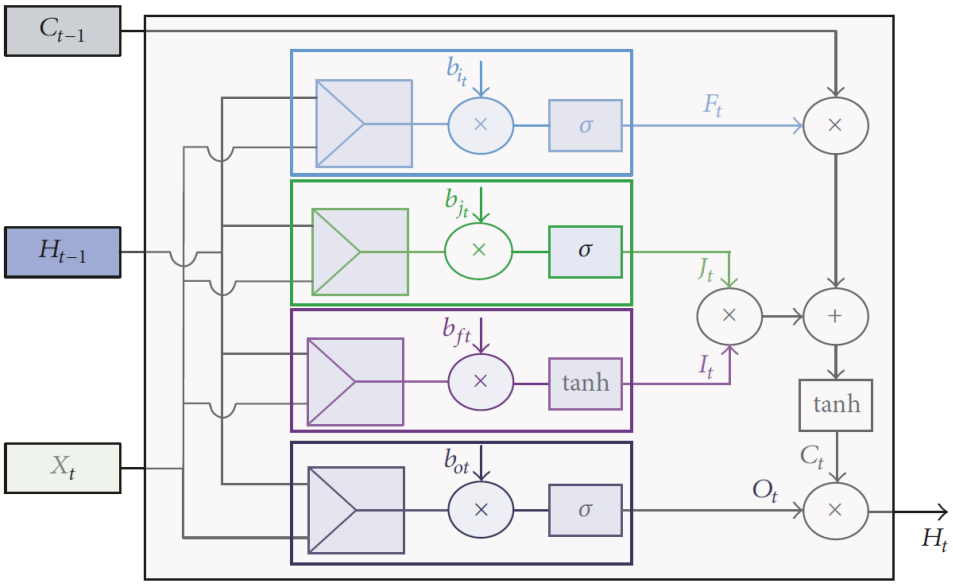

Figure 3.7. A generalized cell structure of an LSTM.

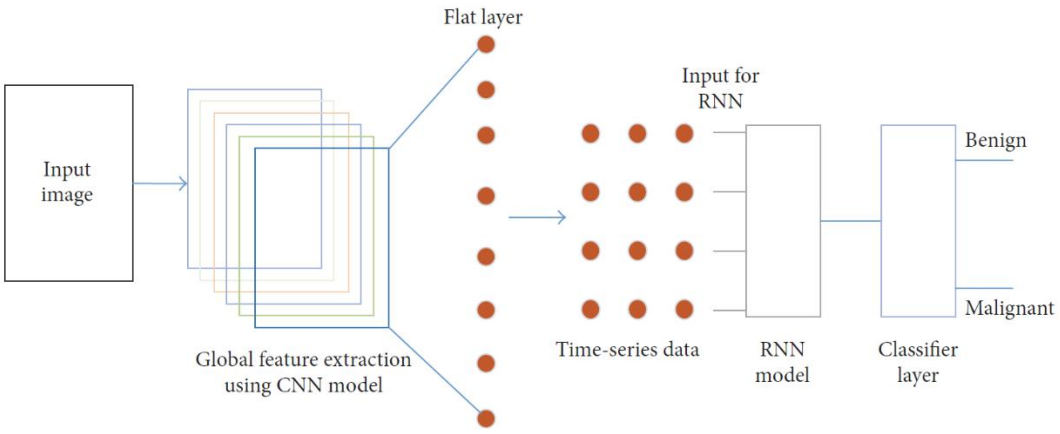

Figure 3.8. CNN and LSTM models combined.

3.4.5. Random Forest

It is generally agreed that Random Forest is one of the algorithms for machine learning that is both the most effective and the most popular. It is made up of a few distinct decision trees, and the manner of class inference is established by using various trees to select which class should be made. The forest is produced as a consequence of creating decision trees based on groupings of samples, which ultimately results in the development of numerous trees. This process is known as "forest generation." This is helpful for classification issues such as these, as well as other challenges like as regression, which functions in the fashion that was explained above, constructing

numerous trees during training and outputting classes or mean predictions for each given tree. This approach is helpful for dealing with issues of this kind. After having several deep decision trees learned on separate subsets of the same data set, the outcomes of these trees are then averaged in order to limit the amount of variation [44].

3.4.6. K-Nearest Neighbors

Classification and regression may be accomplished with the help of a non-parametric technique called the K-nearest Neighbor approach [70]. In spite of its utility, it has a foolproof simplicity that belies its complexity. The findings that are utilized to create the model are generated by using the training set as the data source. These findings have some connection to the function of distance, as well as the function of class selection. A similarity scale is used in the process of categorizing a newly discovered element by first comparing it to elements that are already present in the periodic table. After that, the elements' k closest neighbors are taken into consideration, and the class that appears the most often among these neighbors is selected to serve as the representative for the element that is to be classed. Distance is used as the primary factor in determining neighbor weighting. KNN may need a considerable amount of memory or space to store all the information, but it will only do computations when it needs to make a prediction. This means that it is more efficient than other methods of data storage [73].

3.4.7. Logistic Regression

Logistic regression is one more example of a supervised learning approach that deep learning has appropriated from the discipline of statistics. In contrast to Linear Regression, Logistic Regression uses an independent variable as its output or goal variable. As a result, Logistic Regression is a binary classification technique that assigns each data point to one of the classes represented by the data [10]. The following is the standard equation used in logistic regression:

$$\log\left(\frac{p(s)}{1-p(s)}\right) = \beta_0 + \beta_1 X \tag{3.12}$$

In this equation, $p(s)$ represents the dependent variable, X represents the independent variable, β0 represents the intercept, and β1 represents the slope coefficient. The values of the input (X) are linearly blended with the values of the transaction (Y) in order to determine what the value of the output will be. The output value of the simulation is a binary value, which means that it may either be 0 or 1. Unlike the output value of linear regression, this value cannot be a numeric number. This is the primary distinction between the two of them. The range of possible values for the evaluation value extends from positive infinity to negative infinity, but it must be within the range [0, 1]. As a consequence of this, the transformation is carried out with the assistance of a logistic function, which is also referred to as a nonlinear function. This curve has the shape of an S, and it transforms every real number into a value that falls between 0 and 1. It is extensively used for issues involving binary classification, and it functions more effectively when the number of qualities that are connected to one another is reduced. It sheds light on the findings in the subsequent section of the examination of the data. This model may be taught in a short amount of time and functions effectively when combined with a binary classifier to solve issues.

3.5. NEURAL NETWORK LAYERS

There is just one input layer and one hidden layer in each network. The processed data contains as many input variables as there are neurons in the input layer. Each input has as many corresponding outputs as there are neurons in the output layer. The challenge, however, is in figuring out how many hidden layers there are and how many neurons they contain. The input layer does nothing more than relay data to the hidden layers below it; it does not carry out any computations of its own. However, the input nodes provide data to the output nodes, and the computations take place in the hidden layer. There will be just one visible layer of input and one visible layer of output, but the network may contain any number of hidden levels. He is in charge of the output layer, which is in charge of calculating and delivering data from the network to the outside world. In a typical convolutional neural network, the hidden layers consist of convolutional layers, pooling layers, recurrent neural network, and normalizing layers. Each of the CNN and ANN models in this research uses its own unique collection of

layers, including hidden layers and activation functions like ReLU, Sigmoid, and Softmax. A refined RNN model is shown in Fig. 3.9.

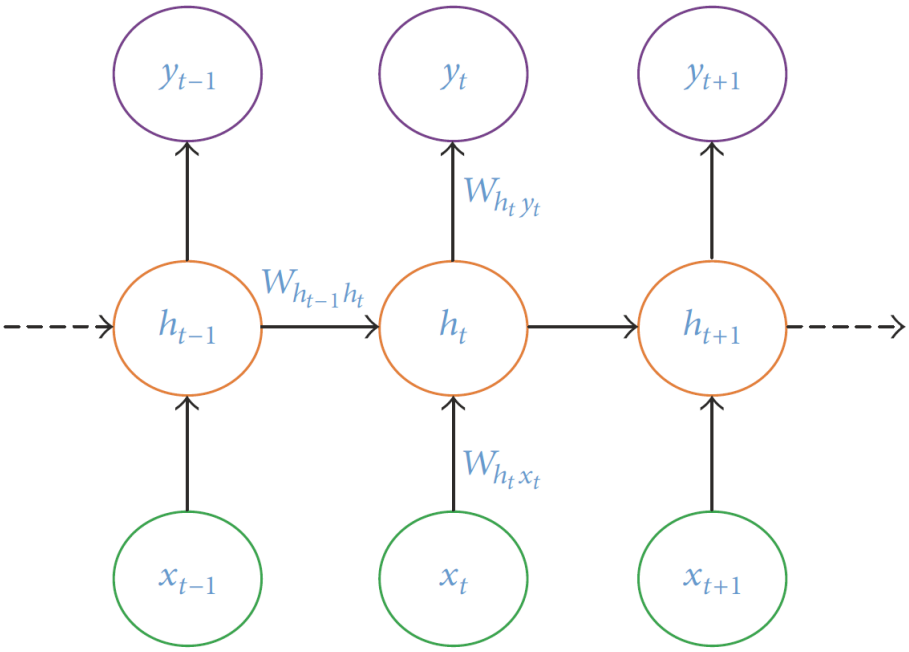

Figure 3.9. A variant of the RNN paradigm that assumes that the hidden neuron both generates the RNN output and receives the data source.

3.6. PROPOSED MODEL

We used a total of three different models in order to carry out the analysis of our data (Figure 3.10). In addition, the Wisconsin Breast Cancer (Diagnostic) Data Set (WBCD) [108], which included a total of 569 patients, was subjected to a number of the leading algorithms in the field in order to improve the accuracy of the early identification of breast cancer. Using a variety of different performance indicators, the outcomes of these apps were evaluated and contrasted with one another. The CNN techniques are used in Model 1, whereas the SVM structure is used in Model 2. The evaluation of the data is performed concurrently using both the CNN and SVM structures.

Model 1. This technique employs a ReLU activation filter at the C-1 level, with the input picture being processed by 3 × 3 cores. For the sake of consistency, we create two new rows and columns with a value of '0' for each core every time it progresses

one step. This guarantees that the freshly generated feature maps also have identical 32 x 32 measurements. After the newly constructed, completely interconnected Layer C-1 comes Layer C-2. It's got the same ReLU rate as the last one and uses a 3x3 basis. A P-1 merging with a kernel size of 2 2 is executed immediately after level C-2. We were able to lower the sizes of both feature maps by opting for a 22-kernel size. The grid size was reduced by half, from 32 by 32 to 16 by 16. The P-1 layer is followed by a C-3 layer, a ReLU rectifier convolutional layer. Layer P-2, a pooling layer with 2x2 kernels, was used to reduce the size of the original 16x16 layer C-3 feature map to 8x2. C-5, the last, completely linked layer, was created after another merging, P-3. The alignment process is carried out when the conv layer is output. There are 256 features in the homogenous layer since each of the 16 feature maps in layer C-5 is 4 by 4. The decision level (SVM/Softmax) was responsible for assessing whether the tumor was benign or malignant, and it received 75% of the information that had been excluded at the exclusion level.

Model 2. In the second iteration, we decided to make use of the SVM technique, which is a component of the CNN model. The picture that we are dealing with right now just has a two-dimensional representation. Our data comes in the form of a matrix, and we have to transform it to the one-dimensional format required by the SVM model. The data vector that we get out is 30721. The one-dimensional data was turned into a time series by our team. We display the data for each time step (TS) from x1 to xu, where the input dimension for each TS is V; that is, we plot the data from c1 to cV, where V u = 3072, such that the value 3072 may be located anywhere within the time series. data. We made use of a two-layer SVM design, with the L-1 layer and the L-2 layer being layered over one another. The output of the SVM-2 layer is responsible for the generation of a total of 42 neurons. The output of the SVM layer is discarded, although there is a one in four chance that it will be reused. Following the addition of the dropout layer, a substantial layer of up of 22 neurons was added. In the end, the classification of disorders as benign or malignant was accomplished by the use of a decision layer.

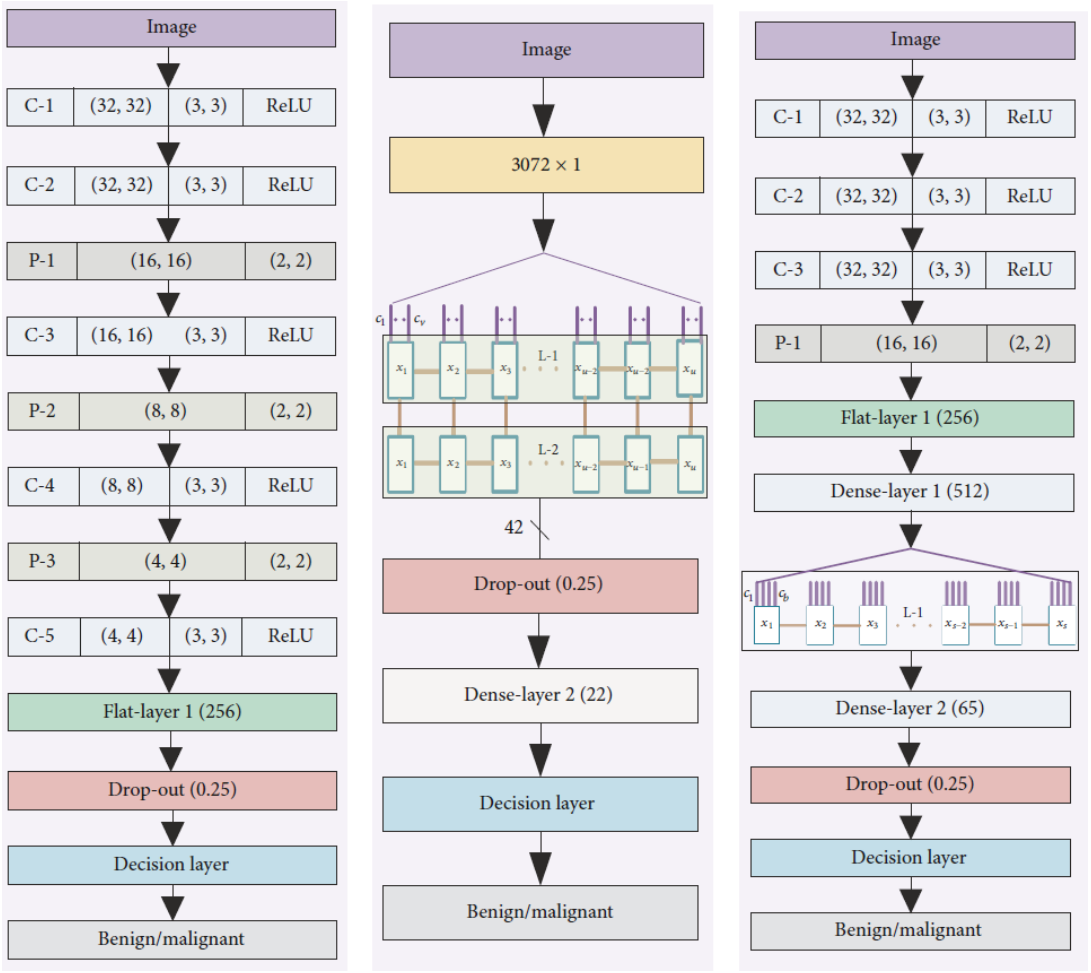

Figure 3.10. Conventional CNN, SVM-based architecture (a, b), and CNN-SVM-based architecture (c).

Model 3. We have achieved this by combining the capabilities of CNN and SVM models. The first layer of the convolutional network, designated C-1, is a convolutional layer that operates on the input picture using a 3x3 kernel and a ReLU modifier. A size of 32 × 32 pixels is assigned to each feature vector that is produced by this layer. On top of the initial layer, two further layers, labeled C-2 and C-3 correspondingly, were each applied in turn. After the C-3 layer comes a pooling layer known as P-1, and the size of its kernel is 22. Due to the fact that the pooling layer employs 22 kernels, the output of P-1 is 1616 kernels. The P-1 layer is followed by the addition of a flat layer, and then 512 neurons are generated by a dense layer. The layer's output is then sent on to the LSTM network to be used as its input layer. Due to the fact that this layer only has one dimension, the vector data that lies underneath it will need to be converted into time series. We created TS data from registers x1 all the way

up to register xs, and each of these registers has an identifier of size q, say c1 all the way up to cb, where s b equals 512. A layer that consists of 65 neurons was formed as a continuation of the SVM layer that was previously created. After then, 25% of the entries were disregarded as invalid. After that, a decision layer was put in place to differentiate between data that may be helpful and data that could be detrimental.

PART 4

EXPERIMENTAL RESULTS AND DISCUSSION

We used the BreakHis dataset [3] and the Breast Cancer Classification Challenges 2020 dataset for binary classification to showcase the efficacy of the CNN-based SVM models and other machine learning models for BC (benign or malignant). The datasets are described in further depth in the next paragraph. Keras [55] and Tensor Flow [70] were utilized for this implementation, with 56GB of RAM and an NVIDIA GEFORCE GTX-980 Ti packed into a single GPU system. In this implementation, we explored an alternative criterion for pathological picture analysis. Whole Slide Images (WSI) often have bigger sizes than standard digital photographs. Additionally, various magnification levels are used to collect diseased pictures. There are situations when the image size exceeds 2000px by 2000px. In this scenario, however, the photos are often supplied to the model in the form of numerous patches. The random crop technique is one of the two most often used procedures for patch selection. In this approach, patches are selected at random from an input sample. The use of non-overlapping, successive patches is the other option. When designing this system, we gave equal consideration to both approaches.

4.1. DATASETS

The proposed strategy was evaluated using the BreaKHis dataset, which consisted of 7909 photos gathered from 82 different patients who wished to remain anonymous. These pictures were all scanned at a resolution of 700 × 460 pixels, which is the same as before. The BreaKHis dataset was separated into benign and malignant tumors. The BreaKHis dataset is freely accessible to the public and is often used for the purpose of researching the breast cancer classification challenge. This dataset consists of 7909 samples that may be divided into two primary categories: benign and malignant. The benign group has a total of 2440 samples, while the malignant subset has a total of

5429 samples. The samples were taken from 82 distinct individuals using a range of magnification factors, including 40x, 100x, 200x, and 400x respectively. A few of the sample photographs that have been used with a magnification factor of 400. Adenosis (A), Fibroadenoma (F), Tubular Adenoma (TA), and Phyllodes Tumor are the four different forms of benign cancer. Each class is further divided into four subtypes (PT). There are four distinct subtypes of cancer that are considered to be malignant: ductal carcinoma (DC), papillary carcinoma (PC), lobular carcinoma (LC), and mucinous carcinoma (MC) (PC). Table 1 contains the statistical information on this collection of data. According to the research in [80,100], we conducted this experiment with 70% of the samples being used for training and 30% of the samples being utilized for testing. We guarantee that the patients that were picked for training are not utilized during testing because we want to be able to generalize the classification task so that we can complete it effectively when testing new patients. In accordance with the experimental methodology presented in [3], we reported the average accuracy after successfully completing five trials. Figure 4.1 provides many instances of BC images, each of which depicts a patient either with a benign tumor or a malignant tumor.

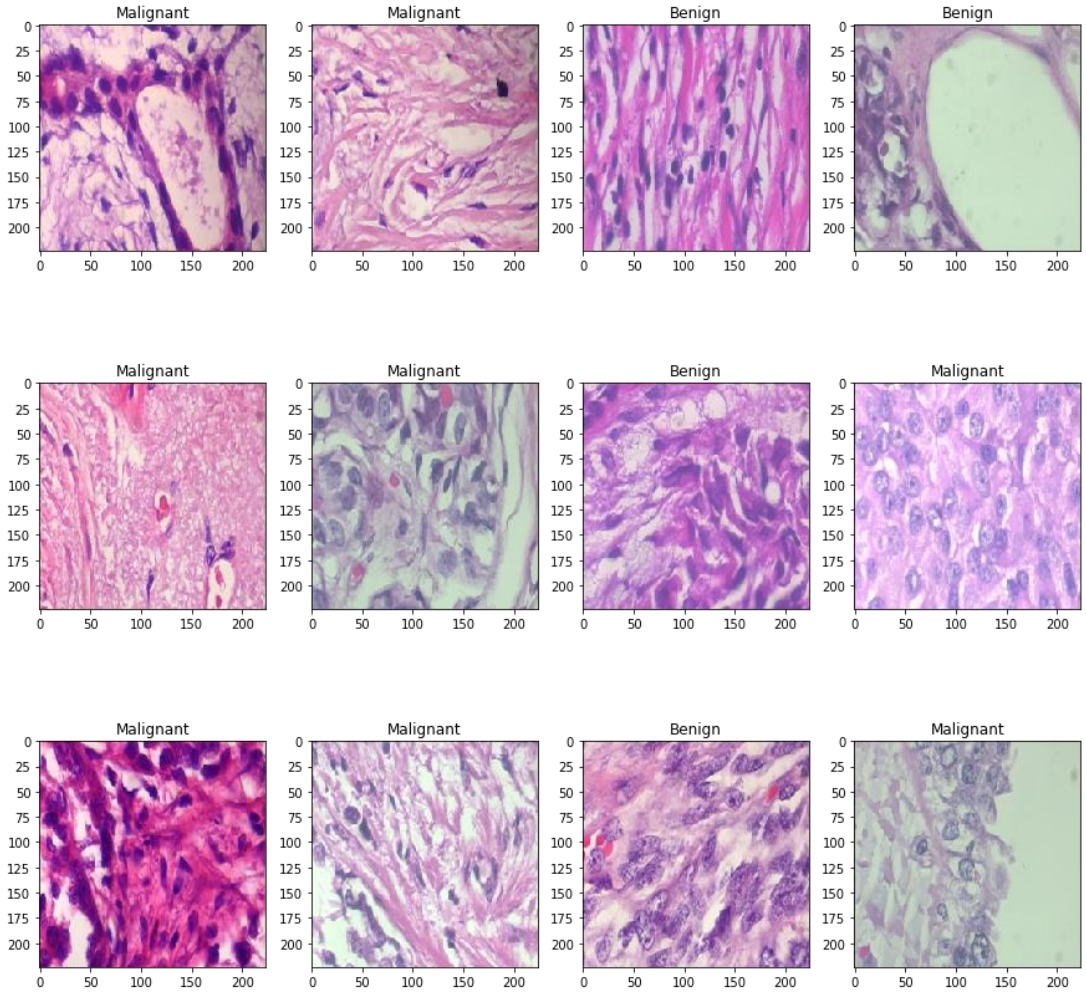

Figure 4.1. Examples of BC images (Benign and Malignant).

4.2. DATA AUGMENTATION

Data from each set was sequentially rotated by 90 degrees, shifted in width by a factor of 0.2, shifted in height by a factor of 0.2, sheared by a factor of 0.2, zoomed in and out by a range of 0.2, flipped horizontally, and flipped vertically. Examples of both types of data are shown in Figure 4.1, along with a variety of enhanced examples. Figure 4.1 shows that there is artificial noise in some of the photos. As a result, we have additionally tested our approach by removing the outside regions from the enhanced samples and focusing on the inner ones. Figure 6 displays the down-sampled and central patches for two distinct input samples. We compared the efficiency of binary breast cancer identification on both the picture and patient levels.

In most cases, the best results and highest accuracy rates are achieved by training on a large number of practice examples. Unfortunately, the quantity of patients means that biomedical datasets have few samples. To that end, data augmentation is any technique for expanding an existing dataset by creating more records from the existing ones. In this publication, we use a technique called rotation to add new information to existing data. The original images are rotated by degrees (0, 90, 180, and 270). That's why we add two more copies of each picture to them.

4.3. RESULTS AND DISCUSSION

In this dissertation, we provide an automated method for identifying breast cancer using two separate datasets. As a result, this version of the binary BC classification issue only takes into account two classes. By combining CNN with SVM and other machine learning models, we improved upon the state-of-the-art testing accuracy for both datasets. There is a correlation between the CNN-based SVM classification performance on the original dataset and the augmented dataset. Each experiment's findings are reported in terms of accuracy (the percentage of properly labeled cases). There are also provided learning curves for both testing and training losses. We also tried the 'CNN Features Classifiers' configuration, which involved taking the 2,000 outputs from the fully-connected layer and classifying them using a K-nearest-neighbor (KNN) classifier, a radial basis support vector machine (RBF SVM), a linear support vector machine (SVM), a random forest, and a 'CNN Features Classifiers' configuration. Activations for input pictures are generated in the layers of convolutional neural networks. These layers may be mined for a collection of characteristics that can be fed into more conventional classifiers. By using the fully connected layer's 2000 outputs to extract features, we used the 'CNN Features Classifier' setup as the basis for feeding those features into K-nearest neighbors, support vector machines, and random forests. The number of neighbors in K-nearest-neighbors is 3. Linear and radial basis kernels were used to evaluate SVMs. Best accuracy was achieved using radial basis kernels with parameters 5 and gamma (1/number of features). A total of 50 trees were employed in the random forest creation process, with the Gini index serving as the split criteria for the characteristics.

All possible scaling factors of the enlarged dataset are tested using the features retrieved from the convolutional neural network. TABLE 11 summarizes the results of the classifiers in terms of their ability to distinguish between several classes. By analyzing the results of the classifiers used for the binary classification, we found that linear SVM yielded the greatest accuracy rates across all four magnification levels. This is because the training data is split at a decision boundary (the data distribution is linearly separable). For multi-class classification, RBF kernel performed best on 40X and 100X photos, KNN performed best on 200X images, and linear SVM performed best on 400X images. So, the data's qualities and the classifiers' discriminatory power are crucial to their success (Benign and Malignant).

4.3.1. Results for BreakHis

We used two metrics to assess the IRRCNN model's efficacy, drawn from the research in [3]. We looked at (1) multi-class classification performance on images and (2) multi-class classification performance on patients for the eight subtypes of breast cancer (either benign or malignant). We have also tested how well a binary class system distinguishes between healthy and unhealthy cases. We disregarded patient context entirely when classifying photos. Images are divided into eight groups, each of which represents a different magnification level for the experiment. In this context, performance is judged using alternative assessment criteria. We examine the effectiveness of the CNN-based SVM machine learning strategy using two distinct performance metrics.

The efficacy of the CNN-Based SVM method for image classification was measured. You can see the produced CNN-Based SVM model shown in Figure 4.2, and its accuracy and loss rate in Figure 4.3, both during training and testing.

Layer (type)	Output Shape	Param #
densenet201 (Functional)	(None, 7, 7, 1920)	18321984
global_average_pooling2d (GlobalAveragePooling2D)	(None, 1920)	0
dropout (Dropout)	(None, 1920)	0
batch_normalization (BatchNormalization)	(None, 1920)	7680
dense (Dense)	(None, 2)	3842

Total params: 18,333,506
Trainable params: 18,100,610
Non-trainable params: 232,896

Figure 4.2. The structure of the trainable model

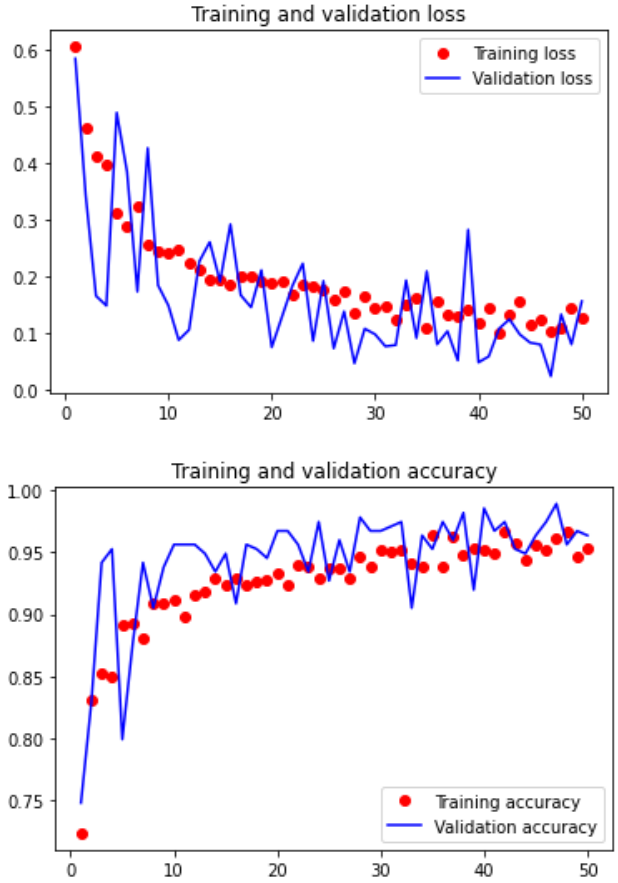

Figure 4.3. Training and validation accuracy for BC classification with 2 classes for the CNN-Based SVM model.

The used tested image is 175 for the Benign tumor and 195 for the Malignant tumor, Figure 4.4 and Figure 4.5 shows the Confusion Matrix (CM) with normalization and without normalization.

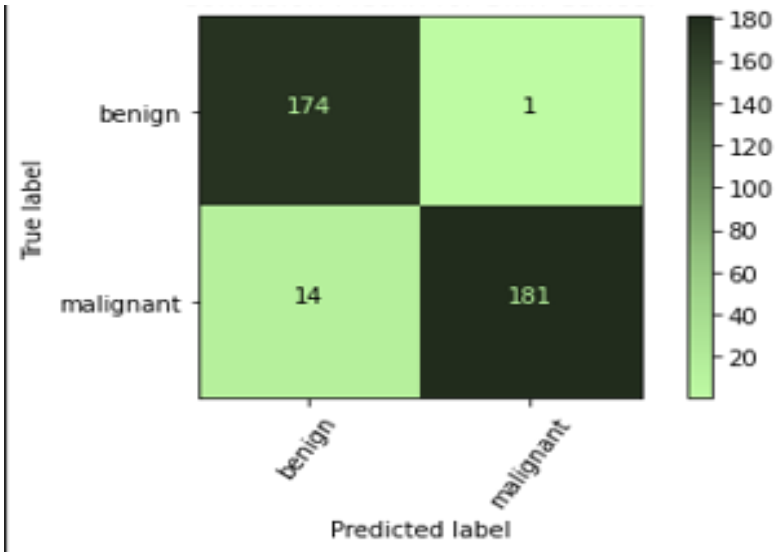

Figure 4.4. CM results without normalization.

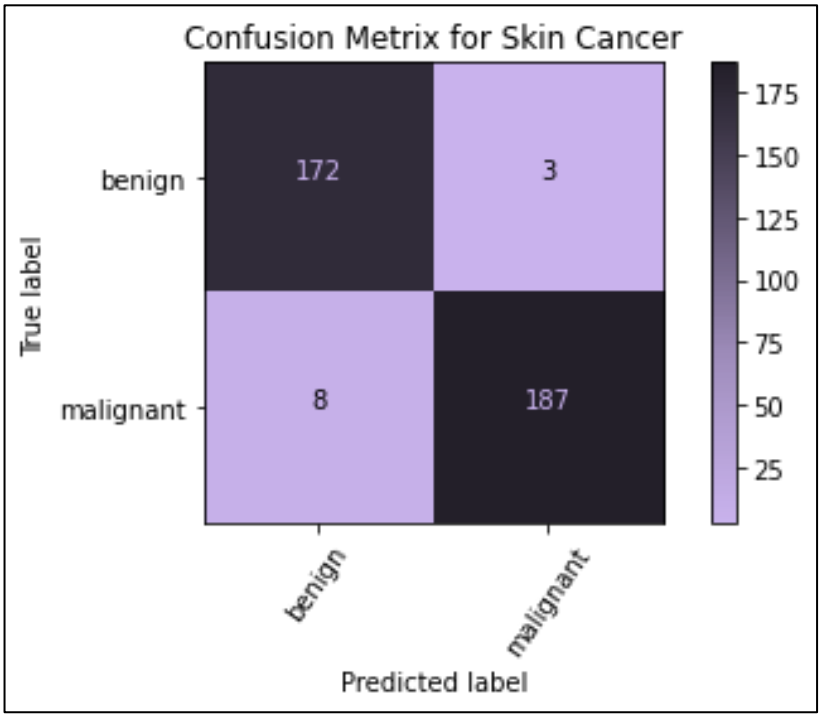

Figure 4.5. CM results with normalization

Figure 4.4 depicts an experiment in which 370 data points were utilized to evaluate a classifier's efficacy. Ninety-five of these 370 samples are cancerous, whereas the other

175 are benign. While the suggested classifier correctly identified all of the benign samples in this experiment, it incorrectly identified one of the malignant samples. Overall, this model was 96.6% accurate.

Figure 4.5 depicts an experiment in which 370 samples were utilized to assess the quality of the suggested classifier. There are a total of 370 samples, 195 of which are cancerous, and 175 of which are benign. Still, 98% precision is achieved with this model.

Figure 4.4 shows that without regularization, benign tumors are more likely to be detected than malignant ones; in this case, the accuracy of correct detection is 99% for the benign breast tumor and 93% for the malignant breast tumor. However, Figure 4.5 shows that this difference is reduced after normalization; the accuracy of correct detection is 98% for the benign breast tumor and 96% for the malignant breast tumor. With normalization, CNN-based SVM classification achieves an average accuracy of 97%, whereas, without normalization, it achieves an accuracy of 96%. As can be shown in Table 4.1, our CNN-based SVM (Benign and Malignant) performs well in testing for binary-class BC classification on BreakHis.

Table 4.1. Performance evaluation of BC tumor classification (CNN-Based SVM) on BreakHis database

BC Tumor	Precision	Recall	F1-Score	Support
Benign 0	**0.96**	**0.98**	**0.97**	**175**
Malignant 1	**0.98**	**0.96**	**0.97**	**195**
accuracy		**0.97**		**370**

Table 4.1 shows that out of 370 evaluated photos on BreakHis, the average accuracy of the CNN-Based Classification technique is 97%, with the Precision, Recall, and F1-Score for Benign BC tumors and Malignant BC tumors being 0.96, 0.98, and 0.97, respectively. The area under the ROC curve (AUC) offers a method for assessing models that are utilized in the medical diagnostic system. The AUC value for a classifier's effectiveness should be around 0 and 1, with a higher AUC value indicating greater performance.

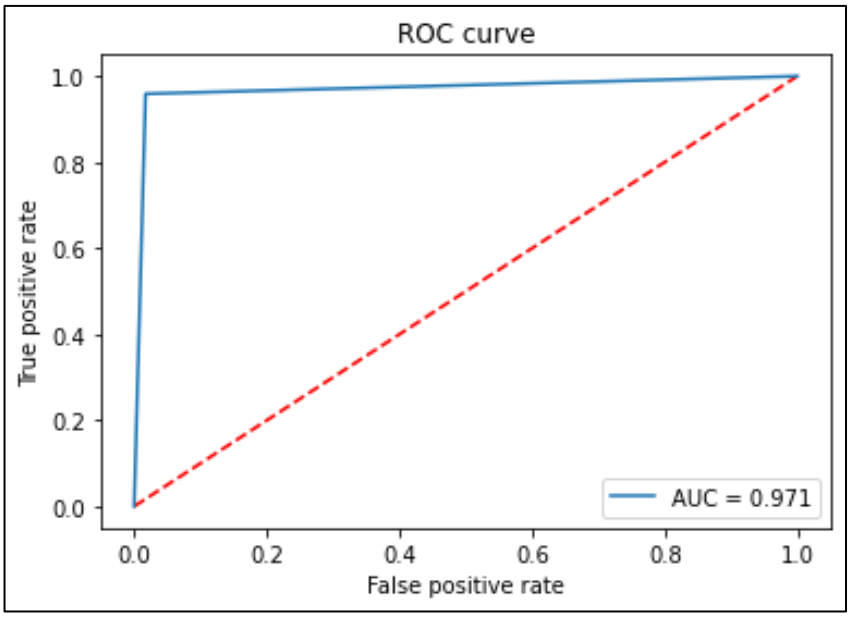

Figure 4.6. The curve with AUC for different magnification factors for 2 class BC classifications.

It has been discovered that the proposed method offers a very high accuracy when compared to other machine learning methods such as DCNN, Based CNN Model, SVM + HOG, SVM + LBP, and CNN Based Rleu on BreakHis. These machine learning methods have all been implemented on the same database and quantity of tested images and evaluated by Precision, Recall, and F1-Score. The results obtained are presented in Table 4.2.

Table 4.2. Evaluation of the performance of implementing machine learning methods on BreakHis

Approaches	**Precision**	**Recall**	**F1-Score**	**Tested Images**
Our CNN-Based SVM	0.97	0.98	0.97	370
DCNN	0.83	0.75	0.74	370
Based CNN Model	0.91	0.77	0.94	370
SVM + HOG	0.78	0.57	0.60	370
SVM + LBP	0.80	0.63	0.65	370
CNN Based Rleu	0.82	0.79	0.79	370

Figure 4.7 shows the predicted results of CNN-Based SVM classification on private images as a Malignant or Benign breast tumor. The top of each sample presents the predicted result and actual results, which shows all are correctly predicted.

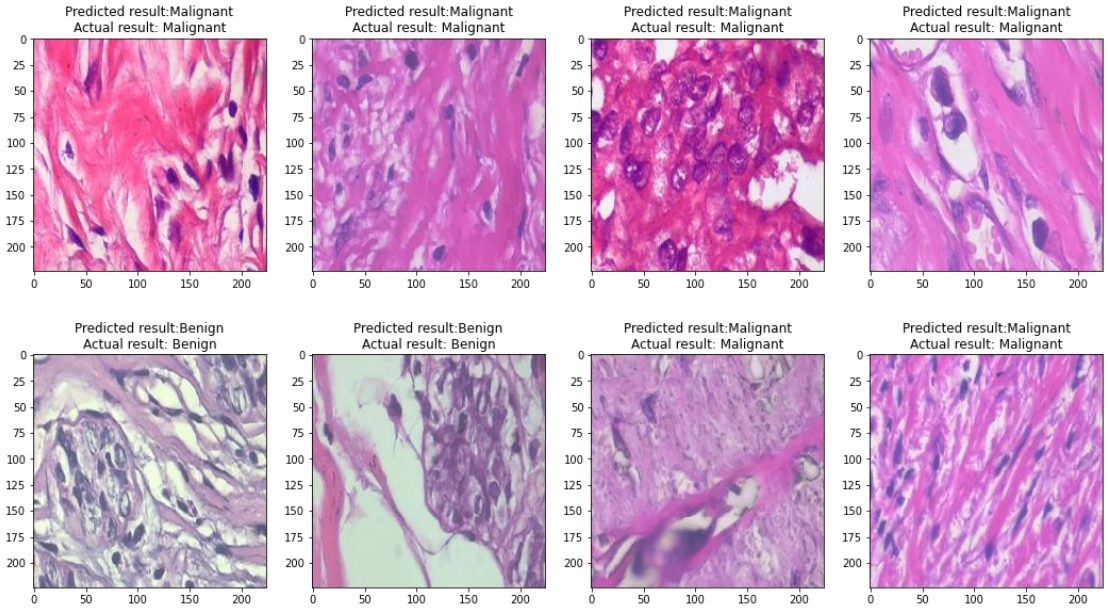

Figure 4.7. The predicted results of CNN-Based SVM classification on private images

A critical step that has an impact on the performance of the CNN when used in conjunction with SVM is the creation of the appropriate convolutional neural networks. However, there are no hard and fast rules about the precise number of layers that have to be used in order to achieve the maximum possible level of performance. Experimentation, prior knowledge, and an understanding of the kind of issue at hand are often used as foundations for it. In this study, we used a topology that consisted of five convolutional layers, each of which was decided upon via the process of cross-validation. Both the binary and the multi-class classification problems are solved by using the same topology. This topology gave the model the flexibility to capture high-level characteristics and non-linearities in the data, while also bringing about promising findings. Moreover, it was able to do so. In order to catch even the most minute of details in the photographs, we have applied minuscule filters with a ratio of three to three to each layer. Now that we've gotten this far, we're going to look at current research that employs CNN and handmade features-based techniques to classify data using the SVM classifier and compare them with our own topology and the results we've gotten so far. The findings of the comparison are shown in Table 4.3.

Table 4.3. Comparative Breast cancer classification results on BreaKHis database

Reference	Reference	Accuracy	Precision	Recall	F1-Score
Our Method	**CNN-Based SVM**	**98.23**	**97.53**	**98.21**	**97.37**
2023 [68]	LBP+SVM	96.91	96.75	94.72	97.10
2023 [96]	FebNet Model	97.10	---	---	---
2022 [115]	deep CNN models	95.85	---	---	94.35
2022 [116]	CNN Models	96.38	---	---	95.00
2022 [117]	DCNN with fficientNetV2-B0	97.82	96.36	97.32	96.08
2022 [118]	6B-Net deep CNN	94.20	---	---	---

The structure that we suggest in order to distinguish between cancerous and noncancerous breast cancers, researchers have developed a CNN-Based SVM. When classifying tumors as benign or malignant, the average accuracy (based on patient score) was 98.23. Breast histology images were classified as benign or cancerous using a CNN fed raw images, custom features, and frequency-domain information in [115]. Results ranged from 95.02% to 95.85%, according to reports. In [116], the authors report achieving 95.5–96.38% accuracy in binary classification and 95.8–96.9% accuracy in multiclass classification using a structured deep learning model. We found that our proposed and retained model is better than state-of-the-art in accuracy in the classification problem, with percentages ranging from 96.15 to 98.33 percent, better than the techniques described in [115-118]. Comparing the CNN topologies used in the research in [117] with our own, we find that our CNN topology (without utilizing an ensemble model) generated the greatest accuracies and F1-scores for pictures. Referring to the research in [118]. The binary-class classification results of [118] were superior (between 92.8% and 93.9%).

Multiple fusion methods, including summing, multiplying, and maxing, are compared and contrasted in [116] to determine the performance level. Consequently, we compared our results to the best accuracy found in [116]. In every scenario, our suggested technique performs better.

For the most part, past research has given classification findings for benign and malignant cases [66] when analyzing the BreakHis dataset's performance. However, some research has shown success in solving the multi-class challenge of classifying

breast cancer [89]. Both binary and multi-class tasks with magnification factors of 40x, 100x, 200x, and 400x have been tested in these tests. Several feature-based methods, like PFTAS, GLCM, QDA, SVM, 1-NN, and RP, were applied to the BreakHis dataset, with reports of accuracy of 85–90% at the patient level [8]. In addition, AlexNet was applied to the problem of binary breast cancer detection over a range of magnification factors, with the best results reaching 95.64.8% for image-level analysis and 90.06.7% for patient-level analysis, respectively [90]. For benign and malignant BC, the highest reported accuracies were 96.91.9% at the picture level and 97.12.8% at the patient level, respectively [114]. The highest test accuracies for multi-class breast cancer classification were 93.91.9% for image-level analysis and 94.73.0% for patient-level analysis [115].

Conversely, the suggested CNN-Based SVM model attained testing accuracies of 98.351.07% and 97.651.20% for benign and malignant BC classification, respectively, in this work's image and patient-level analyses. Accordingly, we have enhanced average performance by 1.08% and 0.65% relative to the best accuracy recorded for picture and patient-level analysis, respectively.

PART 5

CONCLUSION

The proposed CNN-Based SVM model for breast tumor classification combines the power of convolutional neural networks (CNNs) and support vector machines (SVMs) to improve the accuracy and robustness of tumor classification tasks. Feature Extraction with CNNs is CNNs are excellent at automatically extracting hierarchical features from raw data. In the context of breast tumor classification, they can learn relevant patterns and representations from medical images, such as mammograms or histopathological images, without the need for hand-engineered features. This ability allows CNNs to capture complex spatial patterns and contextual information in the images. Transfer Learning is CNNs can be pre-trained on large datasets for general image recognition tasks (e.g., ImageNet dataset) and then fine-tuned on the breast tumor dataset. Transfer learning helps in leveraging the knowledge acquired from the large dataset to improve performance on the target task, even with limited training data, and often leads to better generalization. And last is SVM as Classifier After feature extraction with the CNN, an SVM is used as the classifier. SVMs are a popular choice for classification tasks because they create an optimal hyperplane that maximizes the margin between classes in the feature space. This margin maximization helps in better generalization to unseen data and enhances the model's ability to handle noise and variations in the data.

Machine (SVM) model built on a Convolutional Neural Network (CNN). The convolutional, max-pooling, and fully connected layers were used during the pre-training phase. This pre-training phase was then followed by a classification layer that used SVM to differentiate between benign and cancerous samples. The CNN-Based SVM model was used to conduct experiments on two distinct benchmark datasets: BreakHis and the 2020 Breast Cancer Classification Challenge. The performance of the model was assessed using a variety of performance measures. While putting up

this solution, we considered various factors, including the magnification factor, resized sample inputs, enhanced patches and samples, and patch-based categorization. Compared to all of the findings published in scholarly studies as of 2016, the proposed method exhibits an improvement of around 4.35% and 3.12% in terms of the average recognition accuracy on the BreakHis dataset.

The findings produced for this investigation provide evidence that the classifier's performance is superior to that of other approaches considered state-of-the-art. Despite this, the CNN-Based SVM learning process still requires significant effort to be implemented on outdated hardware. Because of this, developing a CAD system based on CNN and using SVM or other commercial hardware is still a complex undertaking. However, the hardware implementation of neural networks of this kind may aid medical professionals in the early identification of breast cancer. In addition, this technique demonstrates a testing accuracy of 98% for binary breast cancer recognition on the BreakHis dataset. This testing accuracy is significantly higher than any of the other machine learning approaches shown in Table 4.2 for image-based and patch-based recognition performance, respectively.

Therefore, the results of the experiments reveal that the state-of-the-art testing accuracy for breast cancer detection is much higher than that of the techniques currently in use for both datasets that are provided in Table 4.3.

REFERENCES

1. Al Rahhal, M. M., 2018. Breast cancer classification in histopathological images using convolutional neural network. Breast Cancer, 9(3).

2. Bardou, D., Zhang, K., Ahmad, S. M., 2018. Classification of breast cancer based on histology images using convolutional neural networks. Ieee Access, 6, 24680–24693.

3. Benhammou, Y., Achchab, B., Herrera, F., Tabik, S., 2020. BreakHis based breast cancer automatic diagnosis using deep learning: Taxonomy, survey and insights. Neurocomputing, 375, 9–24.

4. Ahmad, N., Asghar, S., & Gillani, S. A. (2022). Transfer learning-assisted multi-resolution breast cancer histopathological images classification. The Visual Computer, 38(8), 2751-2770.

5. Deniz, E., S¸eng¨ur, A., Kadiro˘glu, Z., Guo, Y., Bajaj, V., Budak, U¨ ., 2018. Transfer learning based histopathologic image classification for breast cancer detection. Health information science and systems, 6(1), 1–7.

6. Gour, M., Jain, S., Sunil Kumar, T., 2020. Residual learning based CNN for breast cancer histopathological image classification. International Journal of Imaging Systems and Technology, 30(3), 621–635.

7. Zotin, A., Hamad, Y., Simonov, K., & Kurako, M. (2019). Lung boundary detection for chest X-ray images classification based on GLCM and probabilistic neural networks. Procedia Computer Science, 159, 1439-1448.

8. Hu, Z., Tang, J.,Wang, Z., Zhang, K., Zhang, L., Sun, Q., 2018. Deep learning for image-based cancer detection and diagnosis- A survey. Pattern Recognition, 83, 134–149.

9. Jiang, Y., Chen, L., Zhang, H., Xiao, X., 2019. Breast cancer histopathological image classification using convolutional neural networks with small SE-ResNet module. PloS one, 14(3), e0214587.

10. Zotin, A., Simonov, K., Kurako, M., Hamad, Y., & Kirillova, S. (2018). Edge detection in MRI brain tumor images based on fuzzy C-means clustering. Procedia Computer Science, 126, 1261-1270.

11. Kumar, N., Verma, R., Sharma, S., Bhargava, S., Vahadane, A., Sethi, A., 2017. A dataset and a technique for generalized nuclear segmentation for computational pathology. IEEE transactions on medical imaging, 36(7), 1550–

1560.

12. Li, Y., Wu, J., Wu, Q., 2019. Classification of breast cancer histology images using multi-size and discriminative patches based on deep learning. IEEE Access, 7, 21400–21408.

13. Mehra, R. et al., 2018. Breast cancer histology images classification: Training from scratch or transfer learning? ICT Express, 4(4), 247–254.

14. Nahid, A.-A., Mehrabi, M. A., Kong, Y., 2018. Histopathological breast cancer image classification by deep neural network techniques guided by local clustering. BioMed research international, 2018.

15. Hamad, Y. A., Kadum, J., Rashid, A. A., Mohsen, A. H., & Safonova, A. (2022, October). A deep learning model for segmentation of covid-19 infections using CT scans. In AIP Conference Proceedings (Vol. 2398, No. 1, p. 050005). AIP Publishing LLC.

16. Talo, M., 2019. Automated classification of histopathology images using transfer learning. Artificial intelligence in medicine, 101, 101743.

17. Viale, P. H., 2020. The American Cancer Society's facts & figures: 2020 edition. Journal of the Advanced Practitioner in Oncology, 11(2), 135.

18. Vo, D. M., Nguyen, N.-Q., Lee, S.-W., 2019. Classification of breast cancer histology images using incremental boosting convolution networks. Information Sciences, 482, 123–138.

19. Zotin, A., Kents, A., Simonov, K., & Hamad, Y. (2021, July). Methods of Interpretation of CT Images with COVID-19 for the Formation of Feature Atlas and Assessment of Pathological Changes in the Lungs. In Intelligent Decision Technologies: Proceedings of the 13th KES-IDT 2021 Conference (pp. 173-183). Singapore: Springer Singapore.

20. Safonova, A., Hamad, Y., Dmitriev, E., Georgiev, G., Trenkin, V., Georgieva, M., ... & Iliev, M. (2021). Individual tree crown delineation for the species classification and assessment of vital status of forest stands from UAV images. Drones, 5(3), 77.

21. Safonova, A., Hamad, Y., Alekhina, A., & Kaplun, D. (2022). Detection of Norway Spruce Trees (Picea Abies) Infested by Bark Beetle in UAV Images Using YOLOs Architectures. IEEE Access, 10, 10384-10392.

22. Fan H., Pei J., Zhao Y. An optimized probabilistic neural network with unit hyperspherical crown mapping and adaptive kernel coverage //Neurocomputing. – 2020. – T. 373. – C. 24-34.

23. Hamad, Y. A., Seno, M. E., Al-Kubaisi, M., & Safonova, A. N. (2021). Segmentation and measurement of lung pathological changes for COVID-19

diagnosis based on computed tomography. Periodicals of Engineering and Natural Sciences, 9(3), 29-41.

24. Ahmad W. S. H. M. W. et al. Classification of infection and fluid regions in chest x-ray images //2016 International Conference on Digital Image Computing: Techniques and Applications (DICTA). – IEEE, 2016. – С. 1-5.

25. Kents, A. S., Hamad, Y. A., Simonov, K. V., & Zotin, A. G. (2021). METHODS AND MODELS FOR TEXTURE ANALYSIS OF LUNG PATHOLOGICAL CHANGES BASED ON COMPUTED TOMOGRAPHY FOR COVID-19 DIAGNOSIS. International Archives of the Photogrammetry, Remote Sensing & Spatial Information Sciences.

26. Badawy S. M. et al. Breast cancer detection with mammogram segmentation: a qualitative study //International Journal of Advanced Computer Science and Application. – 2017. – Т. 8. – №. 10.

27. Cadena L. et al. Brain's tumor image processing using shearlet transform //Applications of Digital Image Processing XL. – International Society for Optics and Photonics, 2017. – Т. 10396. – С. 103961B.

28. Zotin, A., Hamad, Y., Simonov, K., Kurako, M., & Kents, A. (2020, June). Processing of CT lung images as a part of radiomics. In Intelligent Decision Technologies: Proceedings of the 12th KES International Conference on Intelligent Decision Technologies (KES-IDT 2020) (pp. 243-252). Singapore: Springer Singapore.

29. Dalmış M. U. et al. Using deep learning to segment breast and fibroglandular tissue in MRI volumes //Medical physics. – 2017. – Т. 44. – №. 2. – С. 533-546.
30. Hamad, Y. A., Simonov, K., & Naeem, M. B. (2018, November). Breast cancer detection and classification using artificial neural networks. In 2018 1st Annual International Conference on Information and Sciences (AiCIS) (pp. 51-57). IEEE.

31. Li X., Chen L., Chen J. A visual saliency-based method for automatic lung regions extraction in chest radiographs //2017 14th International Computer Conference on Wavelet Active Media Technology and Information Processing (ICCWAMTIP). – IEEE, 2017. – С. 162-165.

32. Ramaha, N. T. (2023, February). Review Of Breast Diagnosis Detection and Classification Based on Machine Learning. In International Conference on Trends in Advanced Research (Vol. 1, pp. 222-230).

33. Candemir S., Antani S. A review on lung boundary detection in chest X-rays //International journal of computer assisted radiology and surgery. – 2019. – Т. 14. – №. 4. – С. 563-576.

34. Hamad, Y. A., Simonov, K., & Naeem, M. B. (2018, November). Brain's tumor edge detection on low contrast medical images. In 2018 1st Annual International

Conference on Information and Sciences (AiCIS) (pp. 45-50). IEEE.

35. Hamad, Y. A., Simonov, K. V., & Naeem, M. B. (2019). Detection of brain tumor in MRI images, using a combination of fuzzy C-means and thresholding. International Journal of Advanced Pervasive and Ubiquitous Computing (IJAPUC), 11(1), 45-60.

36. Ragab, M., Albukhari, A., Alyami, J., & Mansour, R. F. (2022). Ensemble deep-learning-enabled clinical decision support system for breast cancer diagnosis and classification on ultrasound images. Biology, 11(3), 439.

37. Dafni Rose, J., VijayaKumar, K., Singh, L., & Sharma, S. K. (2022). Computer-aided diagnosis for breast cancer detection and classification using optimal region growing segmentation with MobileNet model. Concurrent Engineering, 30(2), 181-189.

38. Hamad, Y. A., Simonov, K., & Naeem, M. B. (2020). Lung boundary detection and classification in chest X-rays images based on neural network. In Applied Computing to Support Industry: Innovation and Technology: First International Conference, ACRIT 2019, Ramadi, Iraq, September 15–16, 2019, Revised Selected Papers 1 (pp. 3-16). Springer International Publishing.

39. Hasan S. M. A., Ko K. Depth edge detection by image-based smoothing and morphological operations //Journal of Computational Design and Engineering. – 2016. – Т. 3. – №. 3. – С. 191-197.

40. Webber, J., Mehbodniya, A., Teng, R., & Arafa, A. (2022). Human–Machine interaction using probabilistic neural network for light communication systems. Electronics, 11(6), 932.

41. Khairandish, M. O., Sharma, M., Jain, V., Chatterjee, J. M., & Jhanjhi, N. Z. (2022). A hybrid CNN-SVM threshold segmentation approach for tumor detection and classification of MRI brain images. Irbm, 43(4), 290-299.

42. Terlapu, P. V., Gedela, S. B., Gangu, V. K., & Pemula, R. (2022). Intelligent diagnosis system of hepatitis C virus: A probabilistic neural network based approach. International Journal of Imaging Systems and Technology, 32(6), 2107-2136.

43. Hamad, Y., Mohammed, O. K. J., & Simonov, K. (2019, March). Evaluating of tissue germination and growth rate of ROI on implants of electron scanning microscopy images. In Proceedings of the 9th International Conference on Information Systems and Technologies (pp. 1-7).

44. Wang G. et al. Automatic brain tumor segmentation using cascaded anisotropic convolutional neural networks //International MICCAI brainlesion workshop. – Springer, Cham, 2017. – С. 178-190.

45. Zeinali Y., Story B. A. Competitive probabilistic neural network //Integrated

Computer-Aided Engineering. – 2017. – Т. 24. – №. 2. – С. 105-118.

46. Methods of Interpretation of Data from Isokinetic Tests and MRI Studies during Rehabilitation of Patients after Reconstructive Shoulder Joint Surgery

47. Hamad Y. A., Simonov K. V., Naeem M. B. Detection of Brain Tumor in MRI Images, Using a Combination of Fuzzy C-Means and Thresholding //International Journal of Advanced Pervasive and Ubiquitous Computing (IJAPUC). – 2019. – Т. 11. – №. 1. – С. 45-60.

48. Kabaev, E., Hamad, Y., Simonov, K., & Zotin, A. (2020). Visualization and Analysis of the Shoulder Joint Biomechanics in Postoperative Rehabilitation. In SibDATA (pp. 34-41).

49. Luo, Y., Huang, Q., & Li, X. (2022). Segmentation information with attention integration for classification of breast tumor in ultrasound image. Pattern Recognition, 124, 108427.

50. Mohiyuddin, A., Basharat, A., Ghani, U., Peter, V., Abbas, S., Naeem, O. B., & Rizwan, M. (2022). Breast tumor detection and classification in mammogram images using modified YOLOv5 network. Computational and Mathematical Methods in Medicine, 2022, 1-16.

51. Mathews A. B., Jeyakumar M. K. Performance Analysis of Machine Learning Based Classifiers for the Diagnosis of Lung Cancer & Comparison //Indian Journal of Public Health Research & Development. – 2018. – Т. 9. – №. 12. – С. 2672-2678.

52. Abbasniya, M. R., Sheikholeslamzadeh, S. A., Nasiri, H., & Emami, S. (2022). Classification of breast tumors based on histopathology images using deep features and ensemble of gradient boosting methods. Computers and Electrical Engineering, 103, 108382.

53. Canny J. A computational approach to edge detection //IEEE Transactions on pattern analysis and machine intelligence. – 1986. – №. 6. – С. 679-698.

54. Koprivanac M. et al. Degenerative mitral valve disease-contemporary surgical approaches and repair techniques //Annals of cardiothoracic surgery. – 2017. – Т. 6. – №. 1. – С. 38.

55. Sehgal A. et al. Automatic brain tumor segmentation and extraction in MR images //2016 Conference on Advances in Signal Processing (CASP). – IEEE, 2016. – С. 104-107.

56. Hamad, Y. A., Qasim, M. N., Rashid, A. A., & Seno, M. E. (2020, April). Algorithms of Experimental Medical Data Analysis. In 2020 International Conference on Computer Science and Software Engineering (CSASE) (pp. 112-116). IEEE.

57. Stosic Z., Rutesic P. An improved canny edge detection algorithm for detecting brain tumors in MRI images //International Journal of Signal Processing. – 2018. – Т. 3.

58. Wang G. et al. Automatic brain tumor segmentation based on cascaded convolutional neural networks with uncertainty estimation //Frontiers in computational neuroscience. – 2019. – Т. 13. – С. 56.

59. Хамад, Ю. А., Кириллова, С. В., Курако, М. А., & Симонов, К. В. (2018). Вычислительная методика обработки медицинских изображений, используя вейвлет и нейросети. Медицина и высокие технологии, (3), 5-13.

60. Dalmiya S., Dasgupta A., Datta S. K. Application of wavelet-based k-means algorithm in mammogram segmentation //International Journal of Computer Applications. – 2012. – Т. 52. – №. 15.

61. El Adoui M. et al. MRI Breast Tumor Segmentation Using Different Encoder and Decoder CNN Architectures //Computers. – 2019. – Т. 8. – №. 3. – С. 52.

62. Kanojia M. G., Abraham S. Breast cancer detection using RBF neural network //2016 2nd International Conference on Contemporary Computing and Informatics (IC3I). – IEEE, 2016. – С. 363-368.

63. Liu L. et al. Automated breast tumor detection and segmentation with a novel computational framework of whole ultrasound images //Medical & biological engineering & computing. – 2018. – Т. 56. – №. 2. – С. 183-199.

64. Swanson K. R., Rostomily R. C., Alvord Jr E. C. A mathematical modelling tool for predicting survival of individual patients following resection of glioblastoma: a proof of principle //British journal of cancer. – 2008. – Т. 98. – №. 1. – С. 113-119.

65. Types of tumours - Canadian Cancer Society // http://www.cancer.ca/en/region-selector-page/?url=%2fen%2fabout-us%2fpage-not-found%2f, last accessed: 5 /5/ 2018.

66. Хамад, Ю. А., Симонов, К. В., & Кенц, А. С. (2020). Алгоритмы сегментации и распознавания объектов на медицинских изображениях на основе шиарлет-преобразования и нейронных сетей. Информатизация и связь, (2), 35-45.

67. Zeebaree, D. Q., Haron, H., Abdulazeez, A. M., & Zebari, D. A. (2019, April). Machine learning and region growing for breast cancer segmentation. In 2019 International Conference on Advanced Science and Engineering (ICOASE) (pp. 88-93). IEEE.

68. Chen, H., Ma, M., Liu, G., Wang, Y., Jin, Z., & Liu, C. (2023). Breast Tumor Classification in Ultrasound Images by Fusion of Deep Convolutional Neural Network and Shallow LBP Feature. Journal of Digital Imaging, 1-15.

69. Singh, A. K., & Gupta, B. (2015). A novel approach for breast cancer detection and segmentation in a mammogram. Procedia Computer Science, 54, 676-682.

70. Ye, J., Yang, W., Wang, J., Xu, X., Li, L., Xie, C., ... & Lai, X. (2022). Automated segmentation of mass regions in DBT images using a dilated DCNN approach. Computational Intelligence and Neuroscience, 2022.

71. Ahmed, L., Iqbal, M. M., Aldabbas, H., Khalid, S., Saleem, Y., & Saeed, S. (2020). Images data practices for semantic segmentation of breast cancer using deep neural network. Journal of Ambient Intelligence and Humanized Computing, 1-17.

72. Wang, P., Hu, X., Li, Y., Liu, Q., & Zhu, X. (2016). Automatic cell nuclei segmentation and classification of breast cancer histopathology images. Signal Processing, 122, 1-13.

73. Su, H., Liu, F., Xie, Y., Xing, F., Meyyappan, S., & Yang, L. (2015, April). Region segmentation in histopathological breast cancer images using deep convolutional neural network. In 2015 IEEE 12th International Symposium on Biomedical Imaging (ISBI) (pp. 55-58). IEEE.

74. Zebari, D. A., Zeebaree, D. Q., Abdulazeez, A. M., Haron, H., & Hamed, H. N. A. (2020). Improved threshold based and trainable fully automated segmentation for breast cancer boundary and pectoral muscle in mammogram images. Ieee Access, 8, 203097-203116.

75. Punitha, S., Amuthan, A., & Joseph, K. S. (2018). Benign and malignant breast cancer segmentation using optimized region growing technique. Future Computing and Informatics Journal, 3(2), 348-358.

76. Su, Y., Liu, Q., Xie, W., & Hu, P. (2022). YOLO-LOGO: A transformer-based YOLO segmentation model for breast mass detection and segmentation in digital mammograms. Computer Methods and Programs in Biomedicine, 221, 106903.

77. Chauhan A., Mittal N., Khatri S. K. Reduction of Noise of Cloud Medical Images Using Image Enhancement Technique //Advances in Interdisciplinary Engineering. – Springer, Singapore, 2019. – С. 825-835.

78. Hien N. M., Binh N. T., Viet N. Q. Edge detection based on Fuzzy C Means in medical image processing system //2017 International Conference on System Science and Engineering (ICSSE). – IEEE, 2017. – С. 12-15.

79. Lal S. et al. Efficient algorithm for contrast enhancement of natural images //Int. Arab J. Inf. Technol. – 2014. – Т. 11. – №. 1. – С. 95-102.

80. Симонов, К. В., Зотин, А. Г., & Хамад, Ю. А. (2019). Алгоритмы обнаружения и классификация патологии легких на рентгеновских снимках. Медицина и высокие технологии, (2), 46-53.

81. Shen D., Wu G., Suk H. I. Deep learning in medical image analysis //Annual

review of biomedical engineering. – 2017. – Т. 19. – С. 221-248.

82. Singh A. K., Gupta B. A novel approach for breast cancer detection and segmentation in a mammogram //Procedia Computer Science. – 2015. – Т. 54. – С. 676-682.

83. Suzuki K. Overview of deep learning in medical imaging //Radiological physics and technology. – 2017. – Т. 10. – №. 3. – С. 257-273.

84. S. Rizzo, F. Botta, S. Raimondi, D. Origgi, C. Fanciullo, A.G. Morganti, M. Bellomi, Radiomics: the facts and the challenges of image analysis, Eur Radiol Exp. 2 (2018), https://doi.org/10.1186/s41747-018-0068-z.

85. Z. Liu, S. Wang, D. Dong, J. Wei, C. Fang, X. Zhou, K. Sun, L. Li, B. Li, M. Wang, J. Tian, The applications of radiomics in precision diagnosis and treatment of oncology: opportunities and challenges, Theranostics. 9 (2019) 1303–1322, https://doi.org/10.7150/thno.30309

86.] P. Afshar, A. Mohammadi, K.N. Plataniotis, A. Oikonomou, H. Benali, From handcrafted to deep-learning-Based Cancer radiomics: challenges and opportunities, IEEE Signal Process. Mag. 36 (2019) 132–160.

87. Balha, A., & Singh, C. K. (2023). Comparison of Maximum Likelihood, Neural Networks, and Random Forests Algorithms in Classifying Urban Landscape. In Application of Remote Sensing and GIS in Natural Resources and Built Infrastructure Management (pp. 29-38). Cham: Springer International Publishing.

88. S. Yu, S. Guo, Big Data Concepts, Theories, and Applications, Springer International Publishing, (2016), https://doi.org/10.1007/978-3-319-27763-9.

89. W. Sun, B. Zheng, W. Qian, Computer-aided lung cancer diagnosis with deep learning algorithms, in: Medical Imaging 2016: Computer-Aided Diagnosis, International Society for Optics and Photonics (2016) 97850Z, https://doi.org/10. 1117/12.2216307.

90. Кенц, А. С., Симонов, К. В., & Хамад, Ю. А. (2019). Визуализация и контрастирование медицинских изображений. Медицина и высокие технологии, (4), 26-33.

91. Ma, D., Shang, L., Tang, J., Bao, Y., Fu, J., & Yin, J. (2021). Classifying breast cancer tissue by Raman spectroscopy with one-dimensional convolutional neural network. Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy, 256, 119732.

92. Nahid, A.-A., Mehrabi, M. A., Kong, Y., 2018. Histopathological breast cancer image classification by deep neural network techniques guided by local clustering. BioMed research international, 2018.

93. Jiang, Y., Chen, L., Zhang, H., Xiao, X., 2019. Breast cancer histopathological image classification using convolutional neural networks with small SE-ResNet module. PloS one, 14(3), e0214587.

94. Talo, M., 2019. Automated classification of histopathology images using transfer learning. Artificial intelligence in medicine, 101, 101743.

95. Mehra, R. et al., 2018. Breast cancer histology images classification: Training from scratch or transfer learning? ICT Express, 4(4), 247–254.

96. Amin, M. S., & Ahn, H. (2023). FabNet: A Features Agglomeration-Based Convolutional Neural Network for Multiscale Breast Cancer Histopathology Images Classification. Cancers, 15(4), 1013.

97. Ouyang, Y., Yu, C., Yan, G., & Chen, J. (2021). Machine learning approach for the prediction and optimization of thermal transport properties. Frontiers of Physics, 16(4), 1-16.

98. Aljuaid, H., Alturki, N., Alsubaie, N., Cavallaro, L., & Liotta, A. (2022). Computer-aided diagnosis for breast cancer classification using deep neural networks and transfer learning. Computer Methods and Programs in Biomedicine, 223, 106951.

99. Hamad, Y. A., Simonov, K., & Naeem, M. B. (2018, November). Breast cancer detection and classification using artificial neural networks. In 2018 1st Annual International Conference on Information and Sciences (AiCIS) (pp. 51-57). IEEE.

100. Zotin, A., Hamad, Y., Simonov, K., & Kurako, M. (2019). Lung boundary detection for chest X-ray images classification based on GLCM and probabilistic neural networks. Procedia Computer Science, 159, 1439-1448.

101. Hao, Y., Li, Q., Mo, H., Zhang, H., & Li, H. (2018). AMI-Net: convolution neural networks with affine moment invariants. IEEE Signal Processing Letters, 25(7), 1064-1068.

102. Симонов, К. В., Зотин, А. Г., Хамад, Ю. А., Курако, М. А., & Кенц, А. С. (2019). Алгоритмы обнаружения и классификации визуальных данных. Информатизация и связь, (4), 55-63.

103. Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., & Wojna, Z. (2016). Rethinking the inception architecture for computer vision. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 2818-2826).

104. Jaffar, M. A. (2017). Deep learning based computer aided diagnosis system for breast mammograms. International Journal of Advanced Computer Science and Applications, 8(7).

105. Qiu, Y., Wang, Y., Yan, S., Tan, M., Cheng, S., Liu, H., & Zheng, B. (2016, March). An initial investigation on developing a new method to predict short-term breast cancer risk based on deep learning technology. In Medical Imaging

2016: Computer-Aided Diagnosis (Vol. 9785, pp. 517-522). SPIE.

106. Ertosun, M. G., & Rubin, D. L. (2015, November). Probabilistic visual search for masses within mammography images using deep learning. In 2015 IEEE International Conference on Bioinformatics and Biomedicine (BIBM) (pp. 1310-1315). IEEE.

107. Lin, C. H., Lai, H. Y., Chen, P. Y., Wu, J. X., Pai, C. C., Su, C. M., & Ho, H. W. (2022). Breast lesions screening of mammographic images with 2D spatial and 1D convolutional neural network-based classifier. Applied Sciences, 12(15), 7516.

108. Sajiv, G., & Ramkumar, G. (2022, August). Automated Breast Cancer Classification based on Modified Deep learning Convolutional Neural Network following Dual Segmentation. In 2022 3rd International Conference on Electronics and Sustainable Communication Systems (ICESC) (pp. 1562-1569). IEEE.

109. Albalawi, U., Manimurugan, S., & Varatharajan, R. (2022). Classification of breast cancer mammogram images using convolution neural network. Concurrency and Computation: Practice and Experience, 34(13), e5803.

110. Agnes, S. A., Anitha, J., Pandian, S., & Peter, J. D. (2020). Classification of mammogram images using multiscale all convolutional neural network (MA-CNN). Journal of medical systems, 44(1), 1-9.

111. Zheng, Y., Jiang, Z., Xie, F., Zhang, H., Ma, Y., Shi, H., & Zhao, Y. (2017). Feature extraction from histopathological images based on nucleus-guided convolutional neural network for breast lesion classification. Pattern Recognition, 71, 14-25.

112. Öztürk, Ş., & Akdemir, B. (2019). HIC-net: A deep convolutional neural network model for classification of histopathological breast images. Computers & Electrical Engineering, 76, 299-310.

113. Kaushal, C., Bhat, S., Koundal, D., & Singla, A. (2019). Recent trends in computer assisted diagnosis (CAD) system for breast cancer diagnosis using histopathological images. Irbm, 40(4), 211-227.

114. Talo, M. (2019). Automated classification of histopathology images using transfer learning. Artificial intelligence in medicine, 101, 101743.

115. Saxena, S., Shukla, P. K., & Ukalkar, Y. (2023, March). A Shallow Convolutional Neural Network Model for Breast Cancer Histopathology Image Classification. In Proceedings of International Conference on Recent Trends in Computing: ICRTC 2022 (pp. 593-602). Singapore: Springer Nature Singapore.

116. Hao, Y., Zhang, L., Qiao, S., Bai, Y., Cheng, R., Xue, H., ... & Zhang, G. (2022). Breast cancer histopathological images classification based on deep semantic features and gray level co-occurrence matrix. Plos one, 17(5), e0267955.

117. Sheela, R. K., Nagaraju, Y., & Sahu, D. A. (2022, May). Histopathological Image Classification of Breast Cancer using EfficientNet. In 2022 3rd International Conference for Emerging Technology (INCET) (pp. 1-8). IEEE.

118. Umer, M. J., Sharif, M., Kadry, S., & Alharbi, A. (2022). Multi-Class Classification of Breast Cancer Using 6B-Net with Deep Feature Fusion and Selection Method. Journal of Personalized Medicine, 12(5), 683.

RESUME

His name is Zeyad Abdalkareem Khalaf KHALAF. His primary and elementary education in Iraq. He completed his undergraduate studies at Al-Qalam University in the College of Computer Engineering, Communication Network Departement in 2016 Kirkuk-Iraq Then he started his studying for a master's degree in computer engineering in 2021 and completed his studies in 2023 at the University of Karabuk.