

Breast Cancer Detection through Machine Learning Techniques



By

Aqsaa Khattak

2016-NUST-MS-IT-00000170503

Supervisor

Dr. Rafia Mumtaz

Department of Computing

A thesis submitted in partial fulfillment of the requirements for the degree
of Master's in information technology (MS IT)

In

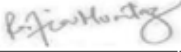
School of Electrical Engineering and Computer Science,
National University of Sciences and Technology (NUST),
Islamabad, Pakistan.

(June 2020)

Approval

It is certified that the contents and form of the thesis entitled "**BREAST CANCER DETECTION THROUGH MACHINE LEARNING TECHNIQUES**" submitted by **AQSAA KHATTAK** have been found satisfactory for the requirement of the degree.

Advisor: Dr. Rafia Mumtaz

Signature:  _____

Date: 13-Jul-2020

Committee Member 1: Dr. Rabia Irfan

Signature:  _____

Date: 13-Jul-2020

Committee Member 2: Ms. Hirra Anwar

Signature:  _____

Date: 12-Jul-2020

Committee Member 3: Asad|Ali Shah

Signature:  _____

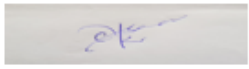
Date: 11-Jul-2020

This dissertation is dedicated to myself; I have worked hard tirelessly for the last 20 years to be where I am today. I appreciate my courage and consistency in the face of struggles. There was no pressure to go faster. I had set my goals at my own pace. Thank you for being me.

Certificate of Originality

I hereby declare that this submission titled "**BREAST CANCER DETECTION THROUGH MACHINE LEARNING TECHNIQUES**" is my own work. To the best of my knowledge it contains no materials previously published or written by another person, nor material which to a substantial extent has been accepted for the award of any degree or diploma at NUST SEecs or at any other educational institute, except where due acknowledgement has been made in the thesis. Any contribution made to the research by others, with whom I have worked at NUST SEecs or elsewhere, is explicitly acknowledged in the thesis. I also declare that the intellectual content of this thesis is the product of my own work, except for the assistance from others in the project's design and conception or in style, presentation and linguistics, which has been acknowledged. I also verified the originality of contents through plagiarism software.

Student Name: AQSAA KHATTAK

Student Signature: 

Acknowledgment

In the name of ALLAH, the most beneficent and the most merciful. First and foremost, all praises and thanks to Allah Almighty for the strength, courage, and His blessing he gave me in completing this dissertation. He who gave me the opportunity, knowledge, ability and helped me all the times I needed Him, without whose will, it was not possible.

Then comes my family, which occupies the most precious place in my life. I would like to thank my parents and siblings for their support and love throughout. Thank you for having faith in me, which kept me going.

Being my Supervisor, I would like to convey my profound and cordial gratitude to my honorable supervisor Dr. Rafia Mumtaz, without whom, I would not have been writing my Master thesis. Ma'am has been a source of inspiration for me. I am thankful to her for pushing me when I was not sure of my potential. I am extremely indebted to my supervisor whose generous suggestions, kind control, guidance, cooperation, encouragement, and advices were greatly useful in completing this crucial task.

I would also like to thank to my junior, Faisal, who stuck to me throughout for bearing me and helping me whenever I was stuck anywhere. This thesis would not have been materialized without him.

I also owe my sincere thanks to my best friends for constantly pushing and telling me that I can do it, because of their affirmations, I did it. I Love you all and you know that.

Table of Contents

Chapter 1: Introduction	1
1.1 Breast Cancer in Pakistan	3
1.2 Motivation.....	4
1.3 Objective and Scope	5
1.4 Thesis Organization	5
Chapter 2: Literature Review.....	6
2.2 Use of Transfer Learning.....	11
2.3 Discussion	11
Chapter 3: Methodology.....	13
3.1 Tools Used.....	14
3.1.1 Anaconda	14
3.1.2 Spyder	15
3.1.3 Jupyter Notebook.....	15
3.2 Data Acquisition and Feature Extraction.....	15
3.2.1 Datasets Description	15
3.2.1.1 Breast Cancer Histopathological Database (BreakHis) Dataset:	16
3.2.1.2 The IDC (Invasive Ductal Carcinoma) Breast Histopathology	
Images Dataset:	17
3.2.1.3 The mini-MIAS dataset of Mammograms:	19
3.2.2 Feature Extraction using Pre-trained Neural Network	21
3.2.2.1 Transfer learning using feature extraction	21

3.2.2.2 Transfer learning using fine-tuning.....	22
3.2.2.3 VGG16 and VGG19.....	22
3.2.3 Data Splitting.....	24
3.2.3.1 Train-Test Split	24
3.3 Machine and Deep Learning Algorithms.....	25
3.3.1 Logistic Regression	25
3.3.2 K-Nearest Neighbors (K-NN)	25
3.3.3 Support Vector Machines (SVM):	25
3.3.4 Naïve Bayes:.....	26
3.3.5 Decision Tree:	26
3.3.6 Random Forest:	26
3.3.7 Artificial Neural Network (ANN):.....	27
3.4 Classification Testing Parameters:	27
3.4.1 Confusion Matrix.....	27
3.4.2 Accuracy:.....	28
3.4.3 Precision:	28
3.4.4 Recall:	28
3.4.5 Specificity.....	28
3.4.6 F1-Score:	29
Chapter 4: Results and Analysis:	30
4.1 Dataset 1: Breast Cancer Histopathological Database (BreakHis).....	30
4.1.1 Winner Techniques: Logistic Regression and Artificial Neural Network	31

4.2 Dataset 2: The IDC (Invasive Ductal Carcinoma) Breast Histopathology Images.....	32
4.2.1 Winner Techniques: Logistic Regression and Artificial Neural Network	33
4.3 Dataset 3: The mini-MIAS dataset of Mammograms	34
4.3.1 Winner Techniques: Logistic Regression and Artificial Neural Network	35
4.4 Discussion	36
Chapter 5: Conclusion and Future Work.....	37
5.1 Limitation/Future Work:	38
References	39

List of Figures

Figure 1: Difference between a normal and cancer cell.....	1
Figure 2 : Trends in cancer incidence rates, females (1975-2016).....	2
Figure 3: No of new cases in Pakistani women (Globocan 2018).....	4
Figure 4 : Methodology Flow.....	14
Figure 5: Sample images BreakHis image dataset	17
Figure 6:Sample Images IDC Breast Histology image dataset.....	19
Figure 7: Sample images Mini Mias Mammographic Image dataset.....	21
Figure 8: General VGG Architecture	23
Figure 9: Deep feature Extraction	24
Figure 10: Graphical Representation of extracted features along with decision boundaries for Logistic Regression and Artificial Neural Network – Dataset 1.....	31
Figure 11: Graphical Representation of extracted features along with decision boundaries for Logistic Regression and Artificial Neural Network - Dataset 2.....	33
Figure 12: Graphical Representation of extracted features along with decision boundaries for Logistic Regression and Artificial Neural Network – Dataset 3.....	35

List of Tables

Table 1: Results and Comparison Dataset 1	30
Table 2: Results and Comparison Dataset 2	32
Table 3: Results and Comparison Dataset 3	34

Abstract

The lifetime probability of developing Breast cancer for a female is 1 in 8. Given this probability, breast cancer becomes one of the most rapidly spreading and the most common type of cancer diagnosed in women. Although the cure is available in most countries and survival rate is increasing due to the advances in medical research, but still there is a huge gap in identifying this disease at initial stages. Pakistan, being a third world nation, alone, has the highest rate of increasing breast cancer in Asia with 90000 new cases every year out of which 40000 die. Various reasons contribute to this, lack of awareness among people, cultural setbacks, lack of research in medical, old treatments etc. Numerous cases go undiagnosed or reach last stage where treatments are ineffective, money is also wasted, and precious lives are lost.

Traditional methods of diagnosis include breast self-examination, clinical examination, Biopsy, Mammography, CAD (Computer Aided Diagnosis), MRI and breast ultrasound. It is very difficult to identify a disease based on visual diagnosis of tissue including multiple features such as this one. In recent past, Machine learning techniques have been proven helpful to radiologists and pathologists for fast and efficient detection of breast cancer.

Present investigation explores several machine learning and deep learning techniques to detect and classify breast cancer using transfer learning via VGG-19 which has not been previously done. Six machine learning techniques, Logistic regression, K nearest Neighbor (KNN), Support Vector Machine (SVM), Naïve Bayes , Decision tree and Random Forest whereas one deep learning technique , Artificial Neural Network (ANN) were applied. Three publicly available breast

cancer image datasets were used in this research and the results were analyzed using various parameters i.e. Accuracy, Precision, Recall, F1 score and Specificity. It was observed from the results that Artificial neural network outperforms all techniques, detecting breast cancer with an average accuracy of 91% followed by Logistic Regression giving an average accuracy up to 90% for all three datasets. In conclusion, the results show the potential of accurate classification of breast cancer images as malignant or benign and proves to be useful in effective treatment of the disease as compared to traditional methods.

Keywords: Machine learning, Feature extraction, Breast Cancer, KNN, SVM, Naïve Bayes, Random forest, Decision Tree, Logistic Regression, Artificial Neural Network

Chapter 1: Introduction

According to the latest estimates of GLOBOCAN, the cancer cases have risen to 18.1 million and death toll has reached 9.6 million in 2018. Around the world, total number of people who are alive within 5 years of cancer diagnosis is assessed to be 43.8 million [19]. Approximately, 29.5 million new cases and 16.5 million deaths are likely to occur by the year 2040 [20].

Cancer is a deadly disease in which the new body cells are not replaced by old ones when they die. The cells inside a normal human body multiply at a steady speed. Our body needs to replace old, damaged cells from time to time. The new cells grow and divide. Cancer is developed when this whole natural process breaks down and starts behaving abnormally. The new cells formed start to multiply abnormally and rapidly, and form clusters which are called lumps. Although the immunity system finds and destroys the harmful cells, but cancer cells are able to evade themselves from the immune system, manage to remain inside the body and grow [38]

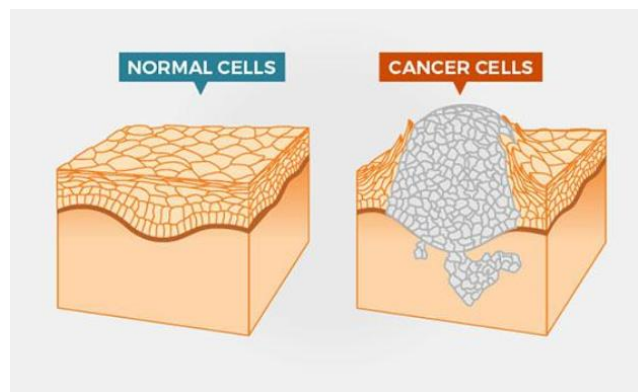


Figure 1: Difference between a normal and cancer cell [38]

Cancer can form in any part of the body but the most common in females after lung cancer is breast cancer. Breast cancer is caused when the tumors are formed inside the breast issues. There are several classifications of breast cancer but to cure it, the most basic one is the binary classification of malignant and benign. Not all tumors are

cancerous. The benign tumors may not be fatal as these do not spread outside of breast tissues, but they certainly increase the probability of being diagnosed with breast cancer during a woman’s lifetime. Second class is malignant which is harmful, and the tumor does not remain confined to the breast tissues only, it can spread to other organs as well.

The incidence of breast cancer is in the first place of malignancy in the world [18]. It is the second major reason of death among women after lung cancer and has found to be more common in women than men. Ratio of occurrence is 99:1. Usually the age in which breast cancer is diagnosed in women is above 50 but it can also occur in younger age depending upon the family history and other factors.

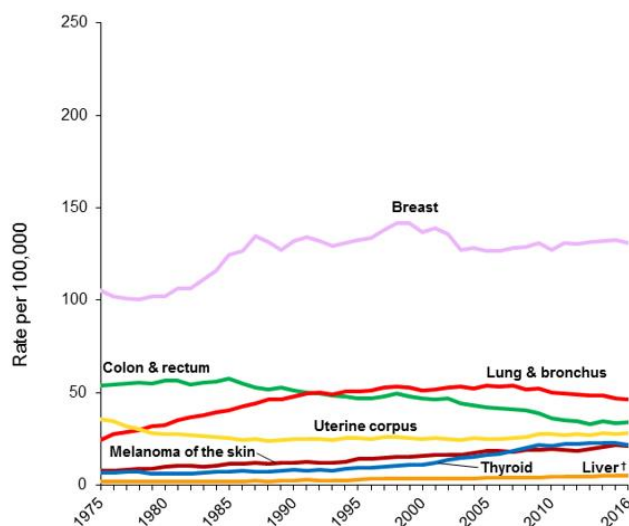


Figure 2 : Trends in cancer incidence rates, females (1975-2016) [20]

There are two major type of risk factors involved in occurrence of this disease. Genetic factors and Environmental/lifestyle factors. Genetic factors include the family history, personal health, menstrual and reproductive history of the patient. The women who had someone with breast cancer in their families are more prone to it. Women having started their menstruation cycle at a very early age and menopause at a very old age have more risks of cancer. Also, if a woman has given birth to child in later period of his life (after 30 years) or does not have children at all throughout her life, she has more chances of being diagnosed with breast cancer. Environmental factors include diet, drug consumption, physical fitness, exposure to pollution etc. The women who don’t exercise and have alleviated lifestyle have been found at a greater risk of breast cancer [39].

There are 15 types of breast cancer out of which six are most common[40].These are Ductal Carcinoma In situ (DCIS) , Lobular Carcinoma In situ (LCIS) , Invasive Ductal Cancer , Invasive lobular breast cancer , Inflammatory breast cancer and Paget's disease. The American joint Committee on Cancer has defined a universal 5 stage guideline (0 , I , II , III , IV) for breast cancer to determine the stage of disease [21].

In 2015, all the united nation member states defined goals for the betterment of humanity and protection of the planet all over the world for the year 2030. These goals are integrated and interdependent. These 17 goals revolve around social, economic, and environmental factors. "Good Health and well-being" is the third goal which ensures that all the people of the world have access to basic health care facilities and can be treated properly creating a healthy environment for people to live. More research in breast cancer in this regard is also important and needed.

1.1 Breast Cancer in Pakistan

Until few years back, it was very difficult to talk about breast cancer in Pakistan because it was associated with female sexuality. According to a research [22], term breast cancer was not properly understood by women or only partially understood. Even some women who knew what the disease was about, they had a very little knowledge and that also because someone from their family or friends was diagnosed with it. Lack of awareness is the major reason of undiagnosed cases and late stage presentation. In Karachi, the incidence of breast cancer is 69.1 per 100000, cases being stage III and IV more than 50% [23]. There is a lack of cancer registries and data collection in Pakistan. only a few major cities like Karachi and Islamabad have registries but they are in their early stages and need special attention from government to study the potential risk factors , predict and help to reduce prevalence of this disease by establishing more cancer centers with proper treatments[24].Over the years as medical research is advancing , steps are being taken to cater this fatal disease, but it will still take a long time to come into proper implementation before it can affect the mortality rate of women at large. Currently, Pakistan ranks on 58th in the world breast cancer ranking [41].

According to a research presented by Shaukat Khanam Cancer Hospital, the current treatments for breast cancer in Pakistan are surgery, radiotherapy, chemotherapy, hormone therapy, targeted therapy [42], but most of the women are presented at later stages of cancer where these treatments are not very effective.

Primary treatment of breast cancer in Pakistan is surgery. It involved removal of the tumor and affected tissue. Sometimes the breasts also have to be removed if the tumor has spread too much in surrounding body. Other procedures are radiotherapy in which tumor is exposed to radiations to kill the cancer cells. This procedure does not have any side effect. Chemotherapy is unfortunately the most required treatment with patients presenting at a later stage. It also includes exposure of cancer cells to high radiations and it has extreme side effects like hair loss, nausea and decreased immunity [26].

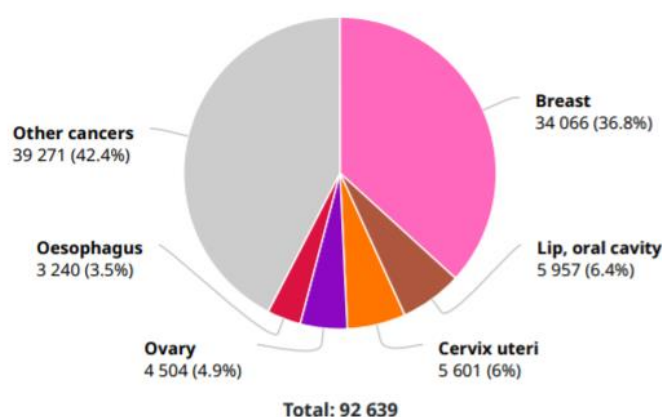


Figure 3: No of new cases in Pakistani women, Globocan 2018 [20]

1.2 Motivation

Pakistan is a developing third world country which is facing a lot of plights especially in health sector. Political instability and economic destruction has generally reduced the overall quality of life. More than half of the population in Pakistan is female. Because most of the population lives in rural areas of Pakistan, medical and healthcare facilities are almost nonexistent. Women are either not diagnosed timely with breast cancer or diagnosed at very last stages which makes it incurable. Lack of education and awareness is also a big cause of this prevalent disease which has been rising in Asia and specially Pakistan since past few years. Although the Government has started fragmented efforts for early breast cancer detection and screening by setting up free diagnostic camps (profit and nonprofit) which are funded through zakats or Baitul maal but these would not be enough to bear the burden in upcoming years if serious measurements are not taken [25].

One in every 9 women are likely to be diagnosed with breast cancer in their lifetime [16]. According to WHO, breast cancer is the 9th out of 50 top death causing disease in Pakistan [17]. Current practices for breast cancer detection in Pakistan are mammography, biopsy and MRI. these techniques can sometimes be not very accurate because of human error and habituation. There is currently no significant research in Pakistan for proactive diagnosis of breast cancer. Machine learning has been very useful in past years in this regard.it is used to solve prognosis and diagnostic problems intelligently and can help doctors to accurately identify the disease.

This thesis is a motivation to facilitate the current medical research on breast cancer in Pakistan through computerized technique so that it can be detected at an early stage efficiently and women can be treated accordingly for increased survival rate. The research is carried out in national interest and will help radiologists/doctors in Pakistan in saving lives of people

1.3 Objective and Scope

The main aim of this research is breast cancer identification through different advanced machine learning and deep learning techniques. Following objectives have been accomplished:

1. To collect breast cancer images datasets containing mammograms and histopathological images from open source online database.
2. To apply various machine learning techniques for classification and detection of breast cancer as malignant and benign.
3. To evaluate results, give comparison and suggest best technique for all the datasets.

1.4 Thesis Organization

This thesis is divided into 5 chapters. Chapter 1 gives the overall introduction, motivation, also describes the objectives and novelties of the research. Chapter 2 sheds light on the literature review. Chapter 3 is Implementation and describes all the steps that were involved in this research. Chapter 4 is Results and Analysis, which explains

and gives comparisons of all the results. Chapter 5 discusses the Conclusion and Future Work for this research.

Chapter 2: Literature Review

Vital contributions in literature has been made over breast cancer diagnostic and treatments. A variety of research has been carried in the past two decades. Researchers have suggested different ways to detect breast cancer at its early stages. In this chapter we have included many national and international papers which we will discuss in a brief and effective manner. We have categorized our surveyed papers into three categories: Research concerning Breast Cancer Detection using deep learning techniques and research concerning breast cancer detection using transfer learning. The brief overview of these chapters is given below.

2.1 Use of Machine Learning Techniques

In [1] Convolutional Neural Network (CNN) was proposed to classify eosin and hematoxylin stained images. Majorly, images used belong to four classes, i.e. benign tissue, normal lesion, invasive carcinoma and situ carcinoma. The proposed architecture was flexible enough and was able to extract information at different levels from network. The system was also employed on histology images. Another popular machine learning algorithm known as Support Vector Machines (SVM) was trained and tested, to classify the images on the features extracted from CNN. 77.8% and 83.3% accuracy was achieved on the above mentioned classes respectively. Ability to detect cancer cases was 95.6%.

In [2], a structured deep learning model was proposed for the prognosis/diagnosis of subclasses (Lobular carcinoma, Ductal carcinoma, Fibroadenoma, etc.) of breast cancer. It resolved the difficulties faced in multi-classification methods used to detect breast

cancer. On a large-scale dataset, structured deep learning model achieved a remarkable performance (92.3% accuracy). In clinical settings, the demonstration of strength of the method provided an effective tool for breast cancer multiclassification.

In [3], a computational approach was developed which uses deep convolutional neural network. Hematoxylin and eosin stained breast histology microscopy image dataset was used. The technique used various deep neural networks. Gradient boosted trees classifiers were also used. An accuracy of 87.2% was reported for 4-class classification task. For 2-classes an accuracy of 93.8% was achieved. The approach outperformed many other common methods to segregate harmful tissues found in breast.

In [4], a new computer aided detection system was proposed to automatically classify malignant and benign tumors using breast mammography images. Two segmentation approaches were used in the system. In the first approach region of interest (ROI) was determined manually while threshold and region based technique was used in the second approach. Feature extraction was done using deep convolutional neural network. A popular DCNN named as AlexNet was employed and was optimized to segregate two classes. Support vector machines classifier was connected to fully connected layer for better accuracy. The datasets relevant to biomedical imaging relatively contains few number of images for training and testing due to small number of patients. As we know that, high accuracy comes with large number of samples, so to enhance size of dataset the technique of data augmentation was used. The form of data augmentation used was rotation. 71.01% accuracy was achieved when ROI was manually cropped from image. The highest AUC was 0.88 (88%). The accuracy was increased to 73.6% when samples obtained from CBIS-DDSM was used. Support vector machines (SVM) enhanced the accuracy of the system to 87.2% and AUC value achieved was 0.94 (94%). As compared to previous work done under similar conditions, the AUC was the highest ever achieved.

In [5], researchers explored a machine-aided system which was based on fusion of features with deep features of Convolutional Neural Network (CNN). A mass detection method was proposed which was based on deep features of a CNN and Unsupervised Extreme Learning Machine (US-ELM) clustering. Morphological features, texture features and density features were fused to build a feature set. Benign and malignant cells were classified by using the new feature set and by applying a classifier. The accuracy and efficiency was demonstrated by extensive experiments.

In [6], an approach which extracts the most vital visual features, by using convolutional layers of various models of deep learning was proposed. To achieve the main goal a novel boosting strategy was also proposed. The system was able to classify the whole set of images into 2 separate groups (carcinomas and non-carcinomas) 4 distinctive classes (benign, normal, in situ and invasive carcinomas). Many state-of-the-art methods (classifiers) were significantly outperformed by the proposed boosting deep learning model.

In [7], an expert diagnostic system was developed and tested, which could distinguish between patients with or without breast cancer was created. The system was able to classify the samples based on characteristics of cell nuclei that are found in digitized image of fine needle aspirate (FNA). UCI machine learning repository was used to collect dataset samples. The total number of samples used from the repository were 699. 99% and 98.9% accuracy was achieved in test and training sets, respectively. The major reason for this high accuracy was Feed Forward Backpropagation single hidden layer neural network with 20 neurons was used along with TANSIG transfer function. In case of multilayered architectures, accuracy was significantly reduced to a range between 74.9% -86.3%, whereas the average accuracy was around 81.37%. A promising method was proposed which could be deployed in any laboratory for classification of harmful tissues.

In [8], a CNN was used for abnormality detection in tissues. Mammograms-MIAS dataset utilized for the testing and training. There were 322 mammograms in the dataset in total, out of which 189 were normal and 133 were abnormal. Quite good results were gathered from experiment, which basically proves, efficiency of the deep learning models for abnormality detection in mammogram images. It also encourages the use of deep learning-based systems in medical imaging especially in breast cancer detection.

In [9], a Deep Neural Network (DNN) was introduced for biomedical image analysis. Statistical and Structural information was gathered from images and was used to classify them by using novel deep neural networks (DNN's). A CNN, a LSTM and a fusion of both of them was proposed for abnormality detection in breasts. Support vector machines and Softmax layers were introduced at the stage of decision making to enhance the performance. The best accuracy that was achieved was 91.00% and the best precision recorded on the dataset was 96.00%.

In [10], a machine-aided system for the detection of abnormality in breast was proposed. In first phase the region of interest was segmented. To obtain the unified

time-frequency spectrum a weighted type fractional Fourier transform (WFRFT) was utilized. Principle component analysis (PCA) was utilized to reduce the spectrum to only 18 components. The classifier was generated using the feed-forward neural network (FNN). Finally, Jaya, a novel algorithm specific technique was used for training purpose. The proposed system achieved accuracy of 92.27%, sensitivity of 92.16% and specificity of 92.28%. It was quite efficient in detecting abnormality in breasts and outperformed 5 best systems of its time. Also, Jaya was quite effective in enhancing the performance of FNN.

In [11], a comparison between two machine learning algorithms for automatic classification between various types of breast cancer was presented. Two techniques were majorly used. In the first technique handcrafted features were extracted. These features were then encoded using bag of words model and locality constrained linear coding model while the second technique used, was based on the design pattern that was akin to Convolutional Neural Network. They have also experimented with data augmentation to optimize performance of CNN and extracted features. Handcrafted feature-based classifier was outperformed by the neural network. 96.15% - 98.33% accuracy was achieved for two classes but for multiclass classification accuracy was reduced to 83.31% - 88.23%.

In [12], robust classification techniques of machine learning such as Decision Tree and Support Vector Machines (SVM) kernels are used to differentiate between cancer mammograms and normal subjects. A variety of features were proposed such as scale invariant feature transform (SIFT), texture, elliptic Fourier Descriptors (EFDs) and morphological entropy. Machine learning algorithms were trained and tested on these samples. Jack-knife 10-fold cross-validation method was utilized for hyper parameter tuning. Performance was tested by means of sensitivity, false positive rate (FPR), specificity, negative predictive value (NPV), positive predictive value (PPV) and ROC. Bayesian approach gave the best performance whereas highest area under the curve (AUC) was also achieved using Bayesian approach.

In [13], a trustworthy diagnosing process was developed to differentiate between malignant and benign tumors. As many researches have proved that machine learning algorithms are efficient and are preferable in detection of abnormal cells, so three machine learning techniques (SVM, K-nearest neighbors and Decision Trees) have been used and the performance of these machine learning classifiers was compared in order to check which classifier outperforms others and works better in detection of cancer. Wisconsin Breast Cancer (Diagnostic) dataset was utilized for experimentation and

research. The results of the study revealed that quadratic support vector machines was reliable enough and achieved accuracy of 98.1% with lowest false discovery rates. MATLAB was used to carry out all the experimentation.

In [14], the system classifies the whole set of images into four major types invasive carcinoma, in situ carcinoma, malignant and benign by deploying deep learning techniques. The proposed system uses two consecutive CNN's. Salient features were extracted in image patches by the first "patch-wise" while the second "image-wise" network was used to classify the whole image. The first network extracts local features. Global info was extracted using second network. The proposed technique yields an accuracy of 95% while the previously developed system reported an accuracy of 77%.

In [15], various machine learning techniques in prognosis/diagnosis of breast cancer were reviewed. Applications of Artificial Neural Networks (ANNs), Support Vector Machines (SVMs), Decision Trees (DTs), and k-nearest neighbors (KNNs) were investigated. Wisconsin breast cancer database (WBCD) was used and the results were compared. Finally, a healthcare system model of the proposed system was also shown.

In [43], Wisconsin Breast Cancer dataset has been used for detection of breast cancer. The paper focuses on various models that are implemented i.e. Logistic Regression, K Nearest Neighbour (KNN), Support Vector Machine (SVM), Multi-Layer perceptron classifier, Artificial Neural Network (ANN). Two evaluation parameters were used i.e. Accuracy and Precision. The accuracies obtained through SVM and Random Forest Classifier was 96.5%. To increase the accuracy deep learning models such as CNN and ANN were implemented, and accuracy was enhanced to 97.3%. And 99.3% respectively. Activation functions such as Relu and sigmoid were used to predict outcomes in terms of probability.

In [44], a system using sparse autoencoder (SAE) was proposed. This whole system was error free. A classifier and SAE were cascaded, and classification was performed on the learned features. Different machine learning algorithms were used like KNN, SVM, decision trees and Random Forest. The achieved performance of each algorithm was compared with other ones. From results, it was observed that Random Forest outperformed other classifiers.

2.2 Use of Transfer Learning

In [45], detection of breast cancer using transfer learning and deep neural network was proposed. The technique segregated normal, benign, malignant insitu and malignant invasive tissues tissues. Features from top layers were used for training and testing purposes. Also, some feature extracted layers were frozen and fine-tuned. Inception Restnet V2 network was used in the proposed methodology. Data augmentation was recommended to overcome shortage of samples.

In [46], many researchers focused to overcome the complexity of CNN's in training a huge number of parameters. In this paper features from images were extracted using transfer learning. Alexnet and VGG16 were used for this purpose. Extracted feature vectors were then segregated into separate classes by Support Vector Machine (SVM). Large number of experiments were carried out on available datasets. After evaluation of results, it was proposed that transfer learning outperformed other deep learning techniques.

In [47], deep convolutional neural network which was based on transfer learning was proposed to carry our multi-class breast cancer classification. A pre-trained deep-CNN was inherited. The strategy used concatenates the features. The approach outperformed previous techniques by producing 94.3% and 97.5% accuracies respectively.

In [48], a framework which uses deep and transfer learning was proposed. The framework detects and segregates the whole cytology dataset into various classes. The technique extracts features from images using pre-trained CNN's, such as Visual Geometry Group Network (VGGNet), Residual Networks (ResNet) and GoogLeNet. The data from these neural networks were fed into fully connected layers for segregation into desired classes. Average pooling classification was used for this purpose. To evaluate the performance, experiments are performed on datasets. The proposed technique outperformed other deep learning techniques for distinguishing cancerous cells.

2.3 Discussion

In this chapter, we have reviewed the research carried out by various researchers from multiple perspectives i.e. Machine Learning, Deep Learning and Transfer Learning techniques. Most of the models proposed previously, majorly uses various machine and

deep learning techniques to differentiate malignant cells from benign one.it was also observed that the architecture used by researchers is mostly Googlenet ,Resnet , Alexnet or a combination of VGG -16 with anyone of these.

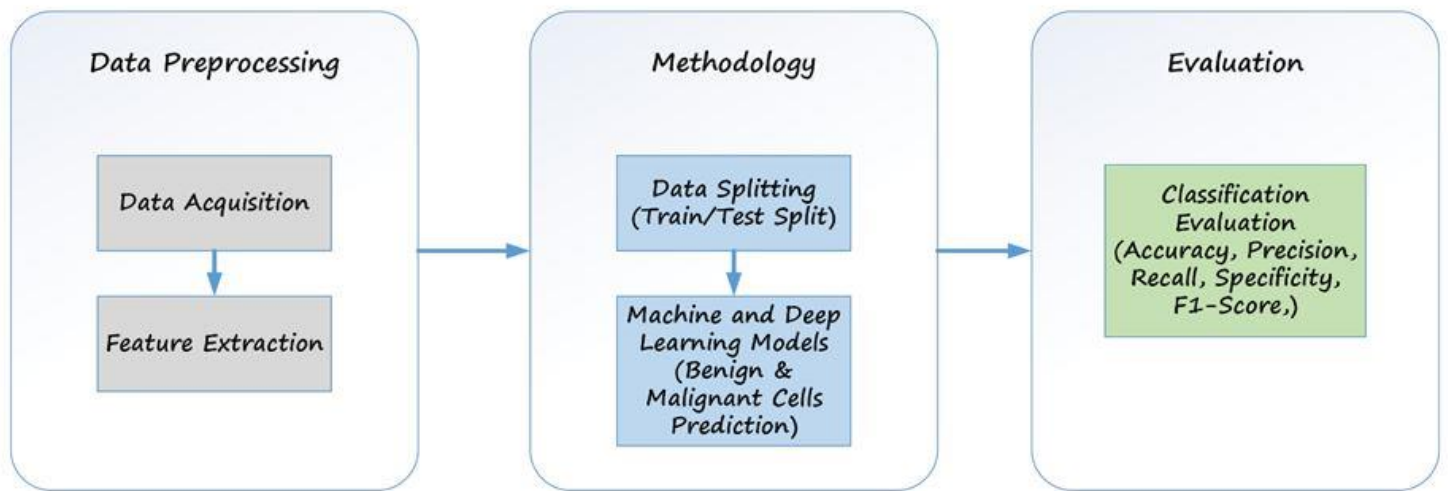
It is observed that the number of datasets and the classification techniques used were limited. There is need for a research which focuses on the application of transfer learning using neural network in which low level features can be extracted and fed to techniques for classification. We also intend to use VGG-19, which was not seen being used in any of the previous work. This can increase the accuracy to of classification. Also, most of the research was carried out internationally, there is a need to do research on breast cancer on national level keeping in view the statistics of our country. Our motivation would be to overcome these gaps through our research.

Chapter 3: Methodology

Before discussing our proposed methodology in detail, we will have a brief overview of the whole concept in the following few paragraphs. We have used concept of feature extraction using transfer learning. In transfer learning we take a pre-trained neural network on any available dataset and use it to distinguish categories of objects/images it was never trained/tested on. Robust and discriminative filters of challenging datasets which are learned by state-of-the-art network can be utilized. We can then apply these networks to recognize objects and images these models were never trained on.

Fist we have selected three open source online available image datasets on which we have run the techniques in our research. The datasets are explained in upcoming section. In the second step, we used the concept of feature extraction using transfer learning, we utilized VGG19 for feature extraction of images available in these datasets. VGG19 was utilized as a pre-trained neural network and also to extract arbitrary feature. We have allowed the image to move through the network in forward direction, we stop the image at a specified layer which is preselected, and outputs of the layer were stored as our features.

In the third & final step, we have applied six machine and deep learning techniques on these extracted features. Major machine learning classifiers that were used are: Logistic regression, K-NN, SVM, Decision Trees, Random Forest and Naïve Bayes. In deep learning, we utilized Artificial Neural Network (ANN) to segregate malignant and benign tumors. Feature extracted from each dataset were provided as training samples to these



datasets and then afterwards their performance was tested on the following parameters i.e. accuracy, precision, recall, specificity, F1-score. The algorithm with the best performance on a specific dataset was proposed to be used on that dataset.

The overall methodology is illustrated in the following diagram:

Figure 4 : Methodology Flow

3.1 Tools Used

The software / tools used in the research are:

3.1.1 Anaconda

Anaconda is the world's most popular environment for Python. All over the world it has a user base of more than 20 million people. Anaconda aims to make package deployment as simple and as easy as possible. Machine learning engineers, data and computer scientists mainly use it for their research purposes. It includes 1500 famous packages of python related to these fields. Also, it is a cross-platform environment i.e. it is available for Windows, Linux and MacOS.

3.1.2 Spyder

Spyder is a powerful environment jotted down in python. It is specifically designed for data scientist, engineers, researchers, and data analysts. It provides users, a combination of advanced analysis, editing debugging and profiling functionality. Data exploration and beautiful visualizations are the distinguishing feature of the development tool. It also enables users to further extend their abilities via API's and plugin systems. It can also be used as a PyQt5 extension library, which allows developers and researchers to build upon its own functionality and to embed further components.

3.1.3 Jupyter Notebook

Jupyter Notebook is the most famous browser based Integrated Development Environment. It is much lighter than other IDE's available in the market. Jupyter Notebook is a web app that is open-source, which allows it's users to manage and share important documents that contains code, text and mathematical equations. Visualization can also be shared using this IDE. Data cleaning, Data analysis, Data transformation, Statistical Modelling, numerical simulation and machine/deep learning are the major uses of Jupyter Notebook.

3.2 Data Acquisition and Feature Extraction

The initial and most important part of any research is a definitive and proper data which defines and drives the research. We collected three major datasets from various sources. In this part we gathered data from the specified sources. We utilized the concept of transfer learning using feature extraction and extracted features by a pretrained neural network (VGG19). These features were fed to train/test various machine/deep learning algorithms.

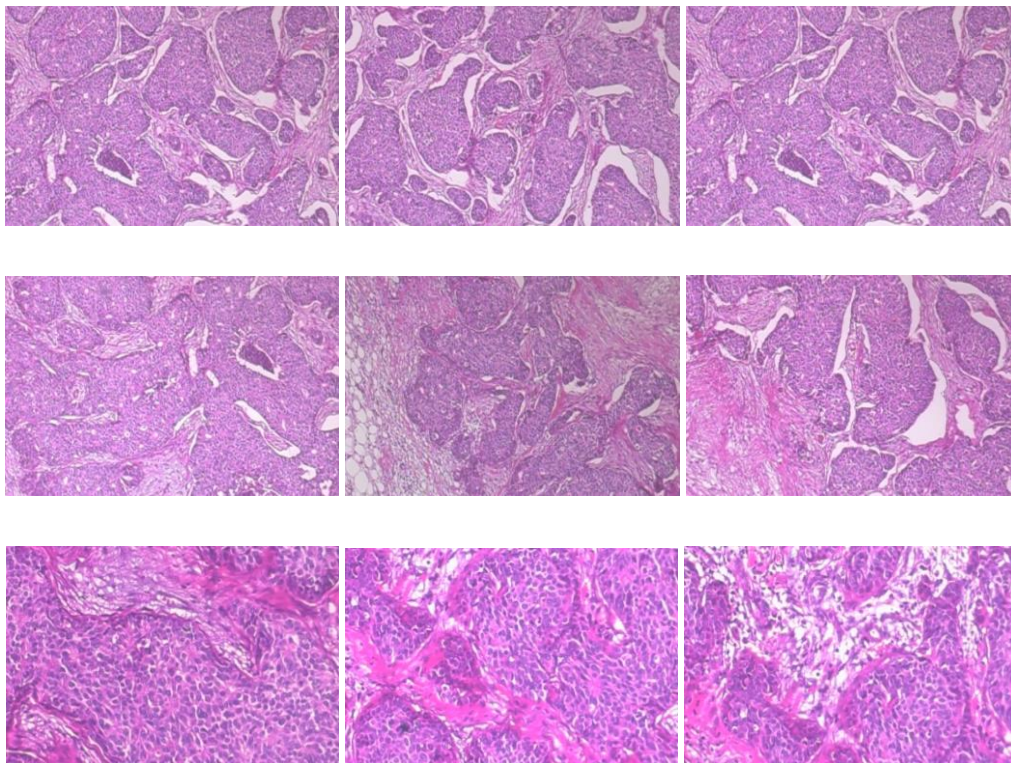
3.2.1 Datasets Description

There are not many open source datasets available on the web related to Breast Cancer images because the hospitals do not provide them readily because of privacy issues. We found the following three datasets and chose them for our study purpose.

3.2.1.1 Breast Cancer Histopathological Database (BreakHis) Dataset:

This dataset consists a total of 7909 microscopic images of breast tissue that contains tumors collected from various patients. It has been made with collaboration of P&D laboratory (Pathological Anatomy and Cytopathology, Parana, Brazil) [27].

The dataset is categorized into 2480 benign and 5429 malignant tumor images and we have used 6000 images from this dataset for our experiment. For testing, 20% and for training 80% data is used.



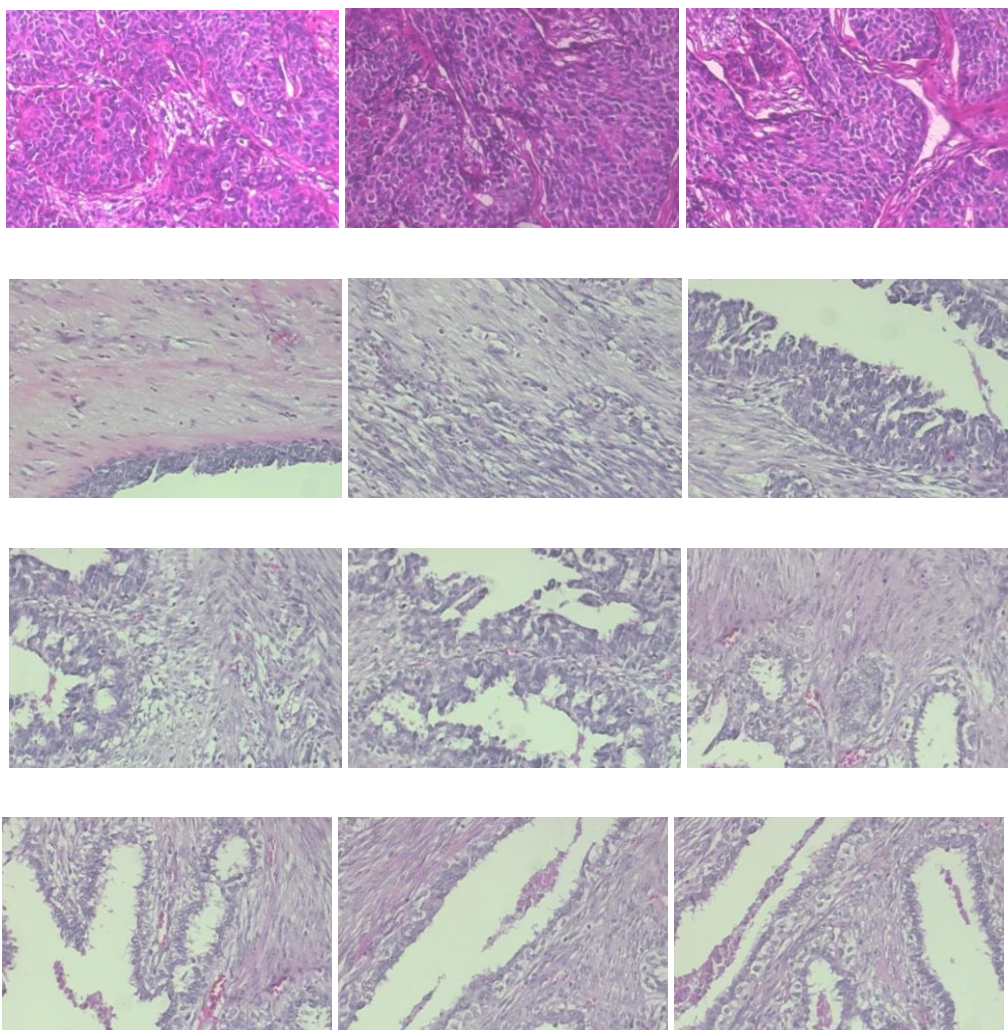


Figure 5: Sample images BreakHis image dataset

3.2.1.2 The IDC (Invasive Ductal Carcinoma) Breast Histopathology Images Dataset:

This dataset originally contains 162 whole mount images of breast cancer specimen using one 40x magnifying factor. 277,524 patches have been extracted from these out of which 198,738 patch level images are negative and 78,786 patch level images are positive. [28]

We have used 11,000 images from this dataset for our research and experiment. For testing, 20% and for training 80% data is used.



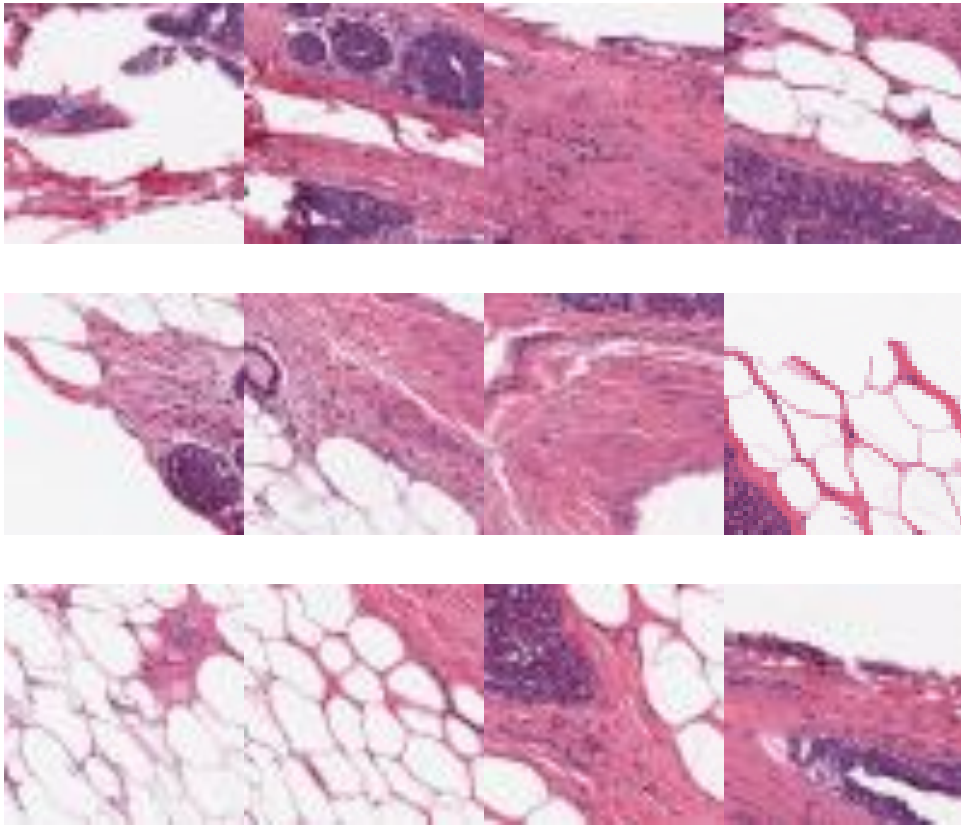
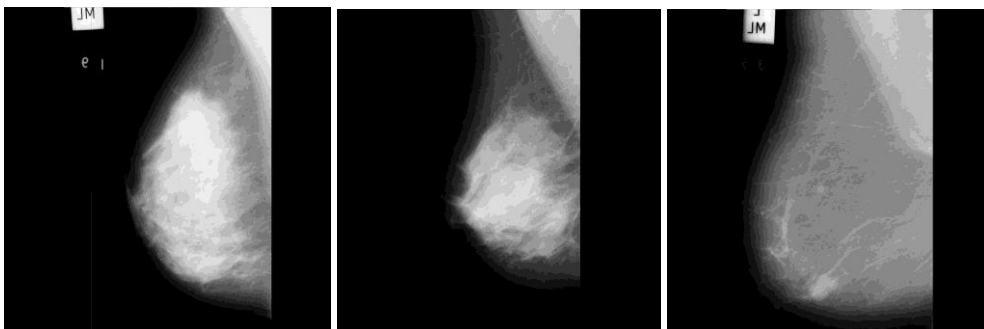
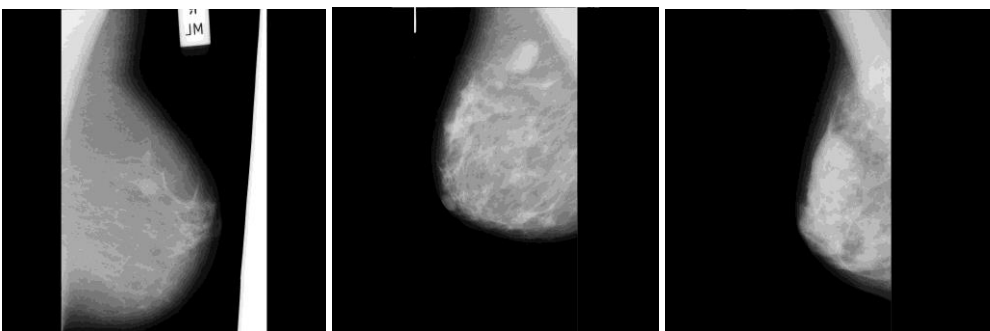
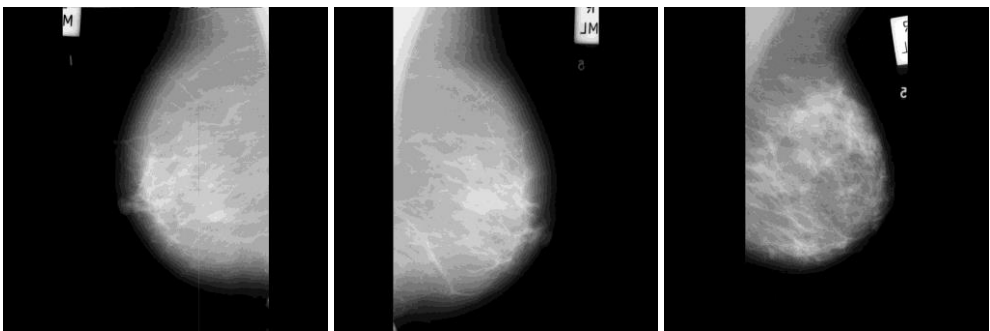
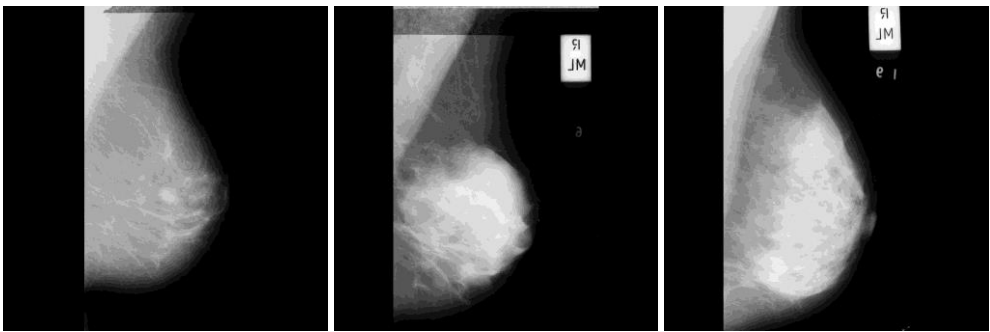


Figure 6: Sample Images IDC Breast Histology image dataset

3.2.1.3 The mini-MIAS dataset of Mammograms:

This dataset contains 323 breast mammogram images taken at 200 micron pixel edge where every image is 1024 x 1024 pixels. This dataset has been provided by the university of Essex United Kingdom. We have used 200 images for our scientific research experiment. For testing, 20% and for training 80% data is used. [29]





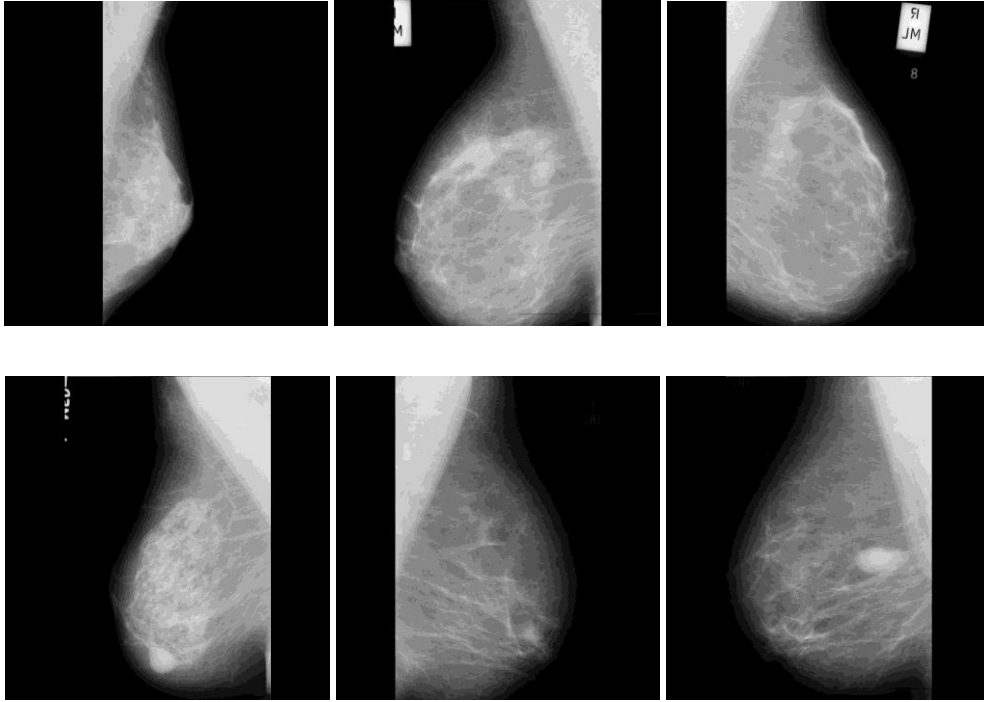


Figure 7: Sample images Mini Mias Mammographic Image dataset

3.2.2 Feature Extraction using Pre-trained Neural Network

In transfer learning we take a pre-trained neural network on any available dataset and use it to distinguish categories of objects/images it was never trained/tested on. Robust and discriminative filters of challenging datasets which are learned by state-of-the-art network can be utilized. We can then apply these networks to recognize objects and images these models were never trained on. In case, we are using deep learning along with transfer learning, transfer learning falls into two major categories.

- Transfer learning using feature extraction
- Transfer learning using fine-tuning

3.2.2.1 Transfer learning using feature extraction

When we perform feature extraction using transfer learning, we utilize a neural network which is pre-trained on some images. In this case, it is also used to extract

arbitrary features. Image is allowed to move forward through the network, we stop the image at a specified layer which is preselected, and outputs of the layer are taken as our features.

3.2.2.2 Transfer learning using fine-tuning

In case of fine-tuning, we must update the architecture of the model itself and heads of previous fully connected layers must be removed. New and freshly initialized heads are provided, and then new layers are trained to predict our input classes.

Deep Neural networks which are trained using large number of samples (large datasets) are quite good at transfer learning. These neural networks can learn, rich and large set of features and can recognize 1000s of segregated classes. It is a good approach if we reuse these neural networks for other tasks for which it was not specifically trained.

Typically, a CNN can act as an image classifier. If an image is provided as an input to network, it can move forward in the network and classification probabilities are obtained wherever the network ends. There is no such rule that defines that we should allow input image to pass through whole network. Propagation of Input image can be halted at any pre-specified layer. We can extract values of the layer and use them as feature vectors.

When we use networks to extract features, we usually “cut off” the network at a given layer which is preselected (the selection of layer varies, most of the networks are chopped off at fully connected layers, but it actually depends on the particular dataset being used).

3.2.2.3 VGG16 and VGG19

Simonyan and Zisserman introduced the architecture of VGG network in a paper published in 2014, the name of their publication was: “*Very Deep Convolutional Neural Networks for Large Scale Image Recognition.*”

The designed network was very simple. Convolutional layers of 3x3 size are placed over each other in sequence of increasing depth. Max pooling reduces volume size. Softmax layer is followed by 4096 nodes and 2 fully connected layers. The “19” and “16” determines quantity of weight layers in network. In 2014, if a network contains 16 or 19 layers in it, it was considered very deep but nowadays ResNet architectures are trained at a depth of 50-200 layers for ImageNet.

Unfortunately, VGGNet comes with 2 drawbacks. They are slow to train, and weights of the architecture are quite large. Due to increased depth and large number of fully connected layers the size of VGG is quite large. VGG16 is 533MB in size while size of VGG19 is 574MB. The larger size makes deployment a bit difficult. VGG can be employed in problems relevant to classification which use deep learning.

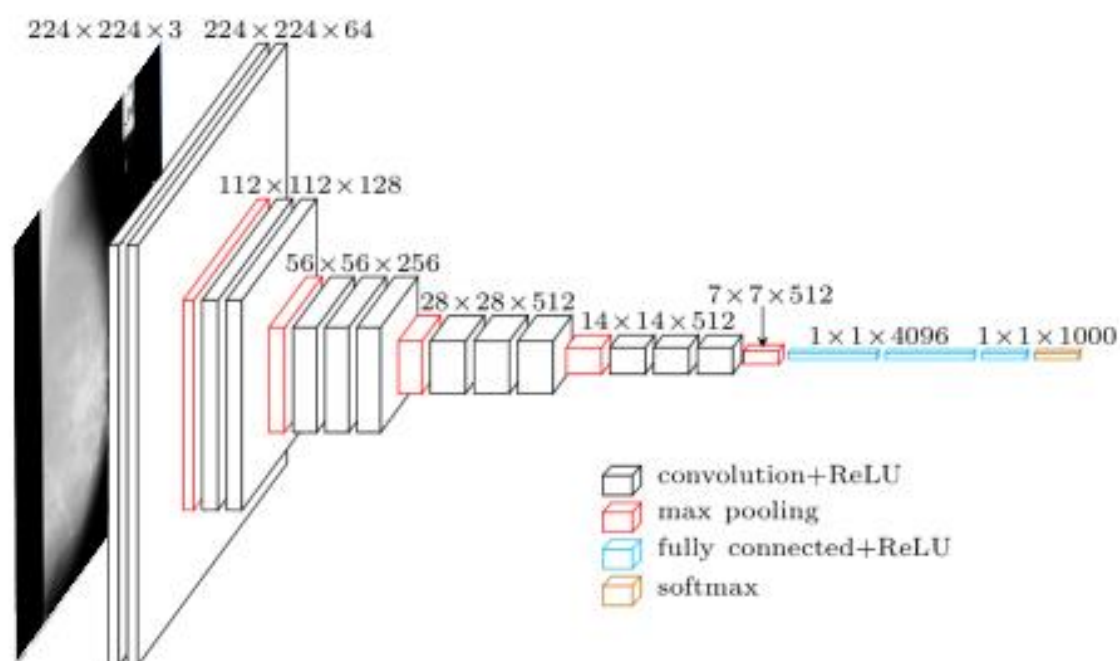


Figure 8: General VGG Architecture [49]

We are using 19-layer VGG19 CNN to extract features. The network was initialized with pretrained ImageNet weights. The activation layers of a pretrained convolutional neural network can be utilized to extract information relevant to feature vectors. The initial layers activations represent features such as edges, these are low level features, while deep layers represent features (high-level features) which are efficient for classification. In VGG19 the activations of fully connected layers represent features which are salient for tasks related to image recognition. The resulting feature vector contains a total of 2048 features. We have extracted these features which are then used as training/testing samples for various machine/deep learning classifiers.

The overall phenomena is illustrated in the following figure.

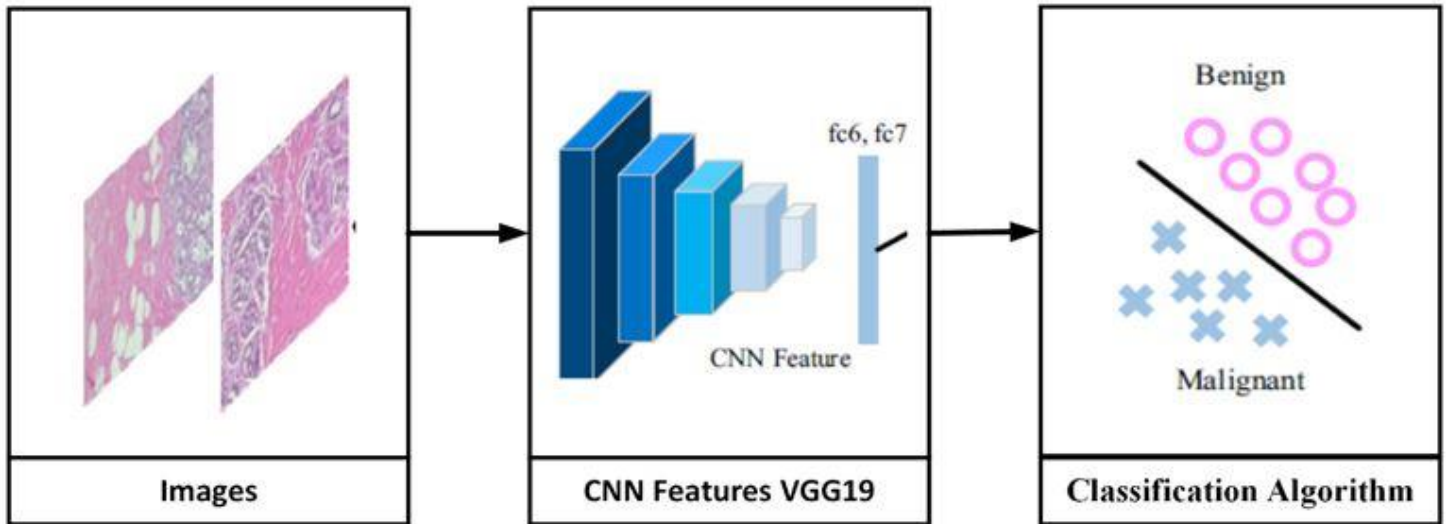


Figure 9: Deep feature Extraction [46]

3.2.3 Data Splitting

The final step before we apply machine learning algorithms to our extracted features is to split the gathered data into training and testing samples. A considerable chunk of the data is used for training and then testing data is used to evaluate its performance. The results of the testing data are passed to accuracy measures to evaluate the performance. This research uses Train-Test split to separate out the training and testing samples.

3.2.3.1 Train-Test Split

In Train-Test split the acquired data is separated into two subsets, one set is named as “Training set” and the other one is named as “Testing Set”. Proposed models are trained on first set i.e. training data set. After training the model is evaluated on second set i.e. “Testing Set”. But it comes with a risk of data not being properly split and test data leaking information into training data. This research uses splits data into 80% training data and 20% test data.

3.3 Machine and Deep Learning Algorithms

The following machine/deep learning algorithms are utilized to segregate the extracted features into required classes i.e. benign and malignant.

3.3.1 Logistic Regression

It is a very basic classification algorithm based on logistic/ sigmoid function. It works on the concept of probability. It is the most popular and simple algorithm for segregating the testing and training data into desired classes. In our case we have used it to classify the benign and malignant cells. [30]

$$p = \frac{1}{1 + b(\beta_0 + \beta_1 x_1 + \dots + \beta_m x_m)} \quad \dots (1)$$

3.3.2 K-Nearest Neighbors (K-NN)

K-NN segregates the given datapoints into various classes. It computes the N nearest neighbors for the point and assigns it to the class having majority of N neighbors in it. K-NN is not used for datasets having very large number of samples as it computes the nearest neighbors each time a class for a sample point is to be computed. [31]

$$R_R (C^{w_{nn}}) - R_R (C^{Bayes}) = (B_1 S_n^2 + B_2 t_n^2) \{1 + o(1)\} \quad \dots (2)$$

3.3.3 Support Vector Machines (SVM):

SVM is majorly used for classifying data into various classes, but sometimes they are employed to perform regression as well. To correctly distinguish datapoints between separate classes SVM maximizes distinction between them by creating a hyperplane which leads to reduced number of miscalculations. [32]

$$\frac{1}{2} \|w\|^2 + \lambda \left[\frac{1}{n} \sum_{i=1}^n \max(0, 1 - y_i(w \cdot x_i - b)) \right] \quad \dots (3)$$

3.3.4 Naïve Bayes:

The algorithm has its foundations in the Bayes theorem which assumes the total probability of presence of any number of features are unrelated to each other. It is a very fast and simple algorithm for classification purposes. [33]

$$p\left(\frac{C_k}{x}\right) = \frac{p(C_k)p\left(\frac{x}{C_k}\right)}{p(x)} \quad \text{--- (4)}$$

3.3.5 Decision Tree:

It is a very self-explanatory and simple classifier majorly used for classifying data into various classes, but sometimes it is employed to perform regression as well. The algorithm keeps all the input parameters in consideration while making predictions and selects root variable using entropy. [34]

$$E(T, X) = \sum_{c \in X} P(c)E(c) \quad \text{--- (5)}$$

3.3.6 Random Forest:

Random forest segregates data into various classes by using various base models which are created using subsets of given data by decision trees classifier. The final decision is based on results accumulated from all the models. It is much more efficient than decision trees classifier because it utilizes all the pros that comes by using decision trees as base model and additional efficiency is achieved by the usage of multiple models. [35]

$$f = \frac{1}{B} \sum_{b=1}^B f_b(x') \quad \text{--- (6)}$$

3.3.7 Artificial Neural Network (ANN):

All neural networks are loosely structured like network of neurons available in our human brain. They contain hundreds of thousands of interconnected nodes with various number of layers. The input and output layers in a neural network are mandatory. The number of hidden layers varies from network to network depending upon the complexity and usage of each of them. Features are passed as predicting parameters to input layers and output layer shows the actual predictions. The model is generalized by updating weights on each node on every iteration. Predictions on test set are made by using these weights. Neural Nets are mostly used for classifying given datapoints but can be used as regressors whenever needed. [36]

3.4 Classification Testing Parameters:

Before discussing various testing parameters, we have to keep a few basic concepts and definitions in minds. They are discussed as follows:

3.4.1 Confusion Matrix

Confusion matrix is shown below:

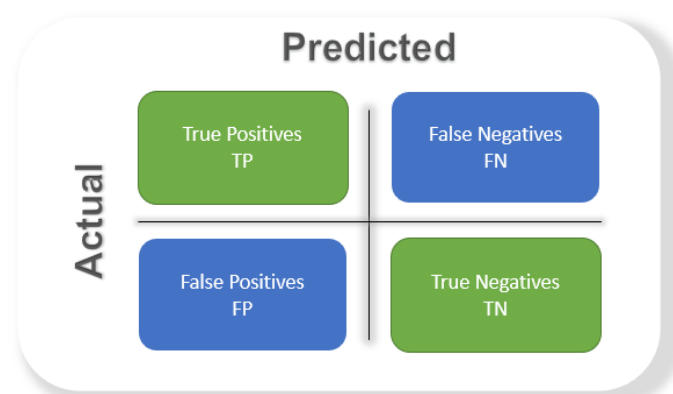


Figure 10 Confusion Matrix

Some common terms to be clear with are:

True positives (TP): Predicted positives which are actually positive.

False positives (FP): Predicted positives which are actually negative.

True Negatives (TN): Predicted negatives which are actually negative.

False Negatives (FN): Predicted negatives which are actually positive.

Now that we know what a confusion matrix is and the actual meaning behind the terms i.e. True positive, True negative, False positive, False negative we will discuss the parameters for classification of images into malignant and benign. For this purpose, following parameters have been used.

3.4.2 Accuracy:

Accuracy is basically, the ratio of total predictions which were correctly predicted to total input samples. [37].

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN} \quad \text{--- (7)}$$

3.4.3 Precision:

It is proportion of positive instances out of total instances which were predicted positive. [37].

$$Precision = \frac{TP}{TP + FP} \quad \text{---(8)}$$

3.4.4 Recall:

It is proportion of positive instances out of total instances which were actually positive. [37].

$$Recall = \frac{TP}{TP + FN} \quad \text{---(9)}$$

3.4.5 Specificity

It is ratio of negative instances out of total instances which were actually negative. Here, the denominator represents (TN+FP) total actual negative instances that are available in the given dataset. [12].

$$\frac{TN}{TN + FP} \text{ ---(10)}$$

3.4.6 F1-Score:

Recall and Precision are valid accuracy measures but individually, they don't cover all aspects. So, we compute harmonic mean to cover all aspects. Harmonic mean of precision and recall is F1 score. Its values always remain between 0 and 1. Higher F1-score relates to better performance. [37].

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

In this chapter we overviewed the methodology. We explained our datasets used for research. Following that, we performed feature extraction by utilizing a pretrained neural network for training/testing various algorithms. We also discussed testing parameters on which we will be evaluating our results. We applied various machine and deep learning algorithms on our datasets. In next chapter we will discuss results and analyze them one by one.

Chapter 4: Results and Analysis:

4.1 Dataset 1: Breast Cancer Histopathological Database (BreakHis)

Since manually differentiating malignant cells from benign cells is complex for human eye and requires a lot experience and expertise in the relevant field. The systems employed for these complex tasks, evaluate their results on a variety of parameters. We have used five parameters to evaluate our results. While employing the following machine and deep learning algorithms, we found Logistic Regression and Artificial Neural Network to be the most efficient algorithms having an accuracy of 93.86% and 93.64% respectively.

Table 1: Results and Comparison Dataset 1

<i>Algorithms</i>	<i>Accuracy</i>	<i>Precision</i>	<i>Recall</i>	<i>Specificity</i>	<i>F1 Score</i>
<i>Logistic Regression</i>	0.9386	0.8664	0.8362	0.9658	0.851
<i>K-Nearest Neighbors</i>	0.9299	0.9007	0.7429	0.9786	0.8142
<i>Support Vector Machines</i>	0.9158	0.7966	0.7966	0.9469	0.7966
<i>Naïve Bayes</i>	0.8188	0.5499	0.7542	0.8357	0.6327
<i>Decision Tree</i>	0.841	0.6288	0.5876	0.9071	0.6047

Random Forest	0.8825	0.9048	0.4831	0.9867	0.6299
Artificial Neural Network	0.9364	0.8311	0.8746	0.9529	0.8523

The following graphs shows these visualizations:

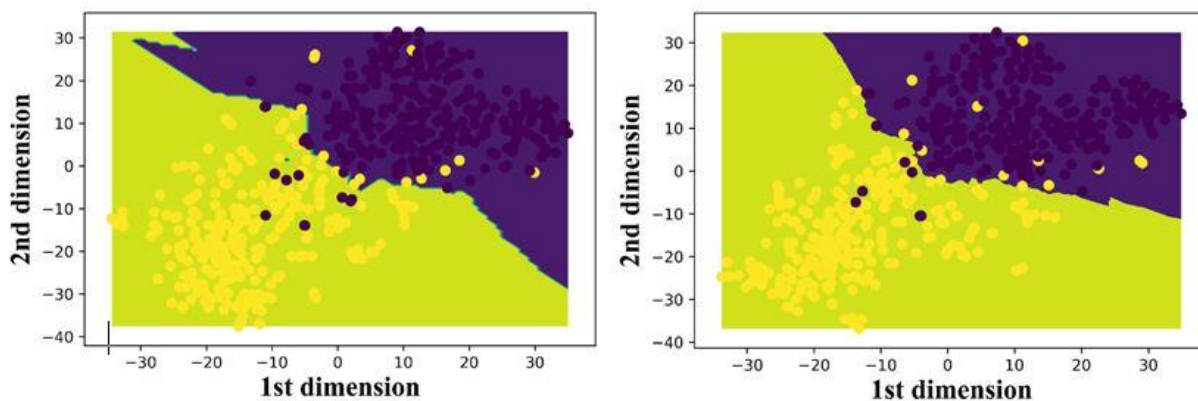


Figure 11: Graphical Representation of extracted features along with decision boundaries for Logistic Regression and Artificial Neural Network – Dataset 1

4.1.1 Winner Techniques: Logistic Regression and Artificial Neural Network

We observed in Figure 11 that the two classes (malignant and benign) are visually distinct and a clear decision boundary along with very few false positives and negatives can be seen in case of best performing models which results in higher performance for these models.

Every algorithm performs different on the same dataset as per its requirement. From the table it can be seen that Naïve Bayes and Decision Tree has performed low as compared to other techniques. This is because the sample images for training were not

enough for these techniques to achieve better accuracy. The more the samples, the more accurate results we have.

The accuracy could have been made better in decision tree , by increasing the depth of trees but in doing so, we only achieve a slight increase in accuracy (approx. 10 percent) which is very insignificant and is a big trade off with algorithm implementation complexity (training testing time of model).

4.2 Dataset 2: The IDC (Invasive Ductal Carcinoma) Breast Histopathology Images

In this dataset, we found Logistic Regression and Artificial Neural Network to be the most efficient algorithms having an accuracy of 92.45% and 93.3% respectively.

Table 2: Results and Comparison Dataset 2

<i>Algorithms</i>	<i>Accuracy</i>	<i>Precision</i>	<i>Recall</i>	<i>Specificity</i>	<i>F1 Score</i>
Logistic Regression	0.9245	0.892	0.8485	0.9566	0.8697
K-Nearest Neighbors	0.9112	0.8763	0.8117	0.9525	0.8428
Support Vector Machines	0.8931	0.8242	0.8131	0.9268	0.8186
Naïve Bayes	0.8681	0.7191	0.9007	0.8515	0.7997

Decision Tree	0.8796	0.8038	0.7862	0.919	0.7949
Random Forest	0.917	0.8948	0.9165	0.9595	0.8539
Artificial Neural Network	0.933	0.8993	0.8721	0.9588	0.8855

The following graphs shows these visualizations:

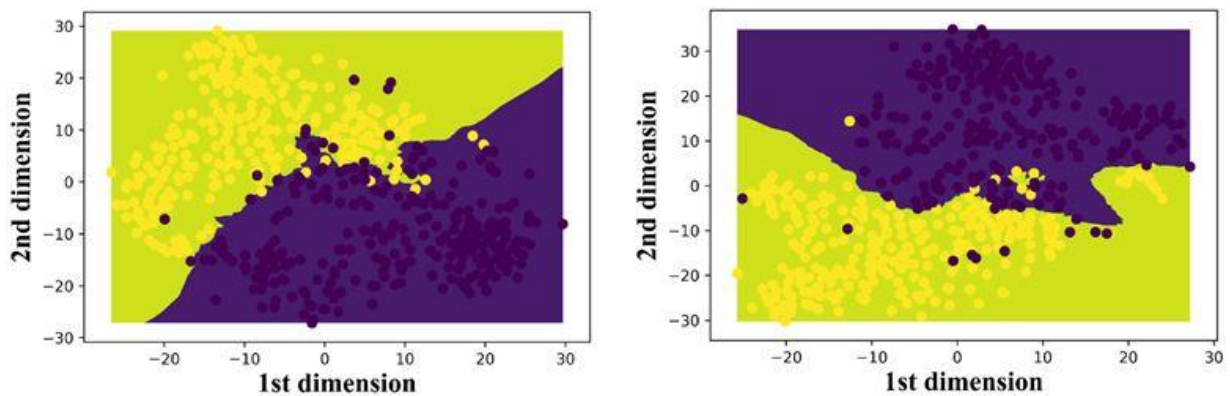


Figure 12: Graphical Representation of extracted features along with decision boundaries for Logistic Regression and Artificial Neural Network - Dataset 2

4.2.1 Winner Techniques: Logistic Regression and Artificial Neural Network

We observed in Figure 12 that the two classes (malignant and benign) are more visually distinct and a clear decision boundary along with very few false positives and negatives

can be seen in case of best performing models which results in higher performance for these models.

As we can see from the table that there is a slight variance in the accuracy of the trained models. This is because that the images in the dataset were of higher resolution and the dataset was equally divided among the two classes (malignant and benign). The accuracy of the low performing algorithms such as Support Vector Machines and Naïve Bayes could be enhanced if we increase the number of samples used for the training and testing of these algorithms but again, at the cost of implementation complexity.

4.3 Dataset 3: The mini-MIAS dataset of Mammograms

In this dataset, we found Logistic Regression and Artificial Neural Network to be the most efficient algorithms having an accuracy of 82.5% and 87.5% respectively.

Table 3: Results and Comparison Dataset 3

<i>Algorithms</i>	<i>Accuracy</i>	<i>Precision</i>	<i>Recall</i>	<i>Specificity</i>	<i>F1 Score</i>
Logistic Regression	0.825	0.8261	0.8636	0.7778	0.8444
K-Nearest Neighbors	0.540	0.5714	0.5926	0.4783	0.5818
Support Vector Machines	0.820	0.8462	0.8148	0.8261	0.8302
Naïve Bayes	0.780	0.8077	0.7778	0.7826	0.7925

Decision Tree	0.580	0.6364	0.5185	0.6522	0.5714
Random Forest	0.4828	0.60	0.6316	0.40	0.6154
Artificial Neural Network	0.875	0.9048	0.8636	0.8889	0.8837

The following graphs shows these visualizations:

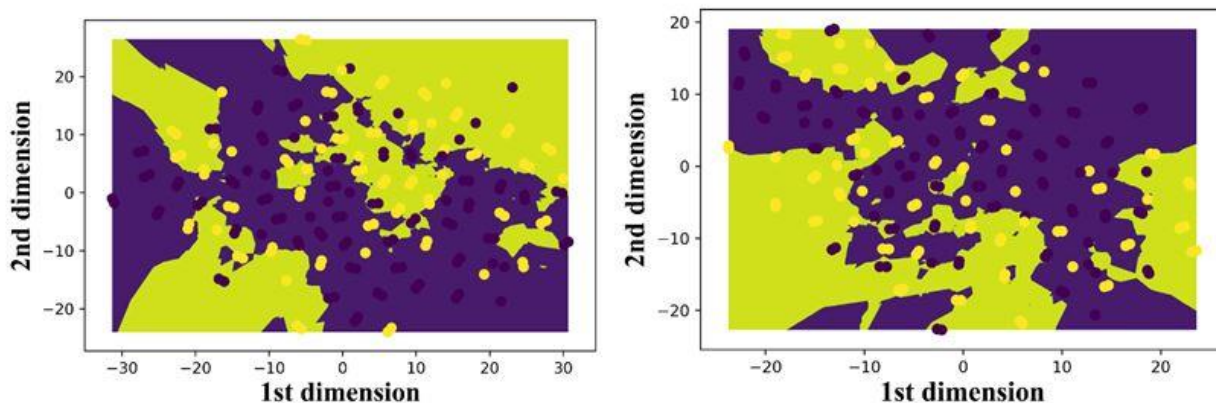


Figure 13: Graphical Representation of extracted features along with decision boundaries for Logistic Regression and Artificial Neural Network – Dataset 3

4.3.1 Winner Techniques: Logistic Regression and Artificial Neural Network

We observed in Figure 13 that the two classes (malignant and benign) are more visually distinct and a clear decision boundary along with very few false positives and negatives

can be seen in case of best performing models which results in higher performance for these models.

As we know that in most cases, the more data we have for training, more is the accuracy we get from the algorithms. For this dataset, we had only a limited number of samples (approx. 250) and due to this fact, some of the above-mentioned classifiers such as k-nearest neighbors, decision trees and random could not perform well. These algorithms require a large number of samples to train themselves on the patterns or features extracted from the deep neural network (i.e. in case of large datasets, performance of algorithms like decision trees and random forests can be enhanced by increasing the depth, which in turn increases the training and testing time exponentially but leads to better accuracy).

4.4 Discussion

In this chapter, we iterated thoroughly, through our study's results and compared the performance of all the classifiers. We used VGG-19 Architecture and calculated Accuracy, Precision, Recall, F1 score and specificity for all datasets and established that VGG-19 gives best results for Logistic Regression Classifier and Artificial Neural Network outperforming other techniques giving an average accuracy of 89.5% and 91.4%.

We also observed that although all the techniques achieved good accuracies but those classifiers which could not perform very well was due to the fact that the image samples were less to train. Also in these techniques, the number of false positives and false negatives were greater and the datapoints were not visually distinctive.

Chapter 5: Conclusion and Future Work

Breast cancer is the most common type of cancer being diagnosed in women in Pakistan and around the world. Due to technological advancements in medical science, the survival rate for patients is increasing day by day, but there is still a long way to go when it will finally come in control. Dense breast tissue, estrogen exposure, body weight, alcohol consumption, age, genetics and radiation exposure are the major factors that can cause breast cancer. Air pollution is also a high risk. Carcinogens, greenhouse gases, pesticide spray on food, diet, increased population, development of industry and alleviated living standard also increases the chances of this fatal disease. In Pakistan, main problem is that patients show up at last stages of cancer which then cannot be cured resulting in increasing deaths. Since Pakistan is a third world country with lack of basic medical facilities, cancer at last stages cannot be cured but it has been observed that initial diagnosis of this disease can reduce morbidity and mortality. There are a few traditional prognosis techniques like mammography, biopsy, MRI etc. In recent years, Machine learning has proved to be vital in finding solutions to medical problems such as these. Various techniques such as Logistic regression, K-NN, SVM, Decision Trees, Random Forest and Naïve Bayes etc. aid to learn trends available in the images of breast cancer, These trends then enable these algorithms to predict anomalies in the cells.

The aim of our research was to predict the malignant tumors from benign one to help doctors to diagnose cancer. A comparison was presented between several techniques. Majorly, we have used concept of feature extraction using transfer learning. In transfer learning we have taken a pre-trained neural network on any available dataset and use it to distinguish categories of objects/images it was never trained/tested on. We have used it to extract features of images. For this purpose, we utilized VGG19 to extract arbitrary features. VGG-19 has not been previously used in research before and it produced excellent results. We have allowed the image, to propagate through network in forward direction, we stop the image at a specified layer which is preselected, and outputs of the layer were stored as deep features for image recognition. Finally, we have applied a variety of machine/deep learning techniques on these extracted features.

The popular machine learning classifiers that were used are: Logistic regression, K-NN, SVM, Decision Trees, Random Forest and Naïve Bayes. In deep learning, we have used Artificial Neural Network (ANN) to classify malignant and benign tumors. Feature extracted from each dataset were provided as training samples to these datasets and then afterwards their performance was tested on basis of precision, recall, accuracy, F1-score and specificity. The algorithm with the best performance on a specific dataset was proposed to be used on that dataset. Logistic regression and Artificial Neural Network outperformed other algorithms in distinguishing cancerous and non-cancerous cells. Because of non-availability of good breast cancer datasets, the sample images were limited and hence rest of the techniques could not perform as much as the ANN and logistic regression. We could still have achieved more results by increasing the implementation complexity but that would be a big cost as compared to a slight improvement in accuracy.

5.1 Limitation/Future Work:

For future advancement, this research can be applied in following ways:

1. Our research focuses only on distinguishing between two types of cancerous cells i.e. benign and malignant. Apart from these two, cancer cells can be classified into more sub-classes.
2. These machine learning techniques can also be used as detection for other cancer like lung cancer, skin cancer etc.
3. The open source breast cancer image datasets are very less and are not easily accessible which limits application of techniques and analysis. More such databases can be made available for better research.

References

- [1] Araújo T, Aresta G, Castro E, Rouco J, Aguiar P, Eloy C, et al. (2017) Classification of breast cancer histology images using Convolutional Neural Networks. *PLoS ONE* 12(6): e0177544. <https://doi.org/10.1371/journal.pone.0177544>
- [2] Han, Z., Wei, B., Zheng, Y. *et al.* Breast Cancer Multi-classification from Histopathological Images with Structured Deep Learning Model. *Sci Rep* 7, 4172 (2017). <https://doi.org/10.1038/s41598-017-04075-z>
- [3] Rakhlin A., Shvets A., Iglovikov V., Kalinin A.A. (2018) Deep Convolutional Neural Networks for Breast Cancer Histology Image Analysis. In: Campilho A., Karray F., ter Haar Romeny B. (eds) *Image Analysis and Recognition. ICIAR 2018. Lecture Notes in Computer Science*, vol 10882. Springer, Cham
- [4] Ragab DA, Sharkas M, Marshall S, Ren J. 2019. Breast cancer detection using deep convolutional neural networks and support vector machines. *PeerJ* 7:e6201
- [5] Z. Wang *et al.*, "Breast Cancer Detection Using Extreme Learning Machine Based on Feature Fusion With CNN Deep Features," in *IEEE Access*, vol. 7, pp. 105146-105158, 2019, doi: 10.1109/ACCESS.2019.2892795.
- [6] Duc My Vo, Ngoc-Quang Nguyen, Sang-Woong Lee, Classification of breast cancer histology images using incremental boosting convolution networks, *Information Sciences*, Volume 482, 2019, Pages 123-138, ISSN 0020-0255, <https://doi.org/10.1016/j.ins.2018.12.089>.
- [7] Osmanović A., Halilović S., Ilah L.A., Fojnica A., Gromilić Z. (2019) Machine Learning Techniques for Classification of Breast Cancer. In: Lhotska L., Sukupova L., Lacković I., Ibbott G. (eds) *World Congress on Medical Physics and Biomedical Engineering 2018. IFMBE Proceedings*, vol 68/1. Springer, Singapore
- [8] S. Charan, M. J. Khan and K. Khurshid, "Breast cancer detection in mammograms using convolutional neural network," *2018 International Conference on Computing, Mathematics and Engineering Technologies (iCoMET)*, Sukkur, 2018, pp. 1-5, doi: 10.1109/ICOMET.2018.8346384.
- [9] Histopathological Breast Cancer Image Classification by Deep Neural Network Techniques Guided by Local Clustering. Abdullah-Al Nahid , Mohamad Ali Mehrabi, and Yinan Kong. *Hindawi BioMed Research International* ,Volume 2018, Article ID 2362108, <https://doi.org/10.1155/2018/2362108>
- [10] Wang, Shuihua et al. 'Abnormal Breast Detection in Mammogram Images by Feed-forward Neural Network Trained by Jaya Algorithm'. 1 Jan. 2017 : 191 – 211.
- [11] D. Bardou, K. Zhang and S. M. Ahmad, "Classification of Breast Cancer Based on Histology Images Using Convolutional Neural Networks," in *IEEE Access*, vol. 6, pp. 24680-24693, 2018, doi: 10.1109/ACCESS.2018.2831280.

- [12] L. Hussain, W. Aziz, S. Saeed, S. Rathore and M. Rafique, "Automated Breast Cancer Detection Using Machine Learning Techniques by Extracting Different Feature Extracting Strategies," *2018 17th IEEE International Conference On Trust, Security And Privacy In Computing And Communications/ 12th IEEE International Conference On Big Data Science And Engineering (TrustCom/BigDataSE)*, New York, NY, 2018, pp. 327-331, doi: 10.1109/TrustCom/BigDataSE.2018.00057.
- [13] Evaluating the Performance of Machine Learning Techniques in the Classification of Wisconsin Breast Cancer Omar Ibrahim Obaid 1* , Mazin Abed Mohammed , Mohd Khanapi Abd Ghani , Salama A. Mostafa , Fahad Taha AL-Dhief
- [14] Nazeri K., Aminpour A., Ebrahimi M. (2018) Two-Stage Convolutional Neural Network for Breast Cancer Histology Image Classification. In: Campilho A., Karray F., ter Haar Romeny B. (eds) *Image Analysis and Recognition. ICIAR 2018. Lecture Notes in Computer Science*, vol 10882. Springer, Cham
- [15] Yue, W.; Wang, Z.; Chen, H.; Payne, A.; Liu, X. Machine Learning with Applications in Breast Cancer Diagnosis and Prognosis. *Designs* **2018**, *2*, 13.
- [16] Sohail S, Alam SN. Breast cancer in Pakistan - awareness and early detection. *J Coll Physicians Surg Pak*. 2007;17(12):711-2.
- [17] WHO, Health Profile of Pakistan. <http://www.worldlifeexpectancy.com/country-health-profile/Pakistan>.
- [18] Siegel, R.L., Miller, K.D., Fedewa, S.A., Ahnen, D.J., Meester, R.G.S., Barzi, A. and Jemal, A. (2017), Colorectal cancer statistics, 2017. *CA: A Cancer Journal for Clinicians*, 67: 177-193. doi:10.3322/caac.21395
- [19] <https://www.who.int/cancer/PRGlobocanFinal.pdf?ua=1>
- [20] <https://gco.iarc.fr/tomorrow/home>
- [21] <https://cancerstaging.org/references-tools/deskreferences/Pages/Breast-Cancer-Staging.aspx>
- [22] Banning M, Hafeez H, Faisal S, Hassan M, Zafar A. 2009. The impact of culture and sociological and psychological issues on Muslim patients with breast cancer in Pakistan. *Cancer Nursing* 32, 317-24.
- [23] Ahmed F, Mahmud S, Hatcher J, Khan SM. 2006. Breast cancer risk factor knowledge among nurses in teaching hospitals of Karachi, Pakistan: a cross-sectional study. *BMC nursing* 5, 6.
- [24] Current situation of breast cancer in Pakistan with the available interventions Shumaila Arshad^{1,2} , Masood ur Rehman² , Farah Abid ¹ , Saleha Yasir¹ , Mehwish Qayyum³ , Kanwal Ashiq^{3*}, Samreen Tanveer³ , Mayyda Bajwa³ , Sana Ashiq⁴
- [25] Begum N. Breast Cancer in Pakistan: A Looming Epidemic. *J Coll Physicians Surg Pak*. 2018;28(2):87-88. doi:10.29271/jcpsp.2018.02.87
- [26] <https://htv.com.pk/womens-health/breast-cancer-treatment-in-pakistan>

- [27] <https://web.inf.ufpr.br/vri/databases/breast-cancer-histopathological-database-breakhis/>
- [28] <https://www.kaggle.com/paultimothymooney/breast-histopathology-images>
- [29] <http://peipa.essex.ac.uk/info/mias.html>
- [30] N. Amral, C. S. Ozveren and D. King, "Short term load forecasting using Multiple Linear Regression," *2007 42nd International Universities Power Engineering Conference*, Brighton, 2007, pp. 1192-1198, doi: 10.1109/UPEC.2007.4469121
- [31] Kevin Beyer., Jonathan Goldstein., Raghu Ramakrishnan. and Uri Shaft. 1999. When is "nearest neighbor" meaningful?. *International conference on database theory* (217-235)
- [32] Simon Tong. & Daphne Koller. 2001 Support Vector Machine Active Learning with Applications to Text Classification. *Journal of Machine Learning Research* (2001). 45-66
- [33] Harry Zhang. 2004 The Optimality of Naive Bayes. *Proceedings of the Seventeenth International Florida Artificial Intelligence Research Society Conference, Miami Beach, Florida, USA*
- [34] J. R. Quinlan, "Decision trees and decision-making," in *IEEE Transactions on Systems, Man, and Cybernetics*, vol. 20, no. 2, pp. 339-346, March-April 1990, doi: 10.1109/21.52545.
- [35] Andy Liaw. & Wiener. 2002 Classification and Regression by randomForest
- [36] A. K. Jain, Jianchang Mao and K. M. Mohiuddin, "Artificial neural networks: a tutorial," in *Computer*, vol. 29, no. 3, pp. 31-44, March 1996, doi: 10.1109/2.485891.
- [37] SanaUllah Khan, Naveed Islam, Zahoor Jan, Ikram Ud Din, Joel J. P. C Rodrigues, A novel deep learning based framework for the detection and classification of breast cancer using transfer learning, *Pattern Recognition Letters*, <https://doi.org/10.1016/j.patrec.2019.03.022>.
- [38] www.cancer.gov
- [39] <https://www.pinkribbon.org.pk/risk-factors/>
- [40] www.pinkribbon.org.pk
- [41] <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4441973/>
- [42] <https://shaukatkhanum.org.pk/bca-articles/>
- [43] TIWARI, MONIKA and Bharuka, Rashi and Shah, Praditi and Lokare, Reena, Breast Cancer Prediction Using Deep Learning and Machine Learning Techniques (March 22, 2020). Available at SSRN: <https://ssrn.com/abstract=3558786> or <http://dx.doi.org/10.2139/ssrn.3558786>

- [44] D. Selvathi and A. AarthyPoornila, "Performance analysis of various classifiers on deep learning network for breast cancer detection," *2017 International Conference on Signal Processing and Communication (ICSPC)*, Coimbatore, 2017, pp. 359-363, doi: 10.1109/CSPC.2017.8305869
- [45] Ferreira C.A. et al. (2018) Classification of Breast Cancer Histology Images Through Transfer Learning Using a Pre-trained Inception Resnet V2. In: Campilho A., Karray F., ter Haar Romeny B. (eds) *Image Analysis and Recognition. ICIAR 2018. Lecture Notes in Computer Science*, vol 10882. Springer, Cham
- [46] Deniz, E., Şengür, A., Kadiroğlu, Z. *et al.* Transfer learning based histopathologic image classification for breast cancer detection. *Health Inf Sci Syst* **6**, 18 (2018). <https://doi.org/10.1007/s13755-018-0057-x>
- [47] T. Kausar, M. Wang and M. S. S. Malik, "Cancer Detection in Breast Histopathology with Convolution Neural Network Based Approach," *2019 IEEE/ACS 16th International Conference on Computer Systems and Applications (AICCSA)*, Abu Dhabi, United Arab Emirates, 2019, pp. 1-5, doi: 10.1109/AICCSA47632.2019.9035244.
- [48] SanaUllah Khan, Naveed Islam, Zahoor Jan, Ikram Ud Din, Joel J. P. C Rodrigues, A novel deep learning based framework for the detection and classification of breast cancer using transfer learning, *Pattern Recognition Letters*, Volume 125, 2019, Pages 1-6, ISSN 0167-8655, <https://doi.org/10.1016/j.patrec.2019.03.022>.
- [49] <https://medium.com/@charlottecullip/a-comparison-of-cnn-architectures-part-2-8d03c67d8ec6>