VISVESVARAYA TECHNOLOGICAL UNIVERSITY

Jnana Sangama, Belagavi – 590 018.

PROJECT PHASE – I REPORT

on

"Breast Cancer Detection System"

Submitted in partial fulfillment of the requirement for the curriculum of the 7th Semester

Bachelor of Engineering

in

Computer Science and Engineering

by

ADITHYA SUNDER	**:**	**1VI18CS002**
NITHIN KUMAR B	**:**	**1VI18CS069**
SANTHOSH B RAO	**:**	**1VI18CS094**
SHASHANK P	**:**	**1VI18CS099**

Under the supervision of

Mrs. Mary Vidya John
Assistant Professor

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING
VEMANA INSTITUTE OF TECHNOLOGY
BENGALURU – 560 034

2021 - 2022

Karnataka ReddyJana Sangha®

VEMANA INSTITUTE OF TECHNOLOGY

Koramangala, Bengaluru-34.

(Affiliated to Visvesvaraya Technological University, Belagavi)

Department of Computer Science and Engineering

Certificate

This is to certify that the project (phase-I) entitled "Breast Cancer Detection System" **is** a bonafide work carried out jointly by **Adithya Sunder (1VI18CS002), Nithin Kumar B (1VI18CS069), Santhosh B Rao (1VI18CS094)** and **Shashank P (1VI18CS099),** during the academic year 2021-22 in partial fulfillment of the requirement for the 7th Semester course work for the **Bachelor of Engineering** in **Computer Science and Engineering** of the **Visvesvaraya Technological University**, **Belagavi.** It is certified that all corrections/suggestions indicated for internal assessment have been incorporated in the report. The project phase - I report has been approved as it satisfies the academic requirements in respect of the project work prescribed for the said degree.

Guide	**HOD**	**Principal**
Mrs. Mary Vidya John	Dr. M. Ramakrishna	Dr. Vijayasimha Reddy. B. G

ACKNOWLEDGEMENT

We sincerely thank **Visvesvaraya Technological University** for providing a platform to do the project phase - I.

Firstly, we would like to express our deep sense of gratitude to our institute "**Vemana Institute of Technology**" that provided us an opportunity to do a project phase - I entitled "**Breast Cancer Detection System**".

We thank **Dr. Vijayasimha Reddy. B. G,** Principal, Vemana Institute of Technology, Bengaluru for providing the necessary support.

We would like to place on record our regards to **Dr. M. Ramakrishna,** Professor and Head, Department of Computer Science and Engineering for his continued support.

We would like to thank our project coordinators **Mrs. A Rosline Mary,** Assistant Professor and **Mrs. Roopalakshmi S,** Assistant Professor, Dept. of CSE for their support and coordination.

We would like to thank our project guide **Mrs. Mary Vidya John,** Assistant Professor, Dept. of CSE for her continuous support and valuable guidance towards successful completion of the project phase – I.

We would be failing in our duty if we do not thank the faculty members, batch mates, lab staffs, technicians and family members for their constant support and guidance.

ADITHYA SUNDER (1VI18CS002)
NITHIN KUMAR B (1VI18CS069)
SANTHOSH B RAO (1VI18CS094)
SHASHANK P (1VI18CS099)

Date:
Place: Bengaluru

ABSTRACT

There have been several empirical studies addressing breast cancer using machine learning and soft computing techniques. Many claim that their algorithms are faster, easier, or more accurate than others are. This study is based on genetic programming and machine learning algorithms that aim to construct a system to accurately differentiate between benign and malignant breast tumors. The aim of this study was to optimize the learning algorithm. In this context, we use artificial neural networks and convolutional neural networks. The performance of the proposed method is based on sensitivity, specificity, precision, accuracy, and the roc curves. The present study proves that genetic programming can automatically find the best model by combining feature preprocessing methods and classifier algorithms. It is alarming that 55% of Indians in 2017 were pushed into poverty due to out-of-pocket medical expenses. Data on quality and accreditation of diagnostic establishments in the country have been described as scanty by many surveys conducted. These statistics are damaging considering the pernicious effects of Covid-19. The pandemic has left millions in disarray and the mounting pressure on the healthcare system isn't helping either. The population has succumbed to the fear of contracting the virus and many people make false assumptions based on their symptoms. Our goal is to get rid of these problems by attacking one major part of healthcare, diagnosis.

Keywords: Breast Cancer, Vital Status Code, Thermal Imaging, ANN, CNN, Cause of Death

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS

(NOTE: in alphabetical order)

ANN	:	Artificial Neural Network
CNN	:	Convolutional Neural Network
COD	:	Cause of Death
SVM	:	Support Vector Machine
TIR	:	Thermal Infrared Ranging
VSC	:	Viral Status Code

CHAPTER 1:

INTRODUCTION

Breast cancer is a prevalent cause of death, and it is the only type of cancer that is widespread among women worldwide. Many imaging techniques have been developed for early detection and treatment of breast cancer and to reduce the number of deaths , and many aided breast cancer diagnosis methods have been used to increase the diagnostic accuracy . In the last few decades, several data mining and machine learning techniques have been developed for breast cancer detection and classification , which can be divided into three main stages: preprocessing, feature extraction, and classification.

To facilitate interpretation and analysis, the preprocessing of mammography films helps improve the visibility of peripheral areas and intensity distribution, and several methods have been reported to assist in this process. Feature extraction is an important step in breast cancer detection because it helps discriminate between benign and malignant tumors. After extraction, image properties such as smoothness, coarseness, depth, and regularity are extracted by segmentation .

Various transform-based texture analysis techniques are applied to convert the image into a new form using the spatial frequency properties of the pixel intensity variations. The common techniques are wavelet transform, fast Fourier transform (FFT) , Gabor transforms , and singular value decomposition (SVD) . To reduce the dimensionality of the feature representation, principal component analysis (PCA) can be applied. Many works have attempted to automate diagnosis of breast cancer based on machine learning algorithms. For example, Malek et al. proposed a method using the wavelet for features extraction and fuzzy logic for classification. Sun et al. studied the problem by comparing features selection methods, whereas Zheng et al. combined K-means algorithm and a support vector machine (SVM) for breast cancer diagnosis.

Several works based on clustering and classification have been conducted. Another approach, introduced by Aličković and Subasi applied a genetic algorithm for feature extraction and rotation forest as a classifier.Feature extraction is an important step in breast cancer detection because it helps discriminate between benign and malignant tumors.

1.1 SCOPE

Screening programs, better treatments and a general increase in awareness have resulted in declining death rates from breast cancer over the past two decades. However, in developed countries, the disease remains the second leading cause of cancer death in women after lung cancer.

The chance of developing invasive breast cancer at some time in a woman's life is approximately 1 in 8, and the chance that breast cancer will be responsible for a woman's death is about 1 in 35.Effective management of this relatively common disease can therefore have a significant impact on survival at the population level, and have a profound effect on lives of those directly affected by the disease and their loved ones.

Breast Cancer Management (ISSN: 1758-1923) addresses key issues in disease management by exploring the best patient-centered clinical research and presenting this information both directly, as clinical findings, and in practice-oriented formats of direct relevance in the clinic. The journal also highlights significant advances in basic and translational research, and places them in context for future therapy.

Breast Cancer Management provides oncologists and other health professionals with the latest findings and opinions on reducing the burden of this widespread disease. Recent research findings and advances clinical practice in the field are reported and analyzed by international experts. The journal presents this information in clear, accessible formats.All articles are subject to independent review by a minimum of two independent experts.

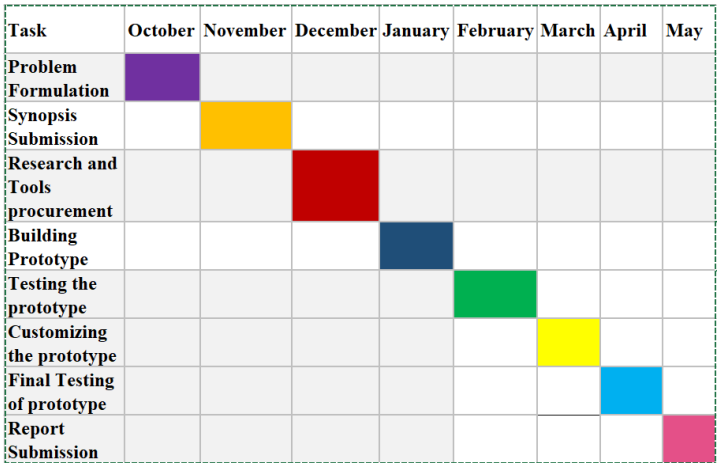
These kinds of diagnostic aids, especially in a diseases like breast cancer where the reason for the occurrence is not totally known, will reduce the false positive diagnosis rate and increase the survival rate among the patients since the early diagnosis of the disease is more curable than in a later stage.

1.2 OBJECTIVES

This analysis aims to observe which features are most helpful in predicting malignant or benign cancer and to see general trends that may aid us in model selection and hyper parameter selection. The goal is to classify whether the breast cancer is benign or malignant. To achieve this I have used machine learning classification methods to fit a function that can predict the discrete class of new input.

1.2.1 Plan of action for the project

Fig 1.1: Timeline Chart



Stage 1 – This is the initial stage of analyzing what is the required in the real world, the problems faced by patients, to draft a problem statement that will be helpful for the hospitals. The team identified the metrics to perform on.

Stage 2 – In this stage, the team identified the software requirements for the project, drafted a plan and collected 4 reference papers to gain information that is required to start the project.

Stage 3- In this stage, the team gained knowledge on neural networks and how to use them. Artificial Neural networks for raw data and Convolutional Networks for imagery.

Stage 4 - The project was proposed to the college and reviewed by the panelists.

Stage 5 – In this stage, we started with the documentation work and drafted the report for project phase-1.

Stage 6- We develop a working prototype and check the initial accuracy.

Stage 7- We customize the prototype based on the results we get during the initial test.

Stage 8- Final testing is done after implementing all the changes

Stage 9- We submit the final report of the project.

1.2.2 Current status of project:

The deadlines set to complete the project phase-1 is end of February and the phase-2 is expected to be completed by June end.

As mentioned earlier, the project phase-1 consists of devising a problem statement, analyzing the requirements, and drafting the report, this is to be successfully completed by February.

The implementation of the application and the main project with outputs matching the expectations is set to be completed by June.

The initial framework is to be completed by the end of February. This gives us more time to focus on the implementation of the main project so that we can get better and accurate results as we planned to spend four months to build the project.

1.2.3 Proposed plan for completion

- **Sourcing and cleaning of datasets**: Storing the medical history of patients and using it as a training set to predict diseases.
- **Use of ANN and CNN:** Use of Artificial Neural Network and Convolutional Neural Network to test the dataset.
- **Training of model against labeled datasets :** Training the neural network based on labeled data
- **Verification of model accuracy and efficiency:** Using a random test case to verify accuracy and efficiency of the model.

1.2.4 Outline of the chapters

Chapter 1: Introduction In this chapter we have a brief introduction to the project. It includes scope of the problem statement and the objectives of the project like plan of action, current status of project and proposed plan for completion of project.

Chapter 2: Literature Survey In this chapter we look at different survey/reference papers that we have considered that is helpful for completion of our project. We have identified the advantages and limitations of each survey paper. This includes a comparative analysis of the papers to understand the how each proposal is different from the other.

Chapter 3: System Requirements In this chapter we have determined all the requirements of the project that include, functional, non-functional, hardware and software requirements.

Chapter 4: Design methodology In this chapter we understand the system architecture and the dataflow diagrams to visualize the workflow.

Chapter 5: Module Description In this chapter we elucidate the working of each module that is built to develop the system.

Chapter 6: Summary In this chapter we have summarized all the work that we have done in the project

CHAPTER 2:

LITERATURE SURVEY

[1]:

Phani Teja Kuruganti, Hairong QiAsymmetry "*Analysis in Breast Cancer Detection Using Thermal Infrared Images*", 2006

Methodology**:** This paper discusses an automated approach for breast cancer detection using Thermal Infrared(TIR) images. Breast cancer is a disease in which only the early diagnosis increases the survival hope. There is use of non-invasive TIR imaging in contrast to the traditional invasive mammography for the early detection. This paper describes a computer-based approach for automating the asymmetry analysis of the thermograms. This kind of approach will help the diagnostics as a useful second opinion.

The use of TIR images for breast cancer detection and the advantages of thermograms over traditional mammograms are discussed. From the experimental results, it can be observed that the Hough transform can be effectively used for breast segmentation. From the presentation of the derived features, the asymmetry can also be effectively identified. In the future with a larger database, supervised pattern classification techniques can be used to attain more accuracy.

These kinds of diagnostic aids, especially in a diseases like breast cancer where the reason for the occurrence is not totally known, will reduce the false positive diagnosis rate and increase the survival rate among the patients since the early diagnosis of the disease is more curable than in a later stage. The cancer cells with their higher metabolic rate are hotter than the normal cells and this property makes the cancerous tumors appear as hotspots in the TIR images. Analysis of the medical images for automatic detection of disease causing abnormality is one of the advanced applications in the fields of image processing and pattern recognition.

Advantage: False positive cases are reduced.
Disadvantage: Chances of activation function are higher, which reduces accuracy.

[2]

Ahmed M. Hassan , Magda El-Shenawee "*Review of Electromagnetic Techniques for Breast Cancer Detection*", 2014

Methodology: Numerous electromagnetic techniques used for detection of disease have been studied and the feasibility has been examined. Techniques like Bio magnetic Detection and Magnetic Resonance Techniques have been studied and clinical records are examined. Each section in this paper provided an up-to-date status of research in seven electromagnetic modalities for breast cancer detection. It is evident that EM techniques show high potential to improve the detection of breast cancer. On the other hand, EM techniques still face major challenges and limitations that need to be overcome before they can be introduced into wide scale utilizations in clinics.

In tegrating more than one EM technique has shown a potential to resolve some of these limitations.Identifying the outer surface of the breast can improve the accuracy and the speed of the consequent steps of imaging the interior of the breast. In addition, the skin layer reflects a considerable portion of the incident microwaves even when a matching liquid between the microwave sources and the breast was utilized. By identifying the outer surface of the breast, the effect of the skin layer can be removed from the signature of the internal structures such as the tumor.

In the algorithm proposed in, the first step involved reconstructing several points on the breast surface equal to the number of antennas used in the imaging. Those points were then interpolated and extrapolated to achieve a larger number of points. Finally, these points were used to generate surface functions to estimate the smooth continuous surface of the breast. Each section in this paper provided an up-to-date status of research in seven electromagnetic modalities for breast cancer detection. It is evident that EM techniques show high potential to improve the detection of breast cancer. On the other hand, EM techniques still face major challenges and limitations.

Advantage: High potential to be best predicted system
Disadvantage: Can't be implemented in a wide scale yet

[3]

Mr. Chintan Shah , Dr. Anjali Jivani "*Comparison of data mining classification algorithms for breast cancer prediction*", 2012

Methodology : Three different data mining classification methods for prediction of breast cancer. Different parameters have been compared to come to this conclusion. The algorithms used are Decision Tree, K-nearest Neighbor and Bayes Classification . Several data mining classification techniques can be applied for the identification and prevention for breast cancer among patients. In this paper, we have used three different data mining classification methods for prediction of breast cancer.

They have compared on different parameters for prediction of cancer. But for superior prediction, we focus on accuracy and lowest computing time. Our studies filtered all algorithms based on lowest computing time and accuracy and we came up with the conclusion that Naïve Bayes is a superior algorithm. A decision tree is a flow-charting like structure, where each internal node denotes a test on an attribute, each branch represents an outcome of the test, and each leaf node holds a class label. This technique separates observation into branches to construct tree on repetition basis.

In most cases, tree classifiers perform classification in two stages: tree-growing and tree-pruning. The tree-growing is top-down approach. In this stage, the tree is split in a recursive manner called recursive partitioning. It is completed when the subset at a node has all the same value of the target variable, or when splitting no longer adds value to the predictions. In the tree-pruning, the tree will be fully-grown, fully-grown tree is cut back to avert over fitting data and this way it improves the correctness of the tree in bottom up manner. This technique is used to improve the estimate and correctness of the algorithm by minimizing the over fitting. Decision tree is widely used in various areas because it is strong enough for data distribution. It is also well suited for developing new machine learning techniques. WEKA comes under the open source software issued under GNU General Public License. A random forest is a collection of unpruned decision tree. Random forest is often used when we have large training data sets and large number of input variables. At the end this method builds many decision trees.

Advantage: Low computation time

Disadvantage: More superior prediction can be achieved. Full potential is not yet reached

[4]

Abdelghani Bellaachia, Erhan Guven "*Predicting Breast Cancer Survivability Using Data Mining Techniques*", 2011

Methodology: An analysis of the prediction of survivability rate of breast cancer patients using data mining techniques. The data used is the SEER Public-Use Data. The preprocessed data set consists of 151,886 records, which have all the available 16 fields from the SEER database. Three data mining techniques investigated are: the Naïve Bayes, the back-propagated neural network, and the C4.5 decision tree algorithms. In this paper they present an analysis of the prediction of survivability rate of breast cancer patients using data mining techniques.

The data used is the SEER Public-Use Data. The preprocessed data set consists of 151,886 records, which have all the available 16 fields from the SEER database. We have investigated three data mining techniques: the Naïve Bayes, the back-propagated neural network, and the C 4.5 decision tree algorithms. Several experiments were conducted using these algorithms. The achieved prediction performances are comparable to existing techniques. However, we found out that C4.5 algorithm has a much better performance than the other two techniques.

The Naïve Bayes technique depends on the famous Bayesian approach following a simple, clear and fast classifier. It has been called 'Naïve' because it assumes mutually independent attributes. In practice, this is almost never true but is achievable by preprocessing the data to remove the dependent categories. This method has been used in many areas to represent, utilize , and learn the probabilistic knowledge and significant results have been achieved in machine learning . The second technique uses artificial neural networks. In this study, a multi-layer network with back-propagation (also known as a multi-layer perceptron) is used.

The third technique is the C4.5 decision-tree generating algorithm. C4.5 is based on the ID3 algorithm. It has been shown that the last two techniques have better performance. This paper has outlined, discussed, and resolved the issues, algorithms, and techniques for the problem of breast cancer survivability prediction in SEER database.

Advantage: Vital Status Recode (VSR) and Cause of Death (COD) is

considered Disadvantage: Analysis doesn't include case of missing data

COMPARITIVE ANALYSIS

Table 2.1: Comparative Analysis

Reference	Algorithm/ Technique	Platform used	Performance Metrics	Advantage	Drawback
[1]	C4.5 decision tree algorithm	Java	Accuracy:86.7%(tested against number of cases positive)	Vital Status Recode (VSR) and Cause of Death (COD) is considered	Analysis does not include records with missing data
[2]	Decision Tree	Java	Accuracy, Specificity, Sensitivity	Low computation time	More superior prediction can be achieved
[3]	EM Algorithm	Java	Accuracy:80-90%(Variety of imaging methods)	High potential to be best predicted system	Can't be implemented in wide scale yet
[4]	Random Forest	Java	Accuracy:85%(Thermograms over Mammograms)	False positive cases are reduced	Chances of activation function is higher due to many options

CHAPTER 3:
SYSTEM REQUIREMENTS

3.1 FUNCTIONAL REQUIREMENTS:

A Functional Requirement is a description of the service that the software must offer. It describes a software system or its component. Requirements are that specifies a function that a system or system component must be able to perform. The functional requirement describes a functionality to be made available to the users of the system, characterizing partially its behavior as an answer to the stimulus that it is subjected to. This type of requirement should not mention any technological issue, that is, ideally functional requirements must be independent of design and implementation aspects.

- The software must be capable of taking inputs that include the clean datasets:

Users are expected to give inputs like, clean datasets for model testing .This is taken by the system and used for further analysis and testing.

- It should be capable of coming to conclusions from the given datasets:

Sometimes the human eye can miss out things that are too minute to be studied. To avoid these mistakes, the model should be accurate and come to conclusions from the data provided to it.

- Perform different operations to the dataset to test the model:

Here we deliberately introduce some inconsistencies into the dataset and find out if the model works accordingly. If it does have comparable accuracy, we try to improve it or we find ways to train the model better.

- Evaluate the operations of the model based on different metrics.
- Output of the project should be a breast cancer report: The series of tests needed to evaluate a possible breast cancer usually begins when a woman or their doctor discover a mass or abnormal calcifications on a screening mammogram, or a lump or nodule in the breast during a clinical or self-examination. Less commonly, a woman might notice a red or swollen breast or a mass or nodule under the arm.

After all the evaluations performed our application should provide the patient a report which elaborates their status of breast cancer. Examination of the tumor under the microscope is used to determine if it is invasive or non-invasive (in situ); ductal, lobular, or another type of breast

cancer; and whether the cancer has spread to the lymph nodes. The report can then be used to seek medical help and treatments at a specialized hospital.

3.2 NON- FUNCTIONAL REQUIREMENTS:

- Reliability: We will provide a reliable service for breast cancer prediction.
- Security : The data will not be compromised as it will be stored on a secure cloud service.
- Performance: We aim to provide a model which is highly accurate.

3.3 SOFTWARE REQUIREMENTS:

- Operating System – Windows 8 or higher
- Platform Used – Jupyter Notebook
- Languages – Python

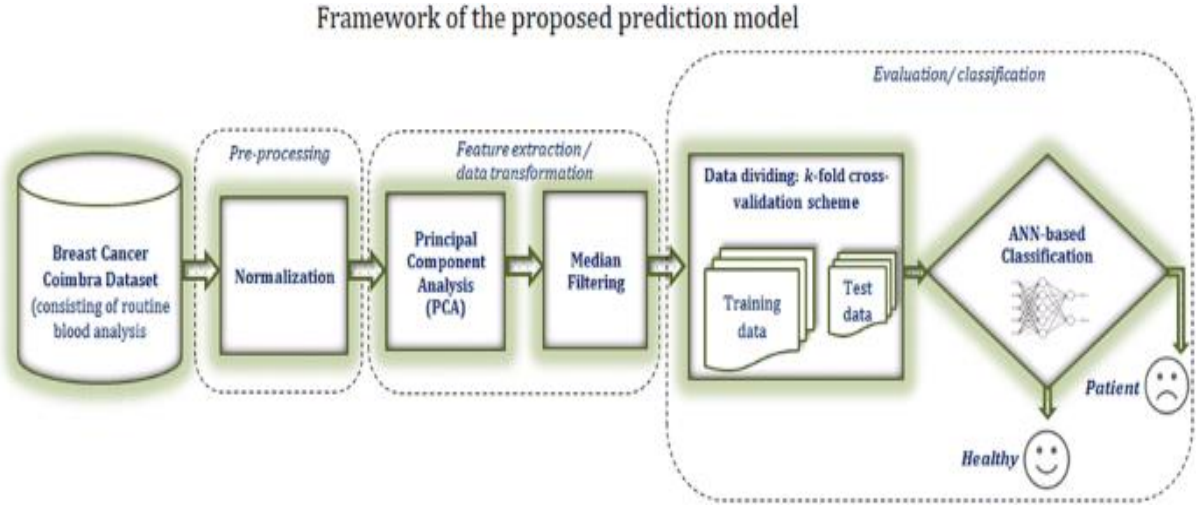
3.4 HARDWARE REQUIREMENTS:

- RAM – 16 GB or higher
- GPU – 1050 or higher
- CPU - Intel i7 processor 10th generation or higher

CHAPTER 4:

DESIGN METHODOLOGY

Fig 4: Design Methodology



As shown in the figure 4, the main modules are:

Pre-processing: Data preprocessing is a process of preparing the raw data and making it suitable for a machine learning model. It is the first and crucial step while creating a machine learning model.When creating a machine learning project, it is not always a case that we come across the clean and formatted data. And while doing any operation with data, it is mandatory to clean it and put in a formatted way. So for this, we use data preprocessing task.A real-world data generally contains noises, missing values, and maybe in an unusable format which cannot be directly used for machine learning models. Data preprocessing is required tasks for cleaning the data and making it suitable for a machine learning model which also increases the accuracy and efficiency of a machine learning model.

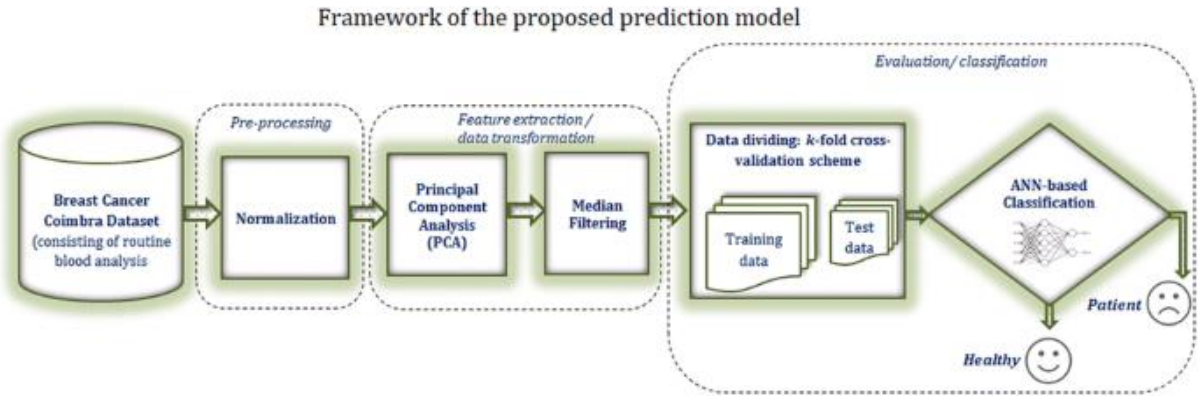
Image segmentation: Image segmentation results in more granular information about the shape of an image and thus an extension of the concept of Object Detection.We segment i.e. divide the images into regions of different colors which helps in distinguishing an object from the other at a finer level

Classification**:** ANN and CNN are used to classify if a person has breast cancer or not. The process of classification is done by taking the flattened weighted feature map obtained from

the final pooling layer, and is used as input to the fully connected network, which calculates the loss and modifies the weights of the internal hidden nodes accordingly.

4.1 SYSTEM ARCHITECTURE

Fig 4.1:System Architecture



As seen in figure 4.1, the architecture is based on:

Dataset: A dataset in machine learning is, quite simply, a collection of data pieces that can be treated by a computer as a single unit for analytic and prediction purposes. This means that the data collected should be made uniform and understandable for a machine that doesn't see data the same way as humans do. For this, after collecting the data, it's important to preprocess it by cleaning and completing it, as well as annotate the data by adding meaningful tags readable by a computer.

Moreover, a good dataset should correspond to certain quality and quantity standards. For smooth and fast training, you should make sure your dataset is relevant and well-balanced. Try to use live data whenever possible and consult with experienced professionals about the volume of the data and the source to collect it from.

Preprocessing: Data preprocessing is a process of preparing the raw data and making it suitable for a machine learning model. It is the first and crucial step while creating a machine learning model.

When creating a machine learning project, it is not always a case that we come across the clean and formatted data. And while doing any operation with data, it is mandatory to clean it and put in a formatted way. So for this, we use data preprocessing task.

A real-world data generally contains noises, missing values, and maybe in an unusable format which cannot be directly used for machine learning models. Data preprocessing is required tasks for cleaning the data and making it suitable for a machine learning model which also increases the accuracy and efficiency of a machine learning model.

It involves below steps:

- Getting the dataset: To create a machine learning model, the first thing we required is a dataset as a machine learning model completely works on data. The collected data for a particular problem in a proper format is known as the dataset.

- Importing libraries: To perform data preprocessing using Python, we need to import some predefined Python libraries. These libraries are used to perform some specific jobs.

- Importing datasets: Now we need to import the datasets which we have collected for our machine learning project. But before importing a dataset, we need to set the current directory as a working directory.

- Finding Missing Data: The next step of data preprocessing is to handle missing data in the datasets. If our dataset contains some missing data, then it may create a huge problem for our machine learning model. Hence it is necessary to handle missing values present in the dataset.

- Encoding Categorical Data: Categorical data is data which has some categories such as, in our dataset; there are two categorical variable, Country, and Purchased. Since machine learning model completely works on mathematics and numbers, but if our dataset would have a categorical variable, then it may create trouble while

building the model. So it is necessary to encode these categorical variables into numbers.

- Splitting dataset into training and test set: In machine learning data preprocessing, we divide our dataset into a training set and test set. This is one of the crucial steps of data preprocessing as by doing this, we can enhance the performance of our machine learning model.

- Feature scaling: Feature scaling is the final step of data preprocessing in machine learning. It is a technique to standardize the independent variables of the dataset in a specific range. In feature scaling, we put our variables in the same range and in the same scale so that no any variable dominate the other variable.

Feature extraction: The goal of feature selection in machine learning is to find the best set of features that allows one to build useful models of studied phenomena.The techniques for feature selection in machine learning can be broadly classified into the following categories:

- Supervised Techniques: These techniques can be used for labeled data, and are used to identify the relevant features for increasing the efficiency of supervised models like classification and regression.

- Unsupervised Techniques: These techniques can be used for unlabeled data.

ANN Based classification: Artificial Neural Networks are a special type of machine learning algorithms that are modeled after the human brain. That is, just like how the neurons in our nervous system are able to learn from the past data, similarly, the ANN is able to learn from the data and provide responses in the form of predictions or classifications.

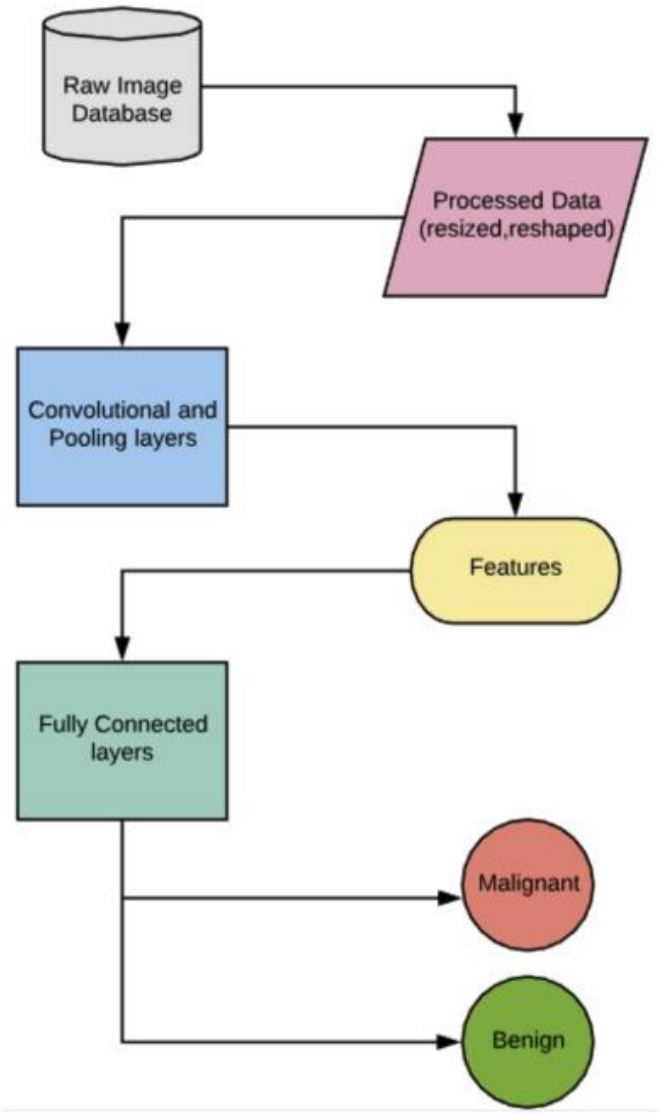
ANNs are nonlinear statistical models which display a complex relationship between the inputs and outputs to discover a new pattern. A variety of tasks such as image recognition,

speech recognition, machine translation as well as medical diagnosis makes use of these artificial neural networks.

An important advantage of ANN is the fact that it learns from the example data sets. Most commonly usage of ANN is that of a random function approximation. With these types of tools, one can have a cost-effective method of arriving at the solutions that define the distribution. ANN is also capable of taking sample data rather than the entire dataset to provide the output result. With ANNs, one can enhance existing data analysis techniques owing to their advanced predictive capabilities.

4.2 DATA FLOW DIAGRAM

Fig 4.2: Data-Flow Diagram



As seen in figure 4.2, the data-flow diagram is based on

Image Processing: Most of the pixels in the image are redundant and do not contribute substantially to the intrinsic information of an image. While dealing with AI networks, it is required to eliminate them to avoid unnecessary computational overhead. This can be achieved by compression techniques. We begin the implementation of our deep net by processing the images in the dataset. This is achieved with the help of the OpenCV library in Python. There are many other modules that can be used in this step e.g. MATLAB or

other image processing libraries or software. This is necessary to remove redundancy from the input data which only contributes to the computational complexity of the network without providing any significant improvements in the result. The aspect ratio of the original slide is preserved since both the dimensions are reduced by a factor of 2, giving an image which is 1/4th in area, that is of dimension 350×230 pixels.

Feature extraction: Feature learning is a crucial step in the classification process for both human and machine algorithm. A study has shown that the human brain is sensitive to shapes, while computers are more sensitive to patterns and texture. Because of this fact, feature learning is entirely different for manual versus machine. In the visual context, malignant tumors tend to have large and irregular nuclei or multiple nuclear structures. The cytoplasm also undergoes changes, wherein new structures appear, or normal structures disappear. Malignant cells have a small cytoplasmic amount, frequently with vacuoles. In this scenario, the ratio of cytoplasm to nucleus decreases .

All of these features are examined by experts, or algorithms are developed to quantify these features to automate detection. This approach is difficult and imprecise as selection and quantification involve various unknown errors that are hard to address. In the case of supervise learning, we do not need to provide these features explicitly. In this case images are fed to an architecture such as CNN, along with its class as a label (Benign or Malignant). From the automatic update of filter values in the training process, CNN is able to extract the computational features. In short, for a given architecture of CNN filters and their weights, are features that are used at the time of testing for model evaluation.

In this approach, CNN takes raw pixels of an image and gives output as learned filter weights. These weights serve input to the dense architecture of the deep neural network for final prediction.

Classification: The process of classification is done by taking the flattened weighted feature map obtained from the final pooling layer, and is used as input to the fully connected network, which calculates the loss and modifies the weights of the internal hidden nodes accordingly.

Benign tumor: Benign (non-cancerous) breast conditions are very common, and most women have them. In fact, most breast changes are benign. Unlike breast cancers, benign breast conditions are not life-threatening. But some are linked with a higher risk of getting breast cancer later on.

Some benign breast changes may cause signs or symptoms (such as breast lumps, pain, or nipple discharge), while others might be found during a mammogram. In either case, sometimes they can be hard to tell apart from breast cancer, so other exams or tests might be needed to find out for sure.

Malignant tumor: malignant tumor that grows in or around the breast tissue, mainly in the milk ducts and glands. A tumor usually starts as a lump or calcium deposit that develops as a result of abnormal cell growth. Most breast lumps are benign but some can be premalignant (may become cancer) or malignant.

Breast cancer is classified as either primary or metastatic. The initial malignant tumor that develops within the breast tissue is known as primary breast cancer. Sometimes, primary breast cancer can also be found when it is spread to lymph nodes that are close by in the armpit. Metastatic breast cancer, or advanced cancer, is formed when cancer cells located in the breast break away and travel to another organ or part of the body.

CHAPTER 5:

MODULE DESCRIPTION

Diagnosis : Module 1 is focused on diagnosis of breast cancer. We use ANN and CNN to test the dataset and find out if the patient has breast cancer. In case the patient has cancer, we can determine which stage they are in for further purposes. ANN is based on a collection of connected units or nodes called artificial neurons, which loosely model the neurons in a biological brain .CNN is used to analyse visual imagery.

Pre-processing: Data preprocessing is a process of preparing the raw data and making it suitable for a machine learning model. It is the first and crucial step while creating a machine learning model.

When creating a machine learning project, it is not always a case that we come across the clean and formatted data. And while doing any operation with data, it is mandatory to clean it and put in a formatted way. So for this, we use data preprocessing task.A real-world data generally contains noises, missing values, and maybe in an unusable format which cannot be directly used for machine learning models. Data preprocessing is required tasks for cleaning the data and making it suitable for a machine learning model which also increases the accuracy and efficiency of a machine learning model.

Image segmentation: Image segmentation results in more granular information about the shape of an image and thus an extension of the concept of Object Detection.We segment i.e. divide the images into regions of different colors which helps in distinguishing an object from the other at a finer level

Classification**:** ANN and CNN are used to classify if a person has breast cancer or not. The process of classification is done by taking the flattened weighted feature map obtained from the final pooling layer, and is used as input to the fully connected network, which calculates the loss and modifies the weights of the internal hidden nodes accordingly.

Treatment : Module 2 is focused on treatment of breast cancer. After finding out which stage the patient is in, we provide appropriate treatment for the patient. Treatments include:

Surgery :Surgery is the removal of the tumor and some surrounding healthy tissue during an operation. Surgery is also used to examine the nearby axillary lymph nodes, which are under the arm. A surgical oncologist is a doctor who specializes in treating cancer with surgery. Learn more about the basics of cancer surgery.Generally, the smaller the tumor, the more surgical options a patient has. The types of surgery for breast cancer include the following:

Lumpectomy: This is the removal of the tumor and a small, cancer-free margin of healthy tissue around the tumor. Most of the breast remains. For invasive cancer, radiation therapy to the remaining breast tissue is often recommended after surgery, especially for younger patients, patients with hormone receptor negative tumors, and patients with larger tumors. For DCIS, radiation therapy after surgery may be an option depending on the patient, the tumor, and the type of surgery. A lumpectomy may also be called breast-conserving surgery, a partial mastectomy, quadrantectomy, or a segmental mastectomy. Women with BRCA1 or BRCA2 gene mutations who have been newly diagnosed with breast cancer may be eligible to receive breast-conserving surgery. So may women with newly diagnosed breast cancer who carry a moderate-risk gene mutation, like CHEK2 or ATM. Your genetic mutation status alone should not determine which course of treatment may be best for you.

Mastectomy. This is the surgical removal of the entire breast. There are several types of mastectomies. Talk with your doctor about whether the skin can be preserved, called a skin-sparing mastectomy, or whether the nipple can be preserved, called a nipple-sparing mastectomy or total skin-sparing mastectomy. A nipple-sparing mastectomy may be a treatment option for certain women with a BRCA1 or BRCA2 gene mutation or for women with a moderate-risk gene mutation, like CHEK2 or ATM. Your doctor will also consider how large the tumor is compared to the size of your breast in determining the best type of surgery.

Lymph node removal, analysis, and treatment:Cancer cells can be found in the axillary lymph nodes in some cancers. Knowing whether any of the lymph nodes near the breast contain cancer can provide useful information to determine treatment and prognosis.Sentinel lymph node biopsy. In a sentinel lymph node biopsy (also called a sentinel node biopsy or SNB), the surgeon finds and removes 1 to 3 or more lymph nodes from under the arm that receive lymph

drainage from the breast. This procedure helps avoid removing a larger number of lymph nodes with an axillary lymph node dissection (see below) for patients whose sentinel lymph nodes are mostly free of cancer. The smaller lymph node procedure helps lower the risk of several possible side effects. Those side effects include swelling of the arm called lymphedema, numbness, and arm movement and range of motion problems with the shoulder. These are long-lasting issues that can severely affect a person's quality of life. Importantly, the risk of lymphedema increases with the number of lymph nodes and lymph vessels that are removed or damaged during cancer treatment.

Doctors may recommend imaging of your lymph nodes with an ultrasound and/or an image-guided biopsy of the lymph nodes before a sentinel lymph node biopsy to find out if the cancer has spread there (see Diagnosis). This is often done if your lymph nodes can be felt during clinical examination or if you are having treatment with chemotherapy before surgery. However, ASCO does not recommend doing this if your cancer is small and your lymph nodes are not able to be felt during clinical examination.

To find the sentinel lymph node, the surgeon usually injects a radioactive tracer and sometimes a dye behind or around the nipple. The injection, which can cause some discomfort, lasts about 15 seconds. The dye or tracer travels to the lymph nodes, arriving at the sentinel node first. If a radioactive tracer is used, it will give off radiation which helps the surgeon find the lymph node. If dye is used, the surgeon can find the lymph node when it turns color.

The pathologist then examines the lymph nodes for cancer cells. If the sentinel lymph node(s) are cancer-free, research has shown that it is likely that the remaining lymph nodes will also be free of cancer. This means that no more lymph nodes need to be removed. In general, for most women with early-stage breast cancer with tumors that can be removed with surgery and whose underarm lymph nodes are not enlarged, sentinel lymph node biopsy is the standard of care. However, in certain situations, it may be appropriate to not undergo any axillary surgery.

Axillary lymph node dissection. In an axillary lymph node dissection, the surgeon removes many lymph nodes from under the arm. These are then examined for cancer cells by a pathologist. The actual number of lymph nodes removed varies from person to person. Women having a lumpectomy and radiation therapy who have a smaller tumor (less than 5 cm) and 2 or less sentinel lymph nodes with cancer may avoid a full axillary lymph node dissection. Also,

some women receiving a mastectomy may avoid an axillary lymph node dissection. This helps reduce the risk of side effects and does not decrease survival. If cancer is found in the sentinel lymph node, whether additional surgery is needed to remove more lymph nodes depends on the specific situation.

Usually, the lymph nodes are not evaluated for people with DCIS and no invasive cancer, since the risk of spread is very low. However, for patients diagnosed with DCIS who choose to have or need a mastectomy, the surgeon may consider a sentinel lymph node biopsy. If some invasive cancer is found with DCIS during the mastectomy, which happens occasionally, the lymph nodes will then need to be evaluated. However, a sentinel lymph node biopsy generally cannot be performed. In that situation, an axillary lymph node dissection may be recommended.

Most people with invasive breast cancer will have either a sentinel lymph node biopsy or an axillary lymph node dissection. For most people younger than 70 with early-stage breast cancer, a sentinel lymph node biopsy will be used to determine if there is cancer in the axillary lymph nodes, since this information is used to make decisions about treatment. For most patients 70 and older with hormone receptor-positive and HER2-negative disease and no clinically apparent cancer in the lymph nodes, ASCO does not recommend sentinel lymph node biopsy. Patients over age 70 with other types of breast cancer or with clinically apparent lymph nodes will generally be recommended to have evaluation of their axillary lymph nodes.No chemotherapy before surgery, and no cancer in the sentinel lymph nodes. For most people in this situation, ASCO does not recommend an axillary lymph node dissection. A small group of patients with tumors located in specific places or with high-risk features may be offered radiation therapy to the lymph nodes.

No chemotherapy before surgery, but there is cancer in the sentinel lymph nodes. For most people in this situation, ASCO recommends radiation therapy instead of axillary lymph node dissection. However, an axillary lymph node dissection may be combined with radiation therapy for people who have 3 or more sentinel lymph nodes containing cancer. For some people in this group, additional radiation therapy to the lymph nodes may be recommended after surgery if the tumors are located in specific places or have high-risk features.

Chemotherapy is given before surgery. Treatment for people who have received chemotherapy before surgery depends on whether the chemotherapy has destroyed the cancer in the lymph nodes. Therefore, after chemotherapy patients are re-staged by sentinel lymph node biopsy. If there was no evidence of cancer in the lymph nodes either before or after chemotherapy , radiation therapy is not recommended . If there was evidence of cancer in the lymph nodes before chemotherapy and there is no longer evidence of cancer in the lymph nodes after chemotherapy, radiation therapy is recommended.If there is evidence of cancer in the lymph nodes after chemotherapy, then both an axillary lymph node dissection and radiation therapy are recommended. Reconstructive (plastic) surgery

Women who have a mastectomy or lumpectomy may want to consider breast reconstruction. This is surgery to recreate a breast using either tissue taken from another part of the body or synthetic implants. Reconstruction is usually performed by a plastic surgeon. A person may be able to have reconstruction at the same time as the mastectomy, called immediate reconstruction. They may also have it at some point in the future, called delayed reconstruction.For patients having a lumpectomy, reconstruction may be done at the same time to improve the look of the breast and to make both breasts look similar. This is called oncoplastic surgery. Many breast surgeons can do this without the help of a plastic surgeon at the same time as the lumpectomy. Surgery on the healthy breast at the same time as the lumpectomy may also be suggested so both breasts have a similar appearance.

The techniques discussed below are typically used to shape a new breast implants. A breast implant uses saline-filled or silicone gel-filled forms to reshape the breast. The outside of a saline-filled implant is made up of silicone, and it is filled with sterile saline, which is salt water . Silicone gel-filled implants are filled with silicone instead of saline . They were thought to cause connective tissue disorders, but clear evidence of this has not been found. Before having permanent implants, a woman may temporarily have a tissue expander placed that will create the correct sized pocket for the implant . Implants can be placed above or below the pectoralis muscle. Talk with your doctor about the benefits and risks of silicone versus saline implants. The lifespan of an implant depends on the woman. However, some women never need to have them replaced. Other important factors to consider when choosing implants include:

Saline implants sometimes "ripple" at the top or shift with time, but many women do not find it bothersome enough to replace.Saline implants tend to feel different than silicone implants. They are often firmer to the touch than silicone implants.There can be problems with breast implants. Some women have problems with the shape or appearance. The implants can rupture or break, cause pain and scar tissue around the implant, or get infected. Implants have also been rarely linked to other types of cancer, including a type called breast implant-associated anaplastic large cell lymphoma (BIA-ALCL). Although these problems are very unusual, talk with your doctor about the risks.

Tissue flap procedures. These techniques use muscle and tissue from elsewhere in the body to reshape the breast. Tissue flap surgery may be done with a “pedicle flap,” which means tissue from the back or belly is moved to the chest without cutting the blood vessels. A “free flap” means the blood vessels are cut and the surgeon needs to attach the moved tissue to new blood vessels in the chest. There are several flap procedures:Transverse rectus abdominis muscle (TRAM) flap. This method, which can be done as a pedicle flap or free flap, uses muscle and tissue from the lower stomach wall.Latissimus dorsi flap. This pedicle flap method uses muscle and tissue from the upper back. Implants are often inserted during this flap procedure.Deep inferior epigastric artery perforator (DIEP) flap. The DIEP free flap takes tissue from the abdomen and the surgeon attaches the blood vessels to the chest wall.External breast forms (prostheses)

An external breast prosthesis or artificial breast form provides an option for women who plan to delay or not have reconstructive surgery. These can be made of silicone or soft material, and they fit into a mastectomy bra. Breast prostheses can be made to provide a good fit and natural appearance for each woman.

Precautions : Module 3 is focused on precautions for breast cancer. Some patients may test falsely positive after a scan. Even benign tumours can end up getting cancerous. The precautions are:

Limit alcohol: Reasons why alcohol consumption may lead to breast cancer include:Alcohol is empty calories and can lead to unwanted weight gain. Excess fat can lead to increased cancer risk. Alcohol can increase levels of estrogen and other hormones associated with breast cancer. Alcohol users are more likely to have increased amounts of folic acid in their systems, which

can lead to increased cancer risk.Men should also limit their drinking, but not because of breast cancer risk. While men can develop breast cancer, alcohol consumption doesn't really increase their risk for breast cancer.

Stay physically active: Findings from observational studies provide much evidence for a link between higher levels of physical activity and lower risk of cancer. However, these studies cannot fully rule out the possibility that active people have lower cancer risk because they engage in other healthy lifestyle behaviors. For this reason, clinical trials that randomly assign participants to exercise interventions provide the strongest evidence because they eliminate bias caused by pre-existing illness and attendant physical inactivity. To confirm the observational evidence and define the potential magnitude of the effect, several large clinical trials are examining physical activity and/or exercise interventions in cancer patients and survivors.

Stop nicotine consumption: Tobacco use is the most preventable cause of all types of cancers, most notably cancers in your lungs, mouth, throat, voice box, and esophagus. The list also includes breast cancer. Cigarette smoke contains toxins, including cancer-causing chemicals. Women who smoke or used to smoke are more likely to get breast cancer than those who don't or never did. Smoking also raises your chances of dying of breast cancer after your diagnosis. And it makes the cancer more likely to come back. It's never too late to stop smoking or give up smokeless tobacco, such as chewing tobacco. If you quit or cut back right after your diagnosis, it'll lessen your chances for lung-related or breathing issues like lung cancer and heart disease.

CHAPTER 6:

SUMMARY

We aim to come up with a model which predicts breast cancer in patients with high accuracy and high prediction speed. We will consider all metrics to come up with an all round development of the product. The product we aim to develop has potential to be used on a wide scale and help millions if not billions around the globe. Our sole aim was to make a project that helps the society and breast cancer is one of the problems that is slowly rising day by day. We need to address this issue as soon as possible and provide cost effective care and early diagnosis to patients.

CONCLUSION AND FUTURE WORK

Breast cancer is a prevalent cause of death, and it is the only type of cancer that is widespread among women worldwide. Many imaging techniques have been developed for early detection and treatment of breast cancer and to reduce the number of deaths , and many aided breast cancer diagnosis methods have been used to increase the diagnostic accuracy . In the last few decades, several data mining and machine learning techniques have been developed for breast cancer detection and classification , which can be divided into three main stages: preprocessing, feature extraction, and classification.

To facilitate interpretation and analysis, the preprocessing of mammography films helps improve the visibility of peripheral areas and intensity distribution, and several methods have been reported to assist in this process. Feature extraction is an important step in breast cancer detection because it helps discriminate between benign and malignant tumors. After extraction, image properties such as smoothness, coarseness, depth, and regularity are extracted by segmentation . Various transform-based texture analysis techniques are applied to convert the image into a new form using the spatial frequency properties of the pixel intensity variations.

The common techniques are wavelet transform, fast Fourier transform (FFT) , Gabor transforms , and singular value decomposition (SVD) . To reduce the dimensionality of the feature representation, principal component analysis (PCA) can be applied. Many works have attempted to automate diagnosis of breast cancer based on machine learning algorithms. For example, Malek et al. proposed a method using the wavelet for features extraction and fuzzy logic for classification. Sun et al. studied the problem by comparing features selection methods, whereas Zheng et al. combined K-means algorithm and a support vector machine (SVM) for breast cancer diagnosis. Several works based on clustering and classification have been conducted. Another approach, introduced by Aličković and Subasi applied a genetic algorithm for feature extraction and rotation forest as a classifier. Market research can be done to cater to the needs of the patients and hospitals as we are still behind on infrastructure and facilities for diagnosis. Future work also includes analysis of cases and finding trends with the results. There is a lot of information that can be gained by finding trends to have a targeted approach on this issue.

REFERENCES

[1] Phani Teja Kuruganti , Hairong QiAsymmetry "*Analysis in Breast Cancer Detection Using Thermal Infrared Images*", 2006

[2] Ahmed M. Hassan, Magda El-Shenawee "*Review of Electromagnetic Techniques for Breast Cancer Detection*", 2014

[3] Mr. Chintan Shah , Dr. Anjali Jivani "*Comparison of data mining classification algorithms for breast cancer prediction*", 2012

[4] Abdelghani Bellaachia, Erhan Guven "*Predicting Breast Cancer Survivability Using Data Mining Techniques*", 2011