BREAST CANCER DETECTION

A MINI PROJECT REPORT

18CSC305J - ARTIFICIAL INTELLIGENCE

Submitted by

VAIBHAV JHA [RA2011027010187]
PUNEET KUMAR RAI [RA2011027010196]
SIVANESH G [RA2011027010177]

Under the guidance of

Dr. PREMALATHA G.

Assistant Professor, Department of Computer Science and Engineering

in partial fulfillment for the award of the degree of

BACHELOR OF TECHNOLOGY

in

COMPUTER SCIENCE & ENGINEERING

of

FACULTY OF ENGINEERING AND TECHNOLOGY

SRM
INSTITUTE OF SCIENCE & TECHNOLOGY
Deemed to be University u/s 3 of UGC Act, 1956

S.R.M. Nagar, Kattankulathur, Chengalpattu District

MAY 2023

BONAFIDE CERTIFICATE

Certified that Mini project report titled **“BREAST CANCER DETECTION”** is the bonafide work of “**VAIBHAV JHA (RA2011027010187), PUNEET KUMAR RAI(RA2011027010196), SIVANESH G(RA2011027010177)”** who carried out the minor project under my supervision. Certified further, that to the best of my knowledge, the work reported herein does not form any other project report or dissertation on the basis of which a degree or award was conferred on an earlier occasion on this or any other candidate.

SIGNATURE

Dr. PREMALATHA G.
GUIDE
Assistant Professor
Department of Data Science and Business System

SIGNATURE

Dr. M. Lakshmi
HEAD OF THE DEPARTMENT
Professor & Head
Department of Data Science and Business System

ABSTRACT

Breast cancer has become the major cause of the increase in mortality rate among women. The causes of breast cancer can be abnormal growth in the breast cell, dividing the cell faster than usual, and gathering of cells to form a mass. This abnormal growth will affect the cells in the nearby tissue also. The causes of breast cancer can be a family history of breast cancer, age factor, exposure to radiation, beginning period at an earlier age, menopause at an older age, having first pregnancy at an older age, drinking alcohol, etc. The symptoms of breast cancer can be the formation of lumps, discoloration, pain, change in size or shape, etc. This paper discusses the various algorithms and methods available in the literature survey for the detection of breast cancer in detail. The indication of cancer is the presence of masses, the presence of calcium deposits in the breast tissue which is seen as a bright spot in the mammogram, change in the shape of the breast. At present detection of breast cancer is done by examining the affected cell under the microscope by an expert and trained pathologist. Since the human intervention is present this method is highly prone to error, so the manual method automatic method of breast cancer classification came into existence. The feature-extract-based automatic breast cancer classification involves pre-processing, segmentation, feature extraction, classification, and validation of the proposed algorithm.

Keywords: *Classification, Data, Segmentation, and Feature extraction.*

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

Chapter 1: Introduction

1.1 Overview

One of the most common cancers in women is breast cancer. In this type of cancer, the breast cells proliferate erratically.

With an expected 2.3 million new cases (11.7%), female breast cancer has surpassed lung cancer as the most frequently diagnosed malignancy worldwide. 90% of breast cancer instances are caused by lifestyle factors, while 10% are genetic. In 15 female PBCRs, there was a discernible rise in the incidence rates of breast cancer. The majority of patients had multimodal treatment, and 97.7% of the tumours were epithelial. Asia's highest incidence of breast cancer was found in Israel (84.6). The district with the highest incidence rate in India was Hyderabad (48.0). Cancer cases are predicted to rise from 13.9 lahks in 2020 to 15.7 lakh in 2025, assuming a 20 percent increase overall, per a research released by the National Cancer Registry Programme (NCRP) [3]. If common malignancies are treated early, death can be avoided. Early detection of bret cancer can result in successful therapy. The research's objective is to more accurately categorise the patients' tumour kinds into benign and malignant ones by using classification algorithms. The Kaggle website provided the dataset. The machine will predict the value for the dependent variable for a given input in the form of an independent variable after the learning process is complete. This technique is known as supervised learning and is a machine learning concept.

The classification techniques used for detecting the tumour are Decision tree, K-Nearest Neighbour (KNN), Support Vector Machine (SVM), Naïve Bayes (NB) classification in Jupyter Notebook along with data visualization.

1.2 PROBLEM DEFINITION AND PURPOSE

One of the leading causes of cancer-related deaths globally is breast cancer. Early diagnosis greatly increases the likelihood of receiving the proper care and surviving, however, this procedure is time-consuming and frequently causes conflict between pathologists. Systems for computer-aided diagnostics had the potential to increase diagnostic precision. However, early detection and prevention can greatly lower the risk of death. Finding breast cancer as soon as possible is crucial. Early and correct diagnosis of the increasing number of breast cancer victims using the already present data can help in saving a large number of lives as they can be furnished with proper treatment timely. Being aware of such a health problem and to be able to investigate more about it, to apply the acquired knowledge over it , it is wise to design a helping aid to detect the same which is an act of humanity in a way.

1.2.1 Client Identification/Need Identification:

Life is valuable and prone to threats. Breast cancer is one of the life damaging threat . With the advent of machine learning in medical field and it has been a powerful helping hand in the decision making process of medical practitioners. With the growing years there is an unfortunate tremendous increase in breast cancer cases which has furnished us with a large set of data for clinical and medical research.

The project makes the use of various machine learning algorithms to detect breast cancer in order to have a purely correct detection of the disease as wrong diagnosis can lead to an incorrect treatment.

1.3 SCOPE OF THE PROJECT

The project takes into consideration ,a large number of datasets of various women and their cell details followed by if they were diagnosed with breast cancer or not.

After the detailed consideration of taking in different datasets and their outcomes, the project comes up with the prediction of breast cancer of a particular person by taking the entire cell details and then by applying machine learning algorithms to it.

The project turns out to be a lifesaver by predicting breast cancer at an appropriate time which provides a helping hand in saving numerous lives.

1.4 IDENTIFICATION OF NEED

There are different major tasks to implement this research –

1. Data Collection and Exploration:

   To collect and read data from various datasets and perform Exploratory data analysis.

2. Model Design:

   A model which is efficient and accurate enough to predict cancer is to be designed under this section.

3. Integration and eployment:

   To integrate the designed ML model and deploy it.

1.5. ORGANIZATION OF REPORT

In Chapter 1, We discussed the need of our project. We tried to identify the problemthat needs a solution. Further we discussed different tasks we are going to perform to make this research relevant and explored different datasets.

In Chapter 2, We will discuss the literature review, wherein we will define the problem and propose solutions. We will also discuss the goals of this research paper and its scope in the future ahead.

In Chapter 3, We will collect raw data to achieve our goal using proper steps. Firstly, we have to analyze the data thoroughly before the arrangement of the data and will create a preliminary design for the project using different models.

In Chapter 4, We will analyze the accuracy of different models used and will use a suitable model out of all.

In Chapter 5, We will integrate and deploy the final outcome of the prediction model Along the description of the future scope.

Chapter 2:Literature Review

2.1 Introduction

Breast cancer is among the 4 leading cancers in women worldwide i.e., lung, breast, and bowel [including, anus], stomach, and prostate cancers). The IARC statistics show that breast cancer accounts for 25% of all cancer cases diagnosed in women worldwide. Around 53% of these cases come from developing countries, which represent 82% of the world’s population. It is reported that 626,700 deaths will occur only in 2018. Breast cancer is the leading cause of cancer death among women in developing countries and the second leading cause of cancer death (following lung cancer) among women in developed countries.

The World Health Organization (WHO) agencies for cancer research (i.e, the International agency for cancer research (IARC) and the American Cancer Society) report that 17.1 million new cancer cases are recorded in 2018 worldwide. WHO estimates that cancer incidences might increase to 27.5 million by 2040, with an estimated 16.3 million deaths expected as a result of cancer.

2.2 Bibliometric Analysis:

Breast cancer is leading cancer in females all over the world. Breast cancer is caused due to the abnormal growth of some cells in the breast. Machine Learning is a field of Artificial Intelligence that uses statistical techniques broadly utilized in bioinformatics and especially in breast cancer growth conclusion. Many researchers work on predicting the earlier way of a breast cancer diagnosis. These are the most widely used algorithm used in breast cancer prediction:

1. **<u>K- Nearest Neighbor: -</u>**

   KNN is a lazy model because it does not learn anything during the training phase and learns in the testing phase. It is instance-based learning. It is non-parametric learning which memorizes the resultant of classifying unseen data. It is used for classification and regression algorithms. The output of the classification is in form of 1, -1and 0. This algorithm is used for pattern recognition and intrusion detection. It takes more time to compute the result so it is less efficient than the others.

2. **<u>Support Vector Machine: -</u>**

   Support Vector Machines belong to the class of supervised learning systems. It is one of the best optimization procedures. This reduces the over-flowing of the trained data. It works by choosing basic examples from all classes referred to as help vectors and isolating the classes by creating a linear function that partitions them as comprehensively as conceivable utilizing these help vectors. In this way, it is regularly said that planning between an input vector to a high dimensionality space is framed utilizing a Support Vector Machine that intends to search out the preeminent reasonable hyperplane that separates the data set into classes.

3. **<u>Random Forest: -</u>**

   Random forests (RF) [20] is one of the most successful ensemble learning techniques which have been proven to be very popular and powerful techniques in pattern recognition and machine learning for high-dimensional classification [21] and skewed problems [20]. These studies used RF to construct a collection of individual decision tree classifiers which utilized the classification and regression trees (CART) algorithms [22]. Many research studies applied the random forests algorithm to construct decision trees.

4. **Artificial Neural Networks: -**

It is the form of Neural Networks. A biologically oriented network is formed to predict breast cancer. It is a computational model that mimics the way nerve cells work in the human brain. It uses learning algorithms (CNN, SNN, etc) that can independently make adjustments as they receive inputs and is a very effective tool for non-linear statistical data modeling. In this ANN accuracy can be as high as 87%.

5. **Naïve Bayes: -**

It is a statically and probabilistic classifier, which is based on Bayes Theorem. Every feature of the attribute is independent of each other attribute. It is a classification technique that was designed to classify high-dimensional datasets. It is suitable for binary and multiclass classification. Naïve Bayes performs well in cases of categorical input variables compared to numerical variables. It is useful for making predictions and forecasting data based on historical results.

2.3 Review of an existing application:

The following is a summary of the existing works by some famous researchers:

Initially, the authors compared the performance of various classification algorithms. The classification algorithms were performed on eight NCD datasets using eight classification algorithms and a 10-fold cross-validation method. These were evaluated using AUC as an indicator of accuracy. To be more specific, breast cancer was detected using hybridization of the guided ABC and neural networks.
Then, According to their study, the majority of NNs have shown promise in detecting tumor cells. However, the imaging approach requires a high computational capacity to pre-process the images.

Also, the authors reviewed different machine learning, deep learning, and data mining algorithms related to breast cancer prediction. Several research papers on breast cancer were reviewed, with a total of 27 papers in machine learning, 4 papers in ensemble techniques, and papers in deep learning techniques. After studying these surveys, our contribution will involve studying genetic sequencing and imaging at the same time to predict breast cancer and to get more information that can help with early diagnosis and treatment of breast cancer.

2.4 Problem Definition

Breast cancer is one of the main causes of cancer death worldwide. Early diagnostics significantly increases the chances of correct treatment and survival, but this process is tedious and often leads to a disagreement between pathologists. Computer-aided diagnosis systems showed the potential for improving diagnostic accuracy. But early prevention can significantly reduce the chances of death.

2.5 Goals/Objectives

It is necessary to improve the screening process in order to reduce the percentage of the female population that is not covered by screening programs and increase the number of early-detected breast cancers. The improvement of the screening program may be reflected in the following: more efficient determination of the list of the women who have to undergo preventive examination, the introduction of the screening program in thermography as a diagnostic method applied in the pre-screening stage, more efficient analysis of mammograms and continuous follow up of patients.

The goal is to increase the proportion of breast cancers identified at an early stage, allowing for more effective treatment to be used and reducing the risks of death from breast cancer. Since early detection of cancer is key to the effective treatment of breast cancer, we use various machine learning algorithms to predict if a tumor is benign or malignant, based on the features provided by the data.

Chapter 3: Design Flow/Process

System design is the solution to the creation of a new system. This phase is composed of several systems. This phase focuses on the detailed implementation of the feasible system. It emphasizes on translating design specifications to performance specifications. System design has two phases of development logical and physical design.

During the logical design phase, the analyst describes inputs (sources), outputs (destinations), databases (data sores) and procedures (data flows) all in a format that meats the uses requirements. The analyst also specifies the user needs and at a level that virtually determines the information flow into and out of the system and the data resources. Here the logical design is done through data flow diagrams and database design.

The physical design is followed by physical design or coding. Physical design produces the working system by defining the design specifications, which tell the programmers exactly what the candidate system must do. The programmers write the necessary programs that accept input from the user, perform necessary processing on accepted data through call and produce the required report on a hard copy or display it on the screen

3.1 Gantt Chart

It is also known as a Bar chart is used exclusively for scheduling purposes. It is a project-controlling technique. It is used for scheduling. Budgeting and resourcing planning. A Gantt is a bar chart with each bar representing activity. The bars are drawn against a timeline. The length of time planned for the activity. The Gantt chart in the figure shows the grey part is the slack time which is the latest by which a task has been finished.

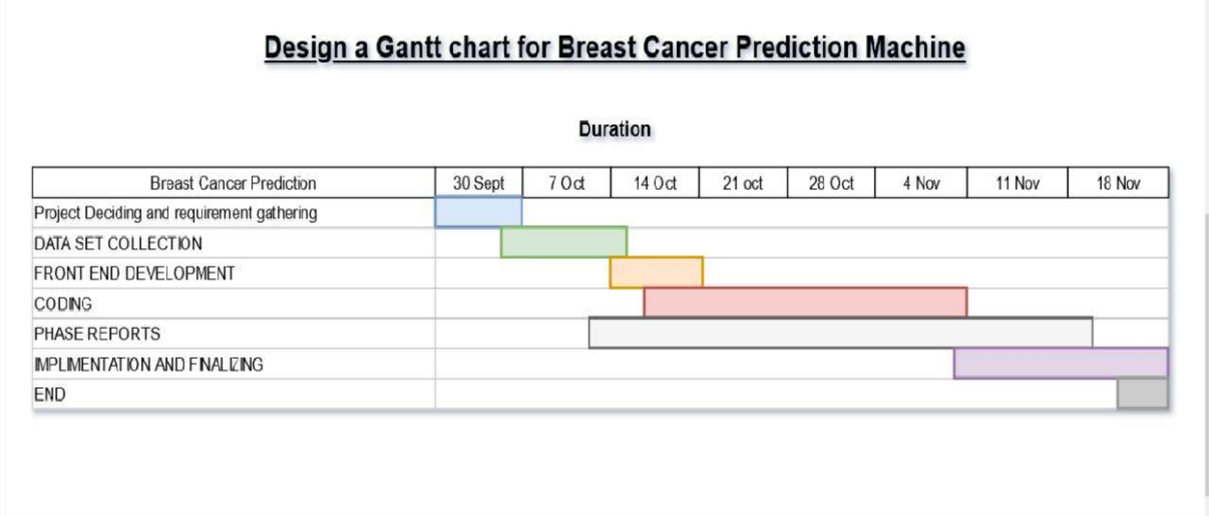


Fig. 3.1: Gantt Chart

3.2 Dataflow Diagram

Data flow diagram is the starting point of the design phase that functionally decomposes the requirements specification. A DFD consists of a series of bubbles joined by lines. The bubbles represent data transformation and the lines represent data flows in the system. A DFD describes what data flow rather than how they are processed, so it does not hardware, software, and data structure, A data-flow diagram (DFD) is a graphical representation of the "flow" of data through an information system. DFDs can also be used for the visualization of data processing. The data flow diagram is a graphical description of a system's data and how to process and transform the data is known as Data Flow Diagram (DFD).

Unlike details flow chart DFDs don't supply detailed descriptions of models that graphically describe a system's data and how the data interact with the system.

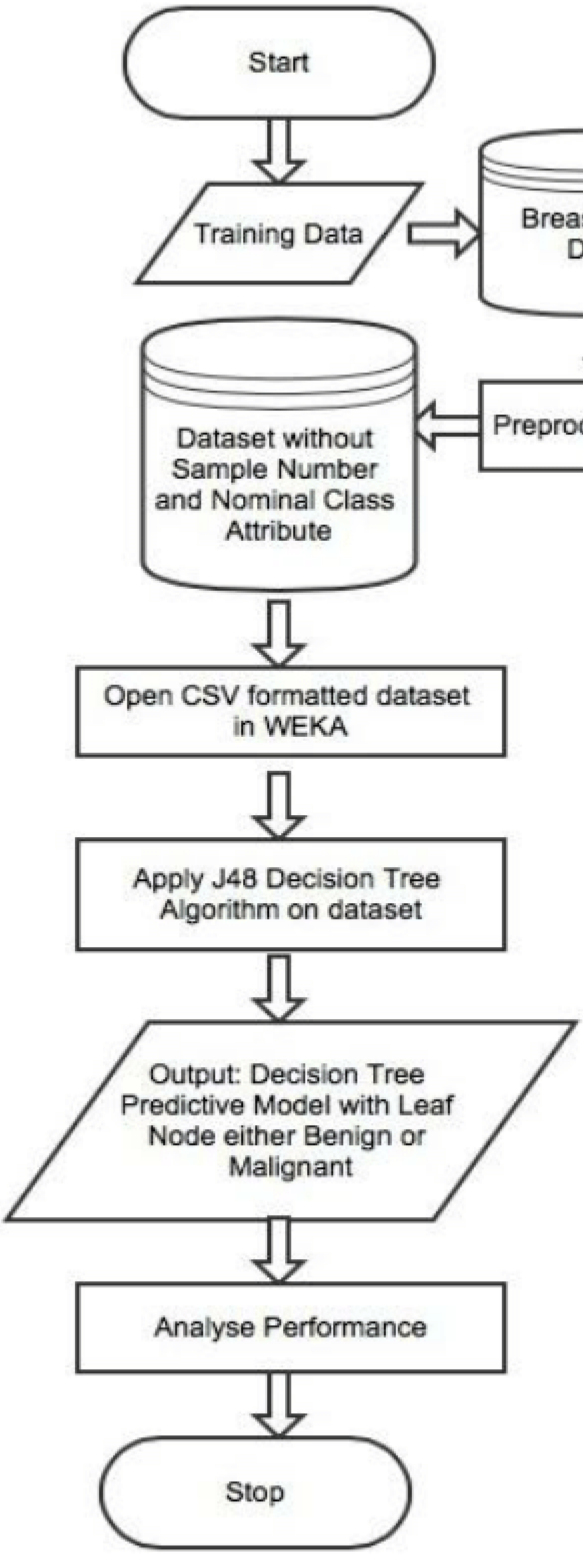

Start
Training Data
Breast Cancer Dataset
Preprocess Data
Dataset without Sample Number and Nominal Class Attribute
Open CSV formatted dataset in WEKA
Apply J48 Decision Tree Algorithm on dataset
Output: Decision Tree Predictive Model with Leaf Node either Benign or Malignant
Analyse Performance
Stop

Fig. 3.2: Data flow diagram

3.3 ER Diagram:

The Entity-Relationship (ER) model was originally proposed by Peter in 1976 [Chen76] as a way to unify the network and relational database views. Simply stated the ER model is a conceptual data model that views the real world as entities and relationships. A basic component of the model is the Entity-Relationship diagram which is used to visually represent data objects. Since Chen wrote his paper the model has been extended and today it is commonly used for database design for the database designer, the utility of the ER model is:

- It maps well to the relational model. The constructs used in the ER model can easily be transformed into relational tables.
- It is simple and easy to understand with a minimum of training. Therefore, the model can be used by the database designer to communicate the design to the end user.
- In addition, the model can be used as a design plan by the database developer to implement a data model in specific database management software.

3.3.1 ER Notation

There is no standard for representing data objects in ER diagrams. Each modelling methodology uses its own notation. The original notation used by Chen is widely used in academic texts and journals but rarely seen in either CASE tools or publications by non-academics. Today, there are a number of notations used; among the more common are Bachman, crow's foot, and IDEFIX.

All notational styles represent entities as rectangular boxes and relationships as lines connecting boxes. Each style uses a special set of symbols to represent the cardinality of a connection. The notation used in this document is from Martin. The symbols used for the basic ER constructs are:

- Entities are represented by labeled rectangles. The label is the name of the entity. Entity names should be singular nouns.
- Relationships are represented by a solid line connecting two entities. The name of the relationship is written above the line. Relationship names should be verbs.
- Attributes, when included, are listed inside the entity rectangle. Attributes that are identifiers are underlined. Attribute names should be singular nouns.
- Cardinality of many is represented by a line ending in a crow's foot. If the crow's foot is omitted, the cardinality is one. Existence is represented by placing a circle or a perpendicular bar on the line. Mandatory existence is shown by the bar (looks like a 1) next to the entity for an instance is required. Optional existence is shown by placing a circle next to the entity that is optional.

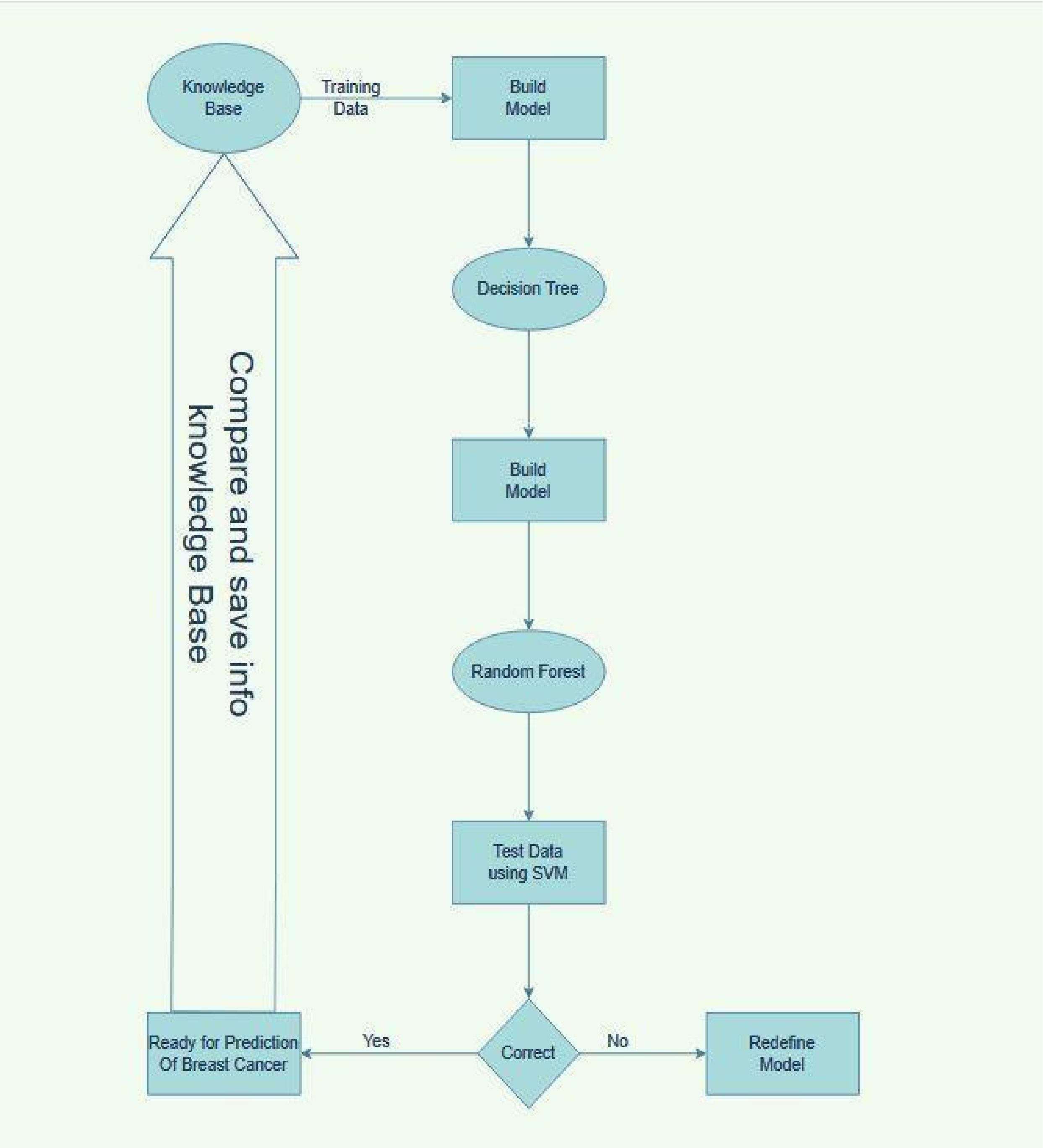


Fig. 3.3: Entity-Relationship Diagram

3.4 UML Class Diagram

In software engineering, a **class diagram** in the Unified Modelling Language (UML) is a type of static structure diagram that describes the structure of a system by showing the system's classes, their attributes, operations (or methods), and the relationships among objects.

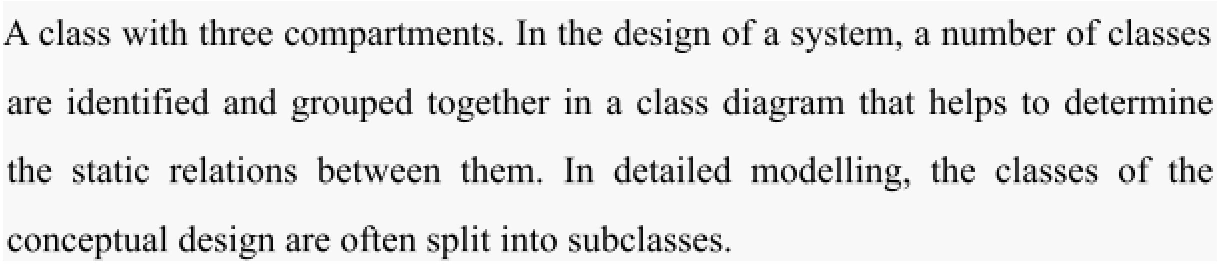
The class diagram is the main building block of object-oriented modelling. It is used for general conceptual modelling of the structure of the application, and for detailed modelling, translating the models into programming code. Class diagrams can also be used for data modelling.[1] The classes in a class diagram represent both the main elements, interactions in the application, and the classes to be programmed.

In the diagram, classes are represented with boxes that contain three compartments:

The top compartment contains the name of the class. It is printed in bold and centred, and the first letter is capitalized.

The middle compartment contains the attributes of the class. They are left-aligned and the first letter is lowercase.

The bottom compartment contains the operations the class can execute. They are also left-aligned and the first letter is lowercase.



A class with three compartments. In the design of a system, a number of classes are identified and grouped together in a class diagram that helps to determine the static relations between them. In detailed modelling, the classes of the conceptual design are often split into subclasses.

Chapter 4: RESULT ANALYSIS AND VALIDATION

4.1. Analysis and Feature Finalization subject to constraints

After going through the evaluation and selection of features and the constraints to that, we came up with the final decisions of what we are going to add as the features, so the final features are as follows:

1. Takes the entire cell details through different columns.
2. Aesthetic and user-friendly appearance and interface.
3. Predicts cancer just by a click after taking all the cell details.
4. Makes efficient use of different machine learning algorithms.

- **METHODOLOGY:-**

4.1.1. Data Set Description

Wisconsin Breast Cancer Dataset repository provided the numerical data set and is composed of FNA (fine needle aspirate) of breast weight/mass. The data was to extract characteristics of cell nuclei of scanned images, and a people of 569 have issues with the treatment. A count of 212 have got a 97% accuracy of malignancy and the other half of benign. The classes are further divided in 4 groups corresponding to benign and malignant case.

4.1.2. Data Pre-Processing

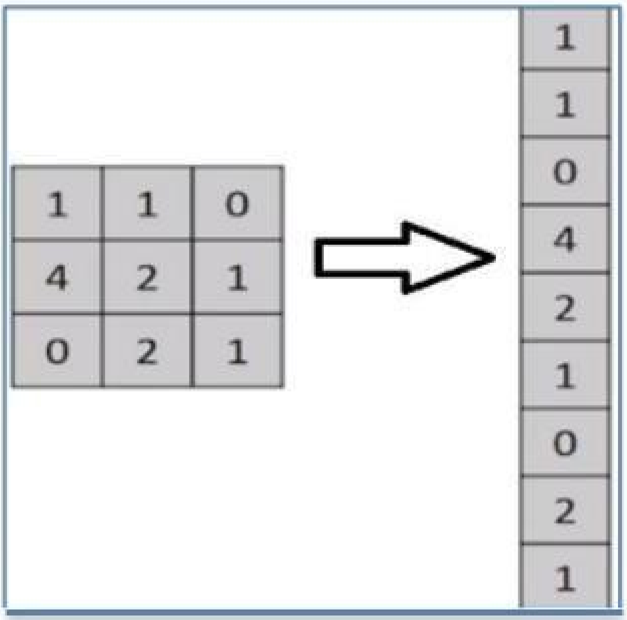
Pre-processing is at an unstructured data which will resize into undesirable data to form structured dataset. The traits are considered to be transferred to mean value. The data is then circulated properly.

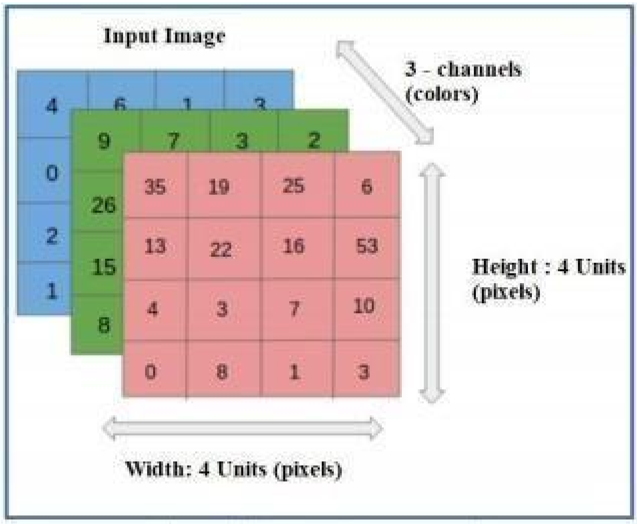
4.1.3. Training and Testing Phase

This phase have the features of dataset whereas testing will return new updated data which will further examined to check if every algorithm is working from its initial stage to give an as much possible accurate prediction. The approach will remain in the testing folds.

4.1.4. Proposed convolution neural network for image dataset analysis

The approach procedure is divided into three system, the first starts with data generation, then analyzation and the prediction. Data will be taken in the form of MRI, X-Rays through IOT devices. The photographs or sensors data for numerical data will be recorded in a database. Medical specialist with help of data analyst forecast the cancer detection by mainly using CNN algorithm. It is necessary to understand detailed levels made in each layer of the process and to get a data in form of single linear vector classification.





The artificial neural association's core design consists of a large number of interconnected neurons grouped in three layers: input, concealment, and yield.

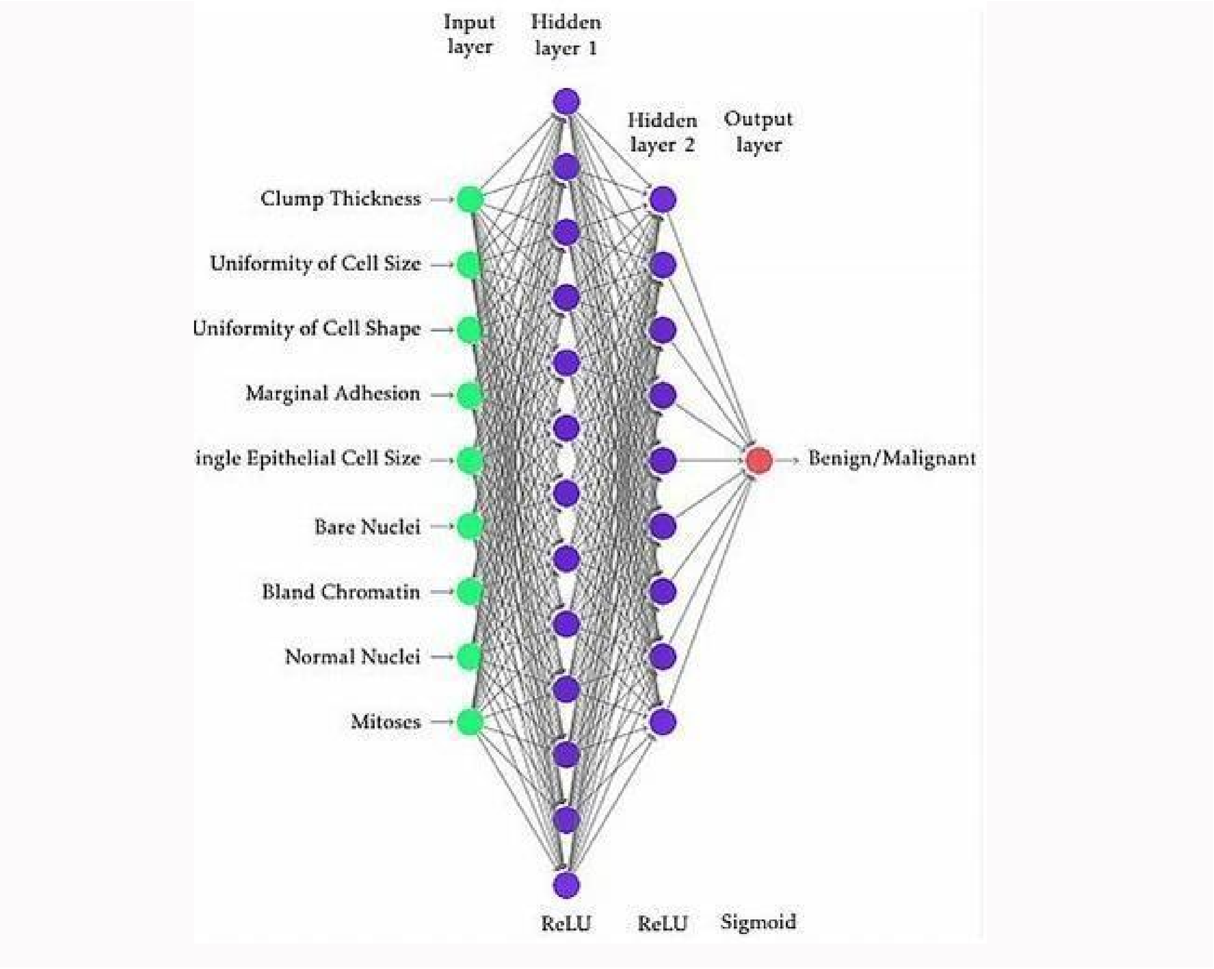
$$\text{Output } = b_i + \sum_{j=1}^{n_x} w_{ij} x_i$$

$$\text{Rectified Linear Unit: Activation}(x) = \begin{cases} 0, & \text{for } x \leq 0 \\ x, & \text{for } x > 0 \end{cases}$$

$$\text{Leaky Rectified Linear Unit: Activation}(x) = \begin{cases} 0.01x, & \text{for } x < 0 \\ x, & \text{for } x \geq 0 \end{cases}$$

$$\text{Softmax: Activation}(x) = \frac{e^{x_i}}{\sum_{j=1}^{J} e^{x_j}}, \text{ where } i = 1, 2, \ldots, j$$

The image of system architecture of ANN as follows:-



Chest data is significantly suitable for analysing evidenced explanation. The proposed CNN model and suggested computation achieved the most essential precision for the dataset, which is greater than the other current algorithms.

4.2. Implementation

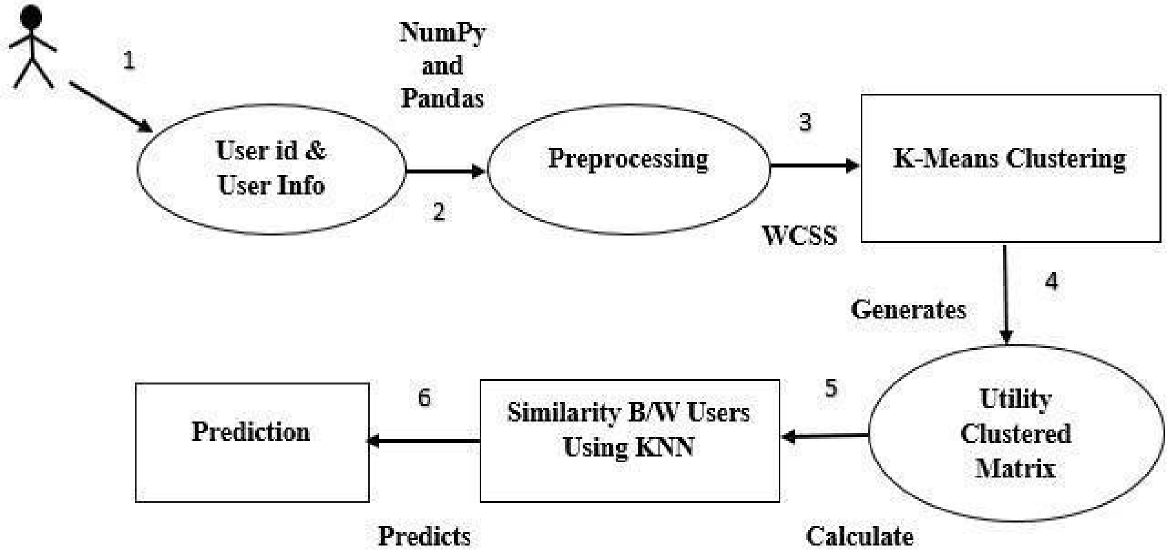


Fig. 4.1: Process Flow Diagram

.

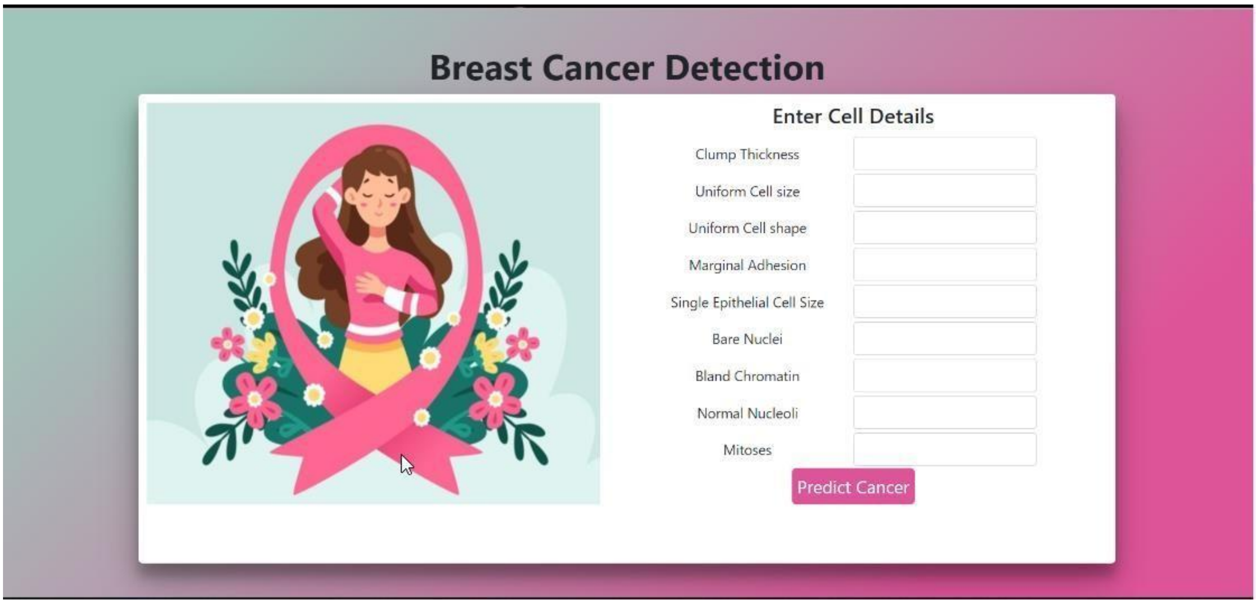


Fig. 4.2: Front end

4.3 Result analysis

During analysis, data collected on various files, decision points and transactions handled by the present system. The commonly used tools are data flow diagram, interviews etc. Training, experience, and common sense are required for the collection of relevant information needed to develop the system. The success of the system depends largely on how clearly the problem is defined, thoroughly investigated, and properly carried out through the choice of solution.

4.4 Project Management

4.4.1. OVERALL PLANNING

We have managed our overall project as described in the following table:

Sr. No	Activity	Description	Period
1.	Planning	In this phase we have plan out what the modules of my project & how to make those modules.	1 weeks
2.	Requirement	In this phase we have noted down all requirement for my project.	1 weeks
3.	Analysis	In this phase we have analyse the old system & solve those limitation into my software.	1 weeks
4.	Database Design	In this phase we have designed my	2 weeks

		database tables for my project.	
5.	Form Design & Report Design	In this phase we have designed all the project forms & project reports.	1 weeks
6.	Coding	In this phase we have started my hard coding of my project	5 weeks
7.	Test & Implementation	After our coding phase is completed, we started with testing of my software. After software works successful, we have implemented that software on the system of an organization.	1 weeks

Table 4.3.1: Milestones

4.3.1. TIME MANAGEMENT FOR CODING PART

At the time of planning, we have divided the time phase of the project where the team can work independently on the project modules. In this we have fractionated our project in four modules. Utmost we have focused on our main module i.e. creating a website which shows best user interface, in this we have designed and exported the data from database. Eventually, we systematize our different modules at one place to give requisite outcomes.

Module	**Description**	**Time taken**
Module A	Created the design for the website.	8-10 days
Module B	Coded the frontend part and breast cancer prediction.	15-20 days
Module C	Added data for the website.	7-9 days
Module D	Trained the machine. Integrated every module to make it one.	5-6 days

Table 4.3.2: Life span of project

Every project has a beginning, a middle period during which activities move the project toward completion, and an ending (either successful or unsuccessful). A standard project typically has the following four major phases (each with its own agenda of tasks and issues): initiation, planning, implementation, and closure. Let's reconnoiter our forethought:

- **Conceptualization Phase:** In this stage, we have identified the basic needs of the users by getting input from all stakeholders, including customers, etc. We have analysed and defined the necessity of the users. In this, we have concentrated on, the requirement has to be precise like what kind of operations will be done, how it will be done, what will be the risk in it, etc. Like, there is no particular app which provides users with everything (hotels, spots, dining, etc) on one platform. We have to navigate through different websites or contact manually for the same.

- **Planning Phase:** In this stage, we have planned our website's interface, how it looks, what functionality it possesses. The team determines the cost and resources required for implementing the analysed requirements. We also detailed the risks involved and provided sub-plans for softening those risks. According to our Website's requirements, we analysed which software will be appropriate and fulfil our needs. In this we scrutinize that Jupyter will be best suited IDE and Flask will meet the needs.
- **Execution Phase:** At this stage, we have implemented our ideas through our Breast Cancer Prediction . After designing the interfaces, we came across the code generation. If the design is performed in a detailed and organized manner, code generation can be accomplished without much hassle. High level programming language i.e. HTML, CSS and JAVASCRIPT are used for coding frontend. The programming language is chosen with respect to the type of website being developed. We have created a prototype of the website to take the significant remarks from the users, where product defects are reported, tracked, fixed and retested.
- **Termination Phase:** We aim that once the team has completed all the tasks, and the supervisor signs off that all deliverables are complete, the project is closed. Any documentation is handed over to the core-supervisor and if required to an ongoing maintenance organization.

Chapter 5: CONCLUSION

Breast Cancer prediction system predicts breast cancer by taking the entire cell details and information including clump thickness, uniform cell site, uniform cell shape , marginal adhesion , single epithelial cell size, bare nuclei ,, bland chromatin, normal nucleoli, mitosis by making use of machine learning algorithms and datasets which proves to be of great significance in saving lives of a large number of women who are a victim of the disease. The project is just a humanitarian act by making wise use of technology.

The proposed machine-learning approaches could predict breast cancer as the early detection of this disease could help slow down the progress of the disease and reduce the mortality rate through appropriate therapeutic interventions at the right time. Applying different machine learning approaches, accessibility to bigger datasets from different institutions (multi-center study), and considering key features from a variety of relevant data sources could improve the performance of modeling.

At the end it is concluded that we have made effort on following points:

We can notice that SVM takes about 0.07 s to build its model unlike K-NN that takes just 0.01 s. It can be explained by the fact that k-NN is a lazy learner and does not do much during training process unlike others classifiers that build the models. In other hand, the accuracy obtained by SVM (97.13%) is better than the accuracy obtained by C4.5, Naïve Bayes and k-NN that have an accuracy that varies between 95.12 % and 95.28 %. It can also be easily seen that SVM has the highest value of correctly classified instances and the lower value of incorrectly classified instances than the other classifiers. After creating the predicted model, we can now analyse results obtained in evaluating efficiency of our algorithms. SVM and C4.5 got the highest value (97 %) of TP for benign class but k-NN correctly predicts 97% of instance that belong to malignant class. The FP rate is lower when using SVM classifiers (0.03 for benign class and 0.02 for malignant class), and then other algorithms follow: k-NN, C4.5 and NB. From these results, we can understand why SVM has outperformed other classifiers In summary, SVM was able to show its power in terms of effectiveness and efficiency based on accuracy and recall.

5.1 Future Work

In a nutshell, it can be summarized that the future scope of the project circles around maintaining information regarding:

- We can add user friendly sites in future.
- We can give more advance software for Breast Cancer Prediction including more facilities.
- We will host the platform on online servers to make it accessible worldwide.
- Integrate multiple load balancers to distribute the loads of the system.
- Create the master and stave database structure to reduce the overload of the database queries.
- Implement the backup mechanism for taking backup of codebase and database on regular basis on different servers.

The future work will focus on exploring more of the dataset values and yielding more interesting outcomes. This will be accompanied by a more effective and reliable disease prediction which will contribute towards better healthcare system by reducing overall cost, time and mortality rate. The analysis of the results signifies that the integration of multidimensional data along with different classification, feature selection and dimensionality reduction techniques can provide auspicious tools for inference in this domain Further research in this field should be carried out for the better performance of the classification techniques so that it can predict on more variables. We are intending how to parametrize our classification techniques hence to achieve high accuracy. We are looking into many datasets and how further Machine Learning algorithms can be used to characterize Breast Cancer. We want to reduce the error rates with maximum accuracy.

References

[1] "What is Breast Cancer? Symptoms, Signs, Types & Stages." Breastcancer.org, 2021.

[2] Ranganathan, Padmapriya, and Usha Rani Poli. "Statistical Analysis of Breast Cancer in India." BioSpectrum India, 5 Oct. 2020.

[3] Boser, Bernhard E., Isabelle M. Guyon, and Vladimir N. Vapnik. "Support Vector Machines: Theory and Applications." ResearchGate, 1992.

[4] Elter, Matthias, et al. "Risk Factors for Breast Cancer: A Systematic Review of Studies with Female Samples among the General Population." Journal of Women's Health, vol. 15, no. 7, 2006.

[5] Sarma, Kandarpa Kumar, et al. "Diagnosis of breast cancer using decision tree and naive Bayes classifier." Procedia Computer Science, vol. 132, 2018.

[6] Odetayo, Moses O., et al. "A Review of Artificial Intelligence Techniques for Breast Cancer Detection Using Mammography." IEEE Access, vol. 1, 2013.

[7] Zhang, Yuan-Ting, et al. "Breast Cancer Detection Using Neural Networks." IEEE Transactions on Biomedical Engineering, vol. 44, no. 12, 1997.

[8] Kannan, Ramasamy, et al. "A Survey on the Role of Artificial Intelligence in Diagnosis and Treatment of Breast Cancer." Computers, vol. 11, no. 9, 2022.