Project Report

on

BREAST CANCER DETECTION USING MACHINE LEARNING

Submitted in partial fulfillment of the

requirement for the award of the degree of

B.Tech CSE

(Established under Galgotias University Uttar Pradesh Act No. 14 of 2011)

Under The Supervision of
S. Kalidass
(Asst. Professor)

Submitted By
VIBHAV YADAV 19SCSE1010039 (19021011248)
VIKAS PRASAD 19SCSE1010040 (19021011249)

SCHOOL OF COMPUTING SCIENCE AND ENGINEERING
DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING
GALGOTIAS UNIVERSITY, GREATER NOIDA
INDIA
DECEMBER,2021

SCHOOL OF COMPUTING SCIENCE AND ENGINEERING GALGOTIAS UNIVERSITY, GREATER NOIDA

CANDIDATE'S DECLARATION

I/We hereby certify that the work which is being presented in the project report, entitled **"BREAST CANCER DETECTION USING MACHINE LEARNING"** in partial fulfillment of the requirements for the award of the **B.TECH CSE**submitted in the School of Computing Science and Engineering of Galgotias University, Greater Noida, is an original work carried out during the 19July 2021- 25Dec 2021, under the supervision of S. Kalidass (Asst. Professor), Department of Computer Science and Engineering , Galgotias University, Greater Noida

The matter presented in the project report has not been submitted by me/us for the award of any other degree of this or any other places.

VIBHAV YADAV 19SCSE1010039
VIKAS PRASAD 19SCSE1010040

This is to certify that the above statement made by the candidates is correct to the best of my knowledge.

S.Kalidass
Asst Professor

CERTIFICATE

The Final project Viva-Voce examination of VIBHAV YADAV 19SCSE1010039 & VIKAS PRASAD 19SCSE1010040 has been held on _____ and his/her work is recommended for the award of **B.TECH CSE**.

Signature of Examiner(s) **Signature of Supervisor(s)**

Signature of Project Coordinator **Signature of Dean**

Date: December, 2021
Place: Greater Noida

Acknowledgement

The satisfaction and the euphoria that accompany the successful completion of any task would be incomplete without the mention of the people who made it possible. The constant guidance of these persons and encouragement provided, crowned my efforts with success and glory. I take this opportunity to express my gratitude to one and all. I am grateful to management and my institute Galgotias University with its very ideals and inspiration for having provided me with the facilities, which made this, project a success. I am indebted with a deep sense of gratitude for the constant inspiration, encouragement, timely guidance, and valid suggestion given to me by my advisor S.Kalidass Sir and my project member, Department of Computer Science, Galgotias University.

Abstract

Breast Cancer is one of the leading cancer developed in many countries including India. Though the endurance rate is high – with early diagnosis 97% women can survive for more than 5 years. Statistically, the death toll due to this disease has increased drastically in last few decades. The main issue pertaining to its cure is early recognition. Hence, apart from medicinal solutions some Data Science solution needs to be integrated for resolving the death causing issue.

Naïve Bayes Classifier produces ill results when the training data is not represented. The SVM classifier is unsuitable for large datasets and also not effective on high computer vision applications. When the data is imbalanced, Bi-clustering and Ada boost Techniques will lead to erroneous classification. RCNN takes more time to train the network. With these limitations, the proposed methodology is introduced.

Data Science solution needs to be integrated for resolving the death causing issue. 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.

There are several benefits to applying the feature selection (Machine Learning Algorithm) methods: it (a) is effective and faster in training the machine learning algorithm, (b) reduces the complexity of a model and makes it easier to interpret, (c) improves the accuracy of a model if the right subset is chosen, and (d) reduces overfitting.

In this project in python, we learned to build a breast cancer tumor predictor on the sklearn dataset and created graphs and results for the same. There are many treatments for a patient based on breast cancer stage; data mining and machine learning can be a very good help in deciding the line of treatment to be followed by extracting knowledge from such suitable databases.

Contents

List of Table

List of Figures

CHAPTER-1 Introduction

1.1 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.

Nowadays, the demand for machine learning is growing until it becomes a service. Unfortunately, machine learning is still a field with high barriers and often requires expert knowledge. Designing an effective machine learning model including the stages of preprocessing, feature selection, and classification processes requires a set of skills and expertise.

CHAPTER-2 Literature Survey

In recent years, several studies have applied data mining algorithms on different medical datasets to classify Breast Cancer. These algorithms show good classification results and encourage many researchers to apply these kind of algorithms to solve challenging tasks. In a convolutional neural network (CNN) was used to predict and classify the invasive ductal carcinoma in breast histology images with an accuracy of almost 88%. Moreover, data mining is used widely in medical fields to predict and classify abnormal events to create a better understanding of any incurable diseases such as cancer. The outcomes of using data mining in classification are promising for breast cancer detection. Therefore, data mining approach is used in this work.

- A list of some literature studies related to this method is presented in Table 1:-

Paper title	**Datasets**	**Algorithms**	**Results**
Integration of data mining classification techniques and ensemble learning for predicting the type of breast cancer recurrence, 2019	Breast Cancer	NB, SVM, GRNN and J48	GRNN & J48 accuracy: 91% NB & SVM: 89%
A study on prediction of breast cancer recurrence using data mining techniques, 2017	WPBC	Classification: KNN, SVM, NB and C5.0, Clustering: K-means, EM, PAM and Fuzzy c-means	Classification accuracy is better than clustering, SVM & C5.0: 81%

Predicting breast cancer recurrence using effective classification and feature selection technique, 2016	WPBM	NB, C4.5, SVM	NB: 67.17%, C4.5: 73.73%, SVM: 75.75%
Using machine learning algorithms for breast cancer risk prediction and diagnosis, 2016	WBC	SVM, C4.5, NB, KNN	SVM outperform others: 97.13%
Study and analysis of breast cancer cell detection using Naïve Bayes, SVM and ensemble algorithms, 2016	WDBC	NB, SVM, Ensemble	SVM: 98.5%, NB & Ensemble: 97.3%

Analysis of Wisconsin breast cancer dataset and machine learning for breast cancer detection, 2015	WDBC	NB, J48	NB: 97.51%, J48: 96.5%
Comparative study on different classification techniques for breast cancer dataset, 2014	Breast Cancer	J48, MLP, rough set	J48: 79.97%, MLP: 75.35%, rough set: 71.36%
A novel approach for breast cancer detection using data mining techniques, 2014	WBC	SMO, IBK, BF Tree	SMO: 96.19%, IBK: 95.90%, BF Tree: 95.46%

Experiment comparison of classification for breast cancer diagnosis, 2012	WBC WDBC WPBC	J48, SMO, MLP, NB, IBK	In WBC: MLP & J48: 97.2818%. In WDBC: SMO: 97.7% or fusion on SMO & MLP: 97.7% In WPBC: fusion of MLP, J48, SMO and IBK: 77%

These cells divide more briskly and disperse faster than healthy cells do and continue to accumulate, form- ing a lump or mass that the may start causing pain. Cells may spread rapidly through your breast to your lymph nodes or to other parts of your body. Some women can be at a higher risk for breast cancer

because of their family history, lifestyle, obesity, radiation, and reproductive factors. In the case of cancer, if the diagnosis occurs quickly, the patient can be saved as there have been advances in cancer treatment. In this study we use four machine learning classifiers which are Naive Bayesian Classifier, k-Nearest Neighbour, Support Vector Machine, Artificial Neural Network and random forest.

Harmonic imaging and real-time compounding has been shown to enhance image resolution and lesion characterisation. More recently, USG elastography seems to be quite encouraging. Initial results show that it can improve the specificity and positive predictive value of USG within the characterisation of breast masses. The reason why any lesion is visible on mammography or USG is that the relative difference within the density and acoustic resistance of the lesion, respectively, as compared to the encompassing breast tissue. [1]

In this paper dierent machine learning algorithms are used for detection of Breast Cancer Prediction. Decision tree, random forest, support vector machine, neural network, linear model, adabost, naive bayes methods are used for prediction.

An ensemble method is used to increase the prediction accuracy of breast cancer. New technique is implemented which is GA based weighted average ensemble method of classification dataset which over- came the limitations of the classical weighted average method. Genetic algorithm based weighted average method is used for the prediction of multiple models. The comparison between Particle swarm optimisation(PSO), Dierential evolution(DE) and Genetic algorithm(GA) and it is concluded that the genetic algorithm outperforms for weighted average methods. One more comparison between classical ensemble method and GA based

weighted average method and it is concluded that GA based weighted average method outperforms. [2]

On Breast Cancer Detection: An Application of Machine Learn ing Algorithms on the Wisconsin Diagnostic Dataset by the Abien Fred M. Agarap. In this paper, six machine learning algorithms are used for detection of cancer . GRU- SVMmodel is used for the diagnosis of breast cancer GRU- SVM, Linear Regression, Multilayer Perceptron (MLP), Nearest Neighbour (NN) search, Softmax Regression, and Support Vector Ma- chine (SVM) on the Wisconsin Diagnostic Breast Cancer (WDBC) dataset by measuring their classification test accuracy, and their sensitivity and specificity values. The said dataset consists of features which were computed from digitised images of FNA tests on a breast mass. For the implementation of the ML algorithms, the dataset was partitioned in the following fashion 70 percent for training phase, and 30 percent for the testing phase Their results were that all presented ML algorithms exhibited high performance on the binary classification of carcinoma, i.e. determining whether benign tumour or malignant tumour. Therefore, the statistical measures on the classification problem were also satisfactory. To further corroborate the results of this study, a CV technique such as k-fold cross-validation should be used. The appliance of such a way won't only provide a more accurate measure of model prediction performance, but it'll also assist in determining the foremost optimal hyper-parameters for the ML algorithms. [3]

Analysis of Machine Learning Techniques for Breast

In this paper, ML techniques are explored in order to boost the accuracy of diagnosis. Methods such as CART, Random Forest, K-Nearest Neighbours are compared. The dataset used is acquired from UC Irvine Machine Learning Repository. It is found that KNN algorithm has much better performance than the other techniques used in comparison. The most accurate model was K-Nearest Neighbour. The classification model such as Random Forest and Boosted Trees showed the similar accuracy. Therefore, the most accurate classier can be used to detect the tumour so that the cure can be found in early stage. [4]

During this paper, four dierent machine learning algorithms are used for the early detection of carcinoma. The aim of this project is to process the results of routine blood analysis with dierent ML methods. Methods used are Artificial Neural Network (ANN), Extreme Learning Machine (ELM), Support Vector Machine (SVM) and Nearest Neighbour (k-NN). Dataset is taken from the UCI library. In this dataset age, BMI, glucose, insulin, homeostasis model assessment (HOMA),

leptin, adiponectin, resistin, and chemokine monocyte chemoattractant protein (MCP1) attributes were used. Parameters that have the best accuracy values were found by using four dierent Machine Learning techniques. This dataset includes age, BMI, glucose, insulin, HOMA, leptin, adiponectin, resistin and MCP1 features that can be acquired in routine blood analysis. The significance of these data in breast cancer detection was investigated by ML methods. The analysis was performed with four dierent ML methods. k-NN and SVM methods are determined using Hyperparameter optimization technique. The highest accuracy and lowest training time were given by ELM which was 80%. and 0.42 seconds. [5]

The study firstly collects the data of the BCCD dataset which contains 116 volunteers with 9 attributes and data of WBCD dataset which contains 699 volunteers and 11 attributes. Then we preprocesses the raw data of WBCD dataset and obtained the info that contains 683 volunteers with nine attributes and therefore the index indicating whether the volunteer has the malignant tumour. After comparing the accuracy, F- measure metric and ROC curve of 5 classification models, the result has shown that RF is chosen as the primary classification model during this study. Therefore, the results of this study provide a reference for experts to distinguish the character of carcinoma .In this study, there are still some limitations that ought to be solved in further work. For instance, though there also exist some indices people haven't found yet, this study only collects the info of 10 attributes during this experiment. The limited data has an impact on the accuracy of results. additionally , the RF can also be combined with other data mining technologies to get more accurate and efficient results in the longer term work physicians

Twenty-four recent research articles have been reviewed to explore the computational methods to predict breast cancer. Chaurasia et al. developed prediction models of benign and malignant breast cancer. Wisconsin breast cancer data set was used. The dataset contained 699 instances, two classes (malignant and benign), and nine integer- valued clinical attributes such as uniformity of cell size. The researchers removed the 16 instances with missing values from the data set to become the data set of 683 instances. The benign were 458 (65.5%) and malignant were 241 (34.5%). The experiment was analyzed by the Waikato Environment for Knowledge Analysis (WEKA). Naive Bayes, RBF Network, and J48 are the three most popular data mining algorithms were used to develop the prediction models. The researchers used 10- fold cross-validation methods to measure the unbiased estimate of the three prediction

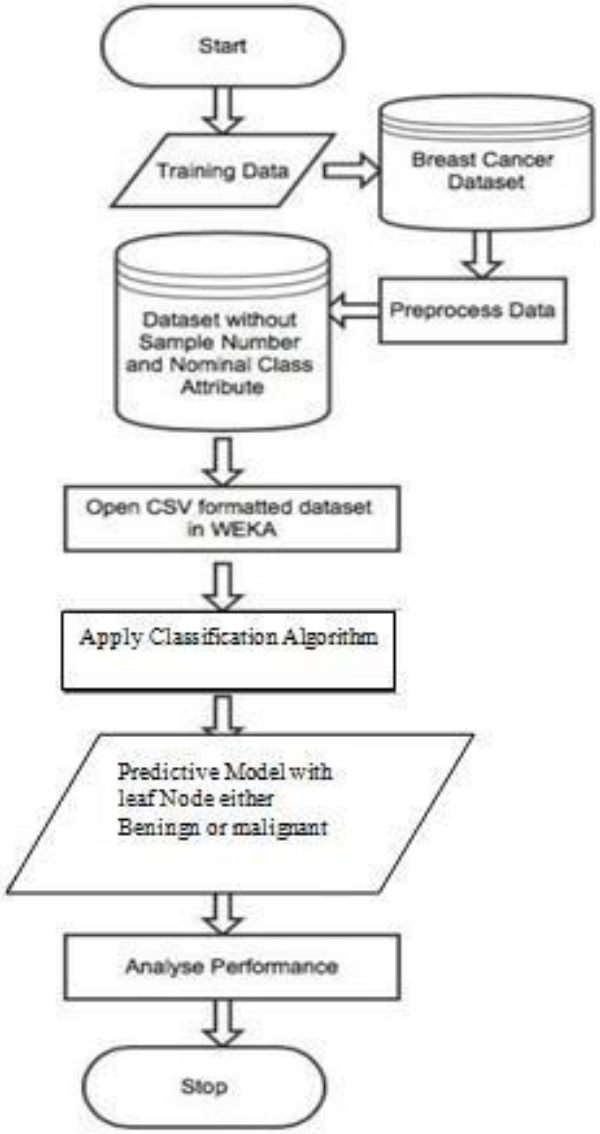
models for performance comparison purposes. The models' performance evaluation was presented based on the methods‟ effectiveness and accuracy. Experimental results showed that the Naive Bayes had gained the best performance with a classification accuracy of 97.36%; followed by RBF Network with a classification accuracy of 96.77% and the J48 was the third with a classification accuracy of 93.41%. In addition, the researchers conducted sensitivity analysis and specificity analysis of the three algorithms to gain insight into the relative contribution of the independent variables to predict survival. The sensitivity results indicated that the prognosis factor Class was by far the most important predictor.

Table 2. Performance analysis of most popular BC detection methods.

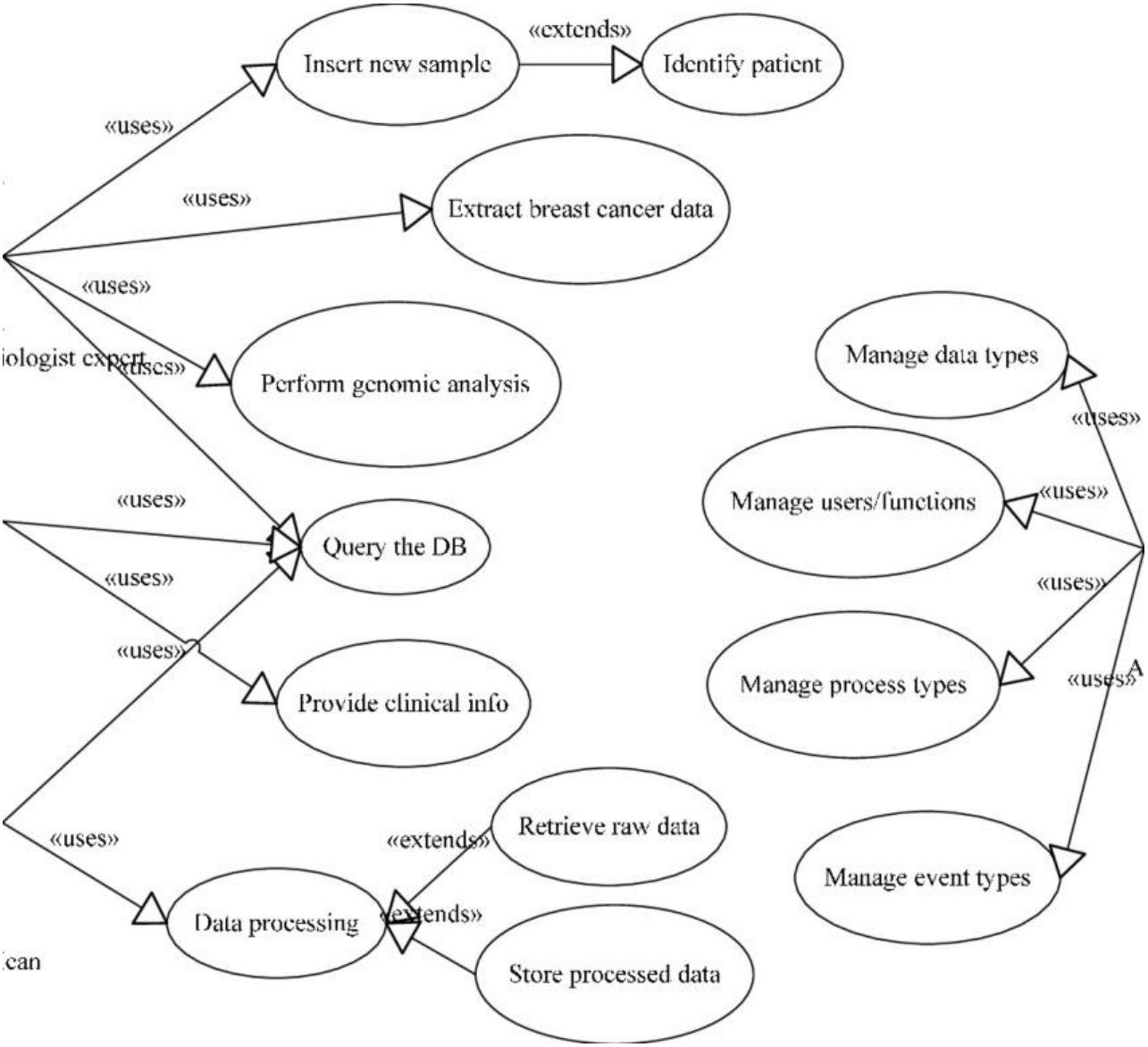
Methodology	**Accuracy**	**Precision**	**Recall**
Naive Bayes classifier	95.61	95.65	93.61
SVM Classifier	95.61	95.65	93.61
Bi-clustering and Ada boost techniques	95.75	95.72	96.26
RCNN classifier	91.30	91.30	89.30
Bidirectional Recurrent Neural Networks (HABiRNN)	82.50	80.90	79.03

CHAPTER-3 Project Design

PROJECT DESIGN



1. Flow chart


Insert new sample
«extends»
Identify patient
«uses»
«uses»
Extract breast cancer data
«uses»
iologist expert
«uses»
Perform genomic analysis
Manage data types
«uses»
Manage users/functions
«uses»
«uses»
Query the DB
«uses»
«uses»
«uses»
Manage process types
«uses»
A
«uses»
Provide clinical info
«uses»
Retrieve raw data
«extends»
Manage event types
Data processing
«extends»
Store processed data
can

2. Use Case

CHAPTER-4 MODULES DESCRIPTION

Well structured designs improve the maintainability of a system. A structured system is one that is developed from the top down and modular, that is, broken down into manageable components. In this project we modularized the system so that they haveminimal effect on each other.

This application is designed into five independent modules which take care of different tasks efficiently.

1. **User Interface Module.**
2. **Admin Module.**
3. **Client Module.**
4. **Database Operations Module.**
5. **Splitting and Merging Module.**
6. **Identify Module.**

User Interface Module:

Actually every application has one user interface for accessing the entire application. In this application also we are providing one user interface for accessing this application. The user interface designed completely based on the end users. It is provide friendly accessing to the users. This user interface has attractive look and feel. Technically I am using the swings in core java for preparing this user interface.

Admin Module: (Table 3)

User requirements	Elaboration	Further Elaboration
Create	Assign new user id & password for an employee.	
Delete	Administrator can delete theuser id & password of unwanted employee.	

Update	First the details of employees are to be obtained by using user id &password.	After obtaining the original details the updated details aresubmitted.

Client Module: (Table 4)

User requirements	Elaboration	Further Elaboration
Login	Employee log in to home page by entering id &password.	
Adding details	Personal details of criminal store in to data base	Images are cropped and saved in database.
Update process	Enter criminal id and obtain his details	Update the details and images of existing criminal
Delete process	Enter criminal id	Delete the details and image of unwanted criminal
Logout	Logout in to the home page	

Splitting and Merging Module: (Table 5)

Requirements	Elaboration	Further Elaboration
View clippings	View all clips and select the clip shown by eyewitness	Compare the clippings with images of criminals
Construction	Construct the face of criminalby clubbing all freezed clippings	

SDLC METHDOLOGIES:

This document play a vital role in the development of life cycle (SDLC) as it describes the complete requirement of the system. It means for use by developers and will be the basic during testing phase. Any changes made to the requirements in the future will have to go through formal change approval process.

SPIRAL MODEL was defined by Barry Boehm in his 1988 article, "A spiral Model of Software Development and Enhancement. This model was not the first model to discuss iterative development, but it was the first model to explain why the iteration models.

As originally envisioned, the iterations were typically 6 months to 2 years long. Each phase starts with a design goal and ends with a client reviewing the progress thus far. Analysis and engineering efforts are applied at each phase of the project, with an eye toward the end goal of the project.

The steps for Spiral Model can be generalized as follows:

- The new system requirements are defined in as much details as possible. This usually involves interviewing a number of users representing all the external or internal users and other aspects of the existing system.
- A preliminary design is created for the new system.

- A first prototype of the new system is constructed from the preliminary design. This is usually a scaled-down system, and represents an approximation of the characteristics of the final product.

- A second prototype is evolved by a fourfold procedure:

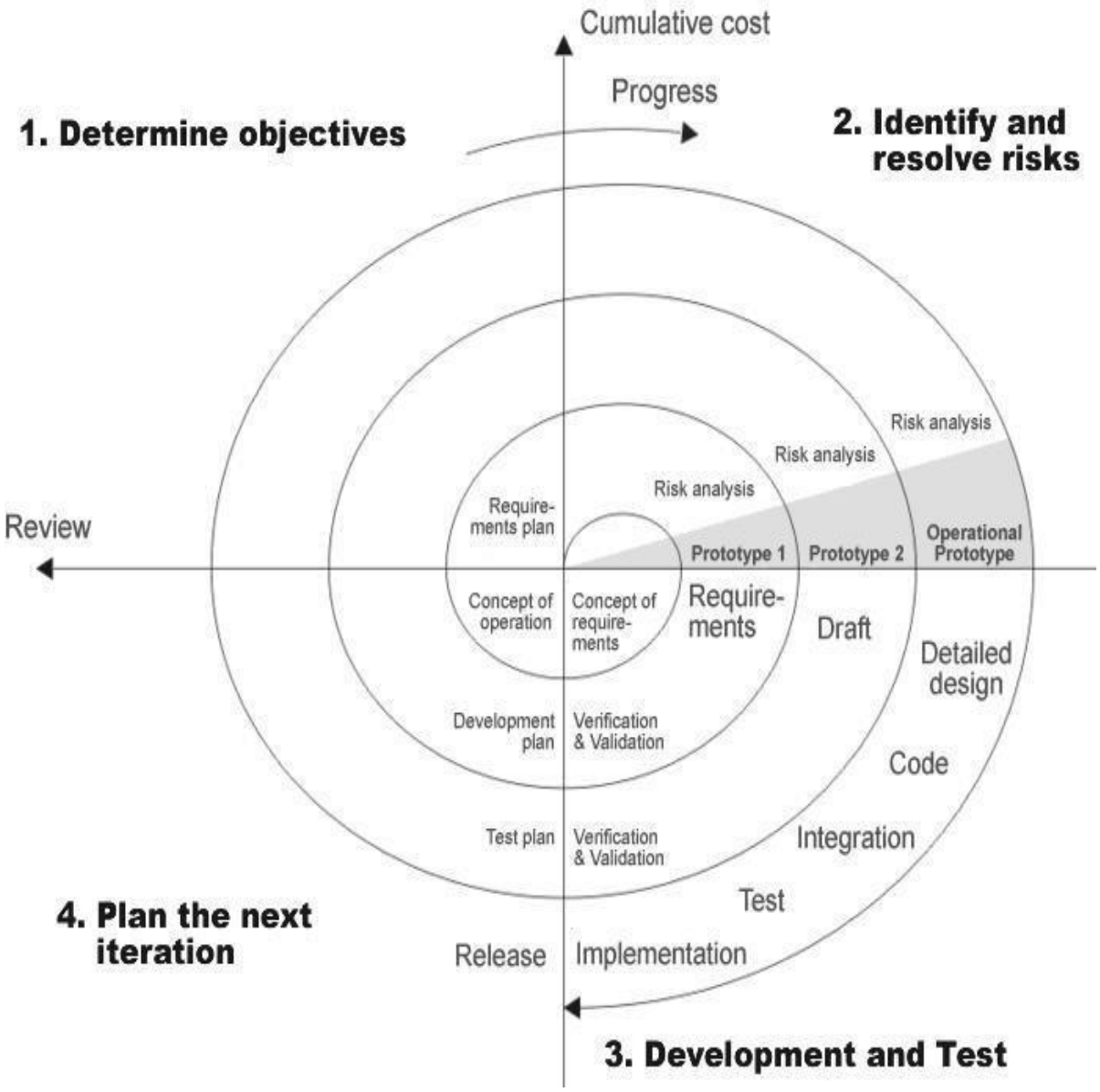
    1. Evaluating the first prototype in terms of its strengths, weakness, and risks.

    2. Defining the requirements of the second prototype.

    3. Planning an designing the second prototype.

4. Constructing and testing the second prototype.

- At the customer option, the entire project can be aborted if the risk is deemed too great. Risk factors might involve development cost overruns, operating-cost miscalculation, or any other factor that could, in the customer's judgment, result in a less-than-satisfactory final product.
- The existing prototype is evaluated in the same manner as was the previous prototype, and if necessary, another prototype is developed from it according to the fourfold procedure outlined above.
- The preceding steps are iterated until the customer is satisfied that the refined prototype represents the final product desired.
- The final system is constructed, based on the refined prototype.
- The final system is thoroughly evaluated and tested. Routine maintenance is carried on a continuing basis to prevent large scale failures and to minimize down time.



ADVANTAGES:

- Estimates(i.e. budget, schedule etc .) become more relistic as work progresses, because important issues discoved earlier .
- It is more able to cope with the changes that are software development generally entails.
- Software engineers can get their hands in and start woring on the core of a project earlier.

APPLICATION DEVELOPMENT N-TIER APPLICATIONS

N-Tier Applications can easily implement the concepts of Distributed Application Design and Architecture. The N-Tier Applications provide strategic benefits to Enterprise Solutions. While 2-tier, client-server can help us create quick and easy solutions and may be used for Rapid Prototyping, they can easily become a maintenance and security night mare

The N-tier Applications provide specific advantages that are vital to the business continuity of the enterprise. Typical features of a real life n-tier may include the following:

- Security
- Availability and Scalability
- Manageability
- Easy Maintenance
- Data Abstraction

The above mentioned points are some of the key design goals of a successful n-tier application that intends to provide a good Business Solution.

DEFINITION

Simply stated, an n-tier application helps us distribute the overall functionality into various tiers or layers:

- Presentation Layer
- Business Rules Layer
- Data Access Layer
- Database/Data Store

Each layer can be developed independently of the other provided that it adheres to the standards and communicates with the other layers as per the specifications. This is the one of the biggest advantages of the n-tier application. Each layer can potentially treat the other layer as a 'Block-Box'.

In other words, each layer does not care how other layer processes the data as long as it sends the right data in a correct format.

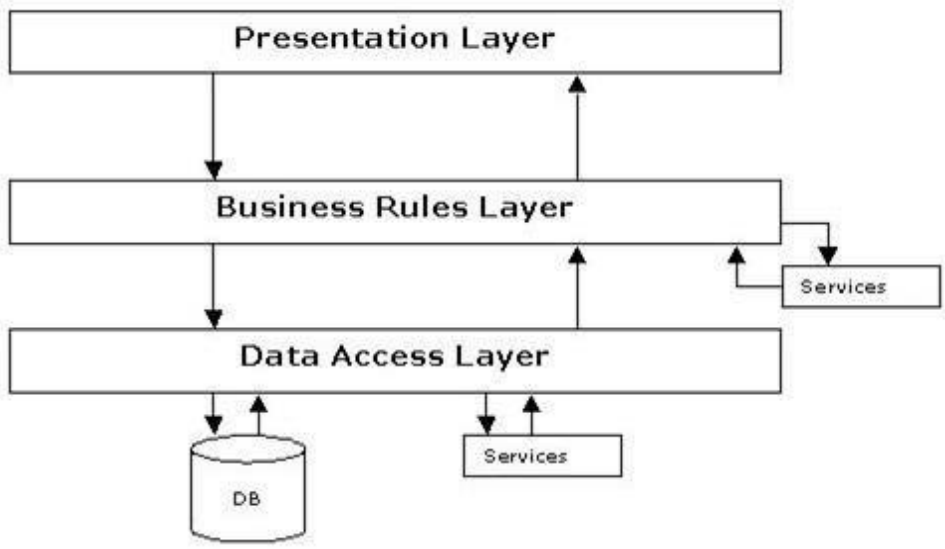


Fig 3-N-Tier Architecture

1-THE PRESENTATION LAYER

Also called as the client layer comprises of components that are dedicated to presenting the data to the user. For example: Windows/Web Forms and buttons, edit boxes, Text boxes, labels, grids, etc.

1 THE BUSINESS RULES LAYER

This layer encapsulates the Business rules or the business logic of the encapsulations. To have a separate layer for business logic is of a great advantage. This is because any changes in Business Rules can be easily handled in this layer. As long as the interface between the layers remains the same, any changes to the functionality/processing logic in this layer can be made without impacting the others. A lot of client-server apps failed to implement successfully as changing the business logic was a painful process. **3THE DATA ACCESS LAYER**

This layer comprises of components that help in accessing the Database. If used in the right way, this layer provides a level of abstraction for the database structures. Simply put changes made to the database, tables, etc do not affect the rest of the application because of the Data Access layer. The different application layers send the data requests to this layer and receive the response from this layer.

4THE DATABASE LAYER

This layer comprises of the Database Components such as DB Files, Tables, Views, etc. The Actual database could be created using SQL Server, Oracle, Flat files, etc.
In an n-tier application, the entire application can be implemented in such a way that it is independent of the actual Database. For instance, you could change the Database Location with minimal changes to Data Access Layer. The rest of the Application should remain unaffected.

MACHINE LEARNING:-

Machine learning (**ML**) is the study of computer algorithms that can improve automatically through experience and by the use of data.[1] It is seen as a part of artificial intelligence.

Machine learning algorithms build a model based on sample data, known as training data, in order to make predictions or decisions without being explicitly programmed to do so.[2] Machine

learning algorithms are used in a wide variety of applications, such as in medicine, email filtering, speech recognition, and computer vision, where it is difficult or unfeasible to develop conventional algorithms to perform the needed tasks.[3]

A subset of machine learning is closely related to computational statistics, which focuses on making predictions using computers; but not all machine learning is statistical learning. The study of mathematical optimization delivers methods, theory and application domains to the field of machine learning. Data mining is a related field of study, focusing on exploratory data analysis through unsupervised learning.[5][6] Some implementations of machine learning use data and neural networks in a way that mimics the working of a biological brain.[7][8] In its application across business problems, machine learning is also referred to as predictive analytics.

Learning algorithms work on the basis that strategies, algorithms, and inferences that worked well in the past are likely to continue working well in the future. These inferences can be obvious, such as "since the sun rose every morning for the last 10,000 days, it will probably rise tomorrow morning as well". They can be nuanced, such as "X% of families have geographically separate species with color variants, so there is a Y% chance that undiscovered black swans exist".[9]

Machine learning programs can perform tasks without being explicitly programmed to do so. It involves computers learning from data provided so that they carry out certain tasks. For simple tasks assigned to computers, it is possible to program algorithms telling the machine how to execute all steps required to solve the problem at hand; on the computer's part, no learning is needed. For more advanced tasks, it can be challenging for a human to manually create the needed algorithms. In practice, it can turn out to be more effective to help the machine develop its own algorithm, rather than having human programmers specify every needed step.[10]

The discipline of machine learning employs various approaches to teach computers to accomplish tasks where no fully satisfactory algorithm is available. In cases where vast numbers of potential answers exist, one approach is to label some of the correct answers as valid. This can then be used as training data for the computer to improve the algorithm(s) it uses to determine correct answers. For example, to train a system for the task of digital character recognition, the MNIST dataset of handwritten digits has often been used.[10]

Association rules

Association rule learning is a rule-based machine learning method for discovering relationships between variables in large databases. It is intended to identify strong rules discovered in databases using some measure of "interestingness".[60]

Rule-based machine learning is a general term for any machine learning method that identifies, learns, or evolves "rules" to store, manipulate or apply knowledge. The defining characteristic of a rule-based machine learning algorithm is the identification and utilization of a set of relational rules that collectively represent the knowledge captured by the system. This is in contrast to other machine learning algorithms that commonly identify a singular model that can be universally applied to any instance in order to make a prediction.[61] Rule-based machine learning approaches include learning classifier systems, association rule learning, and artificial immune systems.

Models

Performing machine learning involves creating a model, which is trained on some training data and then can process additional data to make predictions. Various types of models have been used and researched for machine learning systems.

Artificial neural networks

An artificial neural network is an interconnected group of nodes, akin to the vast network of neurons in a brain. Here, each circular node represents an artificial neuron and an arrow represents a connection from the output of one artificial neuron to the input of

Artificial neural networks (ANNs), or connectionist systems, are computing systems vaguely inspired by the biological neural networks that constitute animal brains. Such systems "learn" to

perform tasks by considering examples, generally without being programmed with any task- specific rules.

An ANN is a model based on a collection of connected units or nodes called "artificial neurons", which loosely model the neurons in a biological brain. Each connection, like the synapses in a biological brain, can transmit information, a "signal", from one artificial neuron to another. An artificial neuron that receives a signal can process it and then signal additional artificial neurons connected to it. In common ANN implementations, the signal at a connection between artificial neurons is a real number, and the output of each artificial neuron is computed by some non- linear function of the sum of its inputs. The connections between artificial neurons are called "edges". Artificial neurons and edges typically have a weight that adjusts as learning proceeds. The weight increases or decreases the strength of the signal at a connection. Artificial neurons may have a threshold such that the signal is only sent if the aggregate signal crosses that threshold. Typically, artificial neurons are aggregated into layers. Different layers may perform different kinds of transformations on their inputs. Signals travel from the first layer (the input layer) to the last layer (the output layer), possibly after traversing the layers multiple times.

The original goal of the ANN approach was to solve problems in the same way that a human brain would. However, over time, attention moved to performing specific tasks, leading to deviations from biology. Artificial neural networks have been used on a variety of tasks, including computer vision, speech recognition, machine translation, social network filtering, playing board and video games and medical diagnosis.

Deep learning consists of multiple hidden layers in an artificial neural network. This approach tries to model the way the human brain processes light and sound into vision and hearing.
Some successful applications of deep learning are computer vision and speech recognition.[68]

Decision trees[

Decision tree learning uses a decision tree as a predictive model to go from observations about an item (represented in the branches) to conclusions about the item's target value (represented in the leaves). It is one of the predictive modeling approaches used in statistics, data mining, and machine learning. Tree models where the target variable can take a discrete set of values are called classification trees; in these tree structures, leaves represent class labels and branches represent conjunctions of features that lead to those class labels. Decision trees where the target variable can take continuous values (typically real numbers) are called regression trees. In decision analysis, a decision tree can be used to visually and explicitly represent decisions and decision making. In data mining, a decision tree describes data, but the resulting classification tree can be an input for decision making

USE OF SUPERVISED MACHINE LEARNING ALGORITHMS:-

Supervised learning is the types of machine learning in which machines are trained using well "labelled" training data, and on basis of that data, machines predict the output. The labelled data means some input data is already tagged with the correct output.

In supervised learning, the training data provided to the machines work as the supervisor that teaches the machines to predict the output correctly. It applies the same concept as a student learns in the supervision of the teacher.

Supervised learning is a process of providing input data as well as correct output data to the machine learning model. The aim of a supervised learning algorithm is to **find a mapping function to map the input variable(x) with the output variable(y)**.

In the real-world, supervised learning can be used for **Risk Assessment, Image classification, Fraud Detection, spam filtering**, etc.

Supervised learning algorithms build a mathematical model of a set of data that contains both the inputs and the desired outputs.[34] The data is known as training data, and consists of a set of training examples. Each training example has one or more inputs and the desired output, also known as a supervisory signal. In the mathematical model, each training example is represented by an array or vector, sometimes called a feature vector, and the training data is represented by a matrix. Through iterative optimization of an objective function, supervised learning algorithms learn a function that can be used to predict the output associated with new inputs.[35] An optimal function will allow the algorithm to correctly determine the output for inputs that were not a part of the training data. An algorithm that improves the accuracy of its outputs or predictions over time is said to have learned to perform that task.[18]

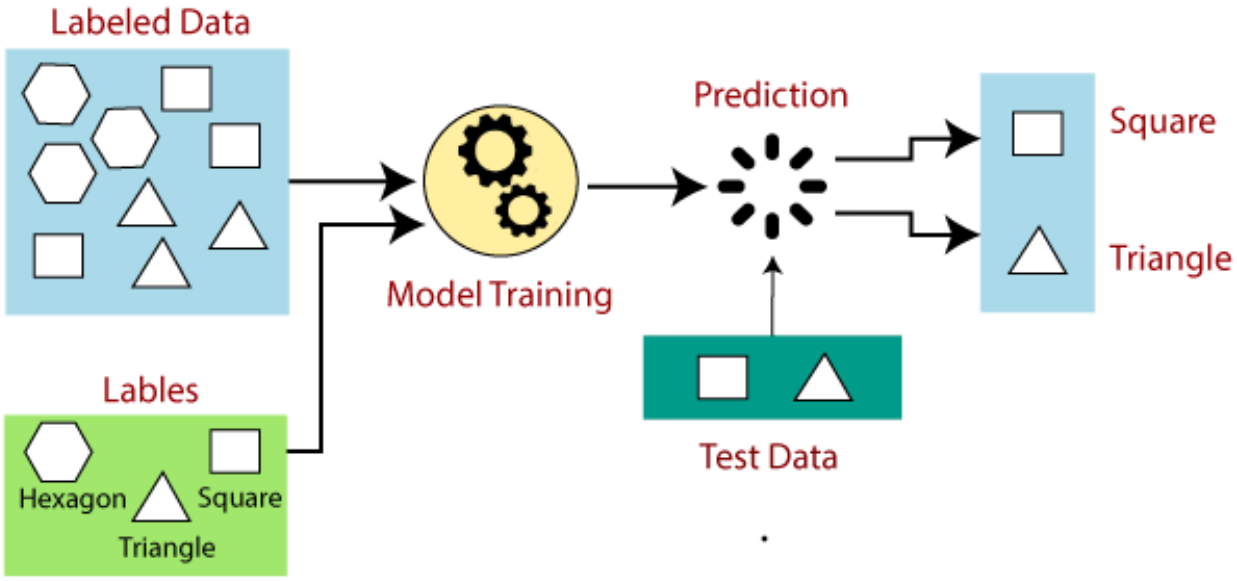
Types of supervised learning algorithms include active learning, classification and regression.

[26] Classification algorithms are used when the outputs are restricted to a limited set of values, and regression algorithms are used when the outputs may have any numerical value within a range. As an example, for a classification algorithm that filters emails, the input would be an incoming email, and the output would be the name of the folder in which to file the email.

Similarity learning is an area of supervised machine learning closely related to regression and classification, but the goal is to learn from examples using a similarity function that measures how similar or related two objects are. It has applications in ranking, recommendation systems, visual identity tracking, face verification, and speaker verification.

How Supervised Learning Works?

In supervised learning, models are trained using labelled dataset, where the model learns about each type of data. Once the training process is completed, the model is tested on the basis of test data (a subset of the training set), and then it predicts the output.

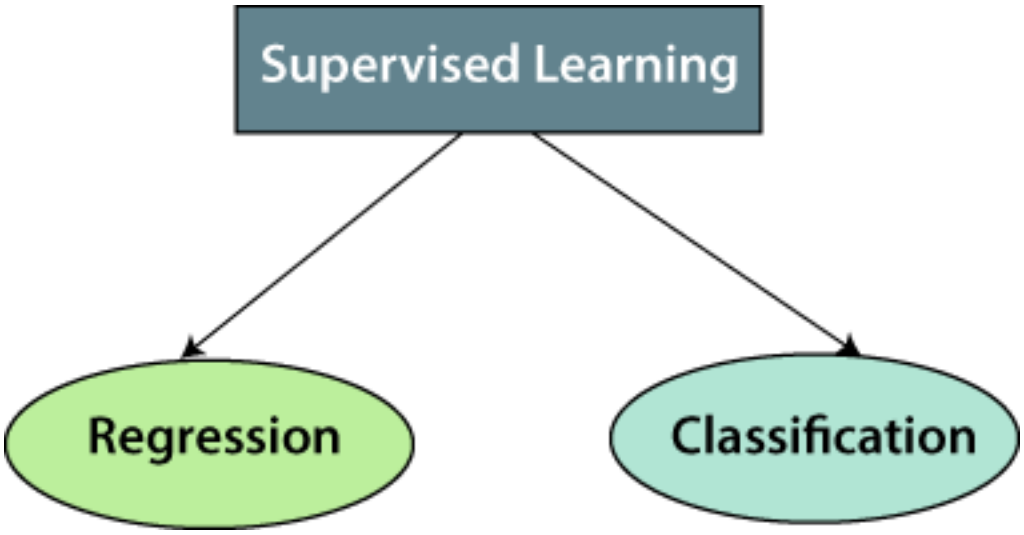


Steps Involved in Supervised Learning:

- First Determine the type of training dataset
- Collect/Gather the labeled training data.
- Split the training dataset into training **dataset, test dataset, and validation dataset**.
- Determine the input features of the training dataset, which should have enough knowledge so that the model can accurately predict the output.
- Determine the suitable algorithm for the model, such as support vector machine, decision tree, etc.
- Execute the algorithm on the training dataset. Sometimes we need validation sets as the control parameters, which are the subset of training datasets.
- Evaluate the accuracy of the model by providing the test set. If the model predicts the correct output, which means our model is accurate.

Types of supervised Machine learning Algorithms:

Supervised learning can be further divided into two types of problems:



1.Regression

Regression algorithms are used if there is a relationship between the input variable and the output variable. It is used for the prediction of continuous variables, such as Weather forecasting, Market Trends, etc. Below are some popular Regression algorithms which come under supervised learning:

Linear Regression

- Regression Trees
- Non-Linear Regression
- Bayesian Linear Regression
- Polynomial Regression

2.Classification

Classification algorithms are used when the output variable is categorical, which means there are two classes such as Yes-No, Male-Female, True-false, etc.

Spam Filtering,

- Random Forest
- Decision Trees
- Logistic Regression
- Support vector Machines

Advantages of Supervised learning:

- With the help of supervised learning, the model can predict the output on the basis of prior experiences.
- In supervised learning, we can have an exact idea about the classes of objects.
- Supervised learning model helps us to solve various real-world problems such as **fraud detection, spam filtering**, etc.

Disadvantages of supervised learning:

- Supervised learning models are not suitable for handling the complex tasks.
- Supervised learning cannot predict the correct output if the test data is different from the training dataset.
- Training required lots of computation times.
- In supervised learning, we need enough knowledge about the classes of object.

CHAPTER-5 Functionality/Working of Project

5.1 RELATED WORK

There have been many studies applying different machine learning techniques on medical analysis. In terms of traditional machine learning methods, Chaurasia et al.[6] used Simple Logistic to reduce the dimension of feature space and applied RepTree and RBF Network to evaluate the performance. Dubey et al. used K-means algorithm to evaluate the impact of clustering using centroid initialization and achieved 92% average positive prediction accuracy [7]. Classification and regression trees (CART) classifier with feature selection and bagging technique was implemented to predict breast cancer in [8]. Wang et al. [9] compared four classifiers: Naive Bayes, Decision Tree, Support Vector Machine and k-nearest neighbor for classification of cancer using gene expression data. These traditional machine learning models has the advantage of low design complexity, but it is not capable of dealing with complex data. In the project, we are going to apply models including these previous mentioned algorithms on the Breast Cancer Wisconsin dataset and compare the performances.

Many deep learning models have also been developed in this objective. In [10], problems in multiple datasets were discussed and Partial Likelihood Artificial Neural Network is applied for prediction of cancer survival. Purwar et al. [11] proposed a hybrid model using a combination of K-means clustering with Multilayer Perceptron with promising results for various medical dataset. Another new classification algorithm for detection of breast abnormalities in digital mammograms using Particle Swarm Optimized Wavelet Neural Network (PSOWNN) is investigated in [12]. These deep learning models are capable of modeling complex and high dimensional data. However, the computational complexity is higher and training time may be long. In this paper we will build a neural network to apply to the same Breast Cancer Wisconsin dataset and compare the performance with traditional models.

5.2 DATASET AND FEATURES

The dataset we used is Breast Cancer Wisconsin dataset which is a widely used dataset in study. It contains 699 instances with 9 features and 2 classes (benign and malignant.) The class distribution is as follows: 458 benign (65.5%) and 241 malignant (34.5%). The features are: 1) radius (mean of distances from center to points on the perimeter). 2) texture (standard deviation of gray-scale values). 3) perimeter. 4) area. 5) smoothness (local variation in radius lengths). 6) compactness ($perimeter^2/area - 1.0$). 7) concavity (severity of concave portions of

the contour). 8) concave points (number of concave portions of the contour). 9) symmetry. 10) fractal dimension (”coastline approximation” - 1).

5.3. Data Preprocessing

We first cleaned the dataset by removing samples with empty values. There are 683 samples after removing invalid samples. We then realized that the dataset with 683 samples is rather small. To enhance the dataset, we generated a new dataset by copying original dataset and add Gaussian noise to it. Afterward, we appended the generated dataset to the original dataset. This process doubles the dataset to 1398 samples. Then we rescaled data to [0,1] using MinMaxScaler.

By plotting scatter matrix of the features, we realized the data is significantly right skewed, which may make the model biased towards the majority of the features.Thus, we took the square root of the feature data to mitigate the data skew problem. The result is shown in Fig 1.

5.4. METHODS

We used 7 traditional models for the classification of breast cancer cases. Feature selection is applied to increase the rate of accurate prediction. Additionally, deep learning model is also built for the diagnosis system. Finally, we compare the performance of all the models applied and choose the one with the highest performance.

5.5. Traditional models

(1) Logistic Regression (LR): Logistic regression predicts the probability of the default class (e.g. Class 2 in this case) and transforms the probability into a binary value (0 or 1) for classification using ”sigmoid” function as shown in Equation 1.

$$f(x) = \frac{1}{1 + e^{-x}} \qquad (1)$$

(2) K-Nearest Neighbor (k-NN): K-nearest neighbor assigns a case to the class that is most common among its k nearest neighbors. The distance between the case and its neighbor is measured by using distance functions like Euclidean: $D_{Euclidean} = \sqrt{\sum_{i=1}^{k} (x_i - y_i)^2}$, Manhattan: $D_{Manhattan} = \sum_{i=1}^{k} |x_i - y_i|$ and Minkowski: $D_{Minkowski} = (\sum_{i=1} (|x_i - y_i|)^q)^{\frac{1}{q}}$

(3) Support Vector Machine (SVM): Support Vector Machine finds an optimal hyperplane that best separates the classes based on the support vectors. The function of kernel for SVM is to take data as input and transform it into the

required form.[13] The kernel function used in SVM model is linear function as shown in Eq.2

$$k(x_i,x_j) = a < x_i,x_j > + b \qquad (2)$$

(4) Kernel Support Vector Machines (Kernel SVM): The kernel SVM in this paper is the SVM algorithm that uses Gaussian radial basis function (RBF) as kernel. The RBF is shown in Eq.3.

$$k(x_i,x_j) = exp(-\gamma||x_i - x_j||^2) \qquad (3)$$

(5) Naive Bayes (NB): NB algorithm makes classifications using the Maximum A Posteriori decision rule (Eq.4) in a Bayesian setting.

$$y = argmax_{c_i}(P(c_i)\prod_{j=1}^{n} P(x_j|c_i)) \qquad (4)$$

where c_i is one of the classes and x_j is one of the features.

(6) Decision Tree (DT): The decision tree are presented with a tree structure. The test objects are classified by their feature values. A node in a decision tree represents an instance, outcomes of the test represented by branch, and the leaf node epitomized the class label[14].

(7) Random Forest (RF): Random forest is a set of individual decision trees. Each decision tree spits out a class prediction. It decides the class of the test object by aggregating the votes from different decision trees[15].

5.6. Feature selection

In many machine learning algorithms, there is a decrease of accuracy when the number of features is redundant [16].In order to improve the accuracy of the models and avoid overfitting, we performed feature selection on the data. For the traditional models, we used two techniques to select features from the dataset. For the Decision Tree and Random Forest model, we generate the feature importance of the last training result and choose features accordingly. Given the importance of the jth feature I_j, we drop two features that has minimum feature importance: $\hat{j}_{drop} = \text{argmin}_j(I_j)$.

For the remaining five models, we used the correlation matrix with heatmap visualization. Correlation represents how the features in the dataset are related to each other. By using the heatmap visualization, it is easier to identify which featuers are highly correlated. Using the seaborn library, we could plot the heatmap for better view. For each group of highly correlated features, we choose only one feature to represent all that are in the group. This way most of the information in

the features is reserved, and the redundant information is dropped to avoid overfitting.

5.7. Deep learning model

The structure is shown in Fig. 2. The first hidden layer has 9 neurons, followed by GaussianNoise layer to improve robustness, and dropout layer to reduce overfitting. Then the process is repeated and the last dense layer has 1 neuron. The activation function is relu.

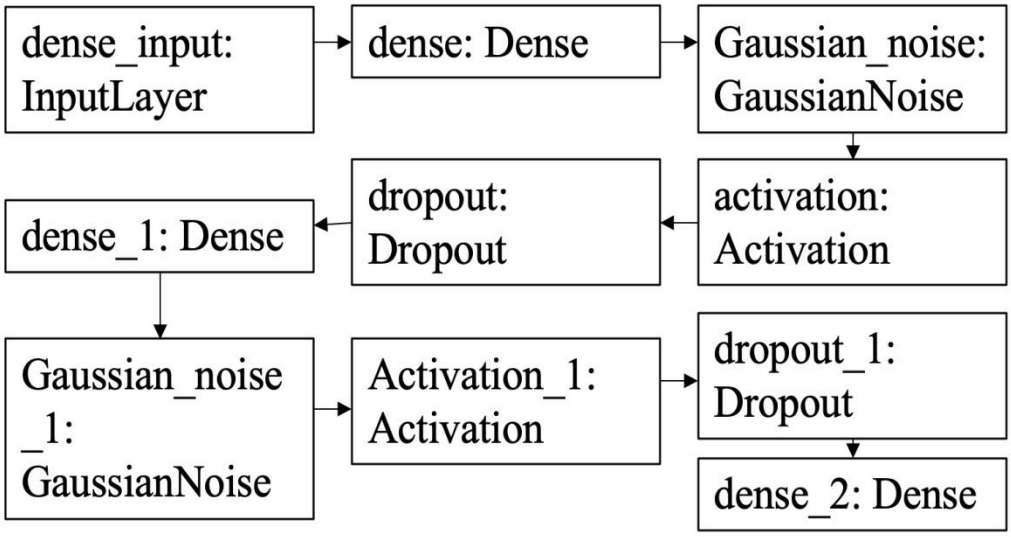


Fig. 4: Deep learning model

Source Code-:

```
import pandas as pd
import numpy as np
from sklearn.datasets import load_breast_cancer
from sklearn.preprocessing import StandardScaler
from keras.models import Sequential
from keras.layers import Dense
data = load_breast_cancer()
data.keys()
print(data['DESCR'])
data['data'].shape
data['feature_names']
```

```
data['data'][0]
j = 0
for i in data['feature_names']:
  print(i,":",data['data'][0][j])
  j+=1
feature = data['data']
label = data['target']
data['target_names']
feature.shape
label.shape
scale = StandardScaler()
feature = scale.fit_transform(feature)
j = 0
for i in data['feature_names']:
  print(i,":",feature[0][j])
  j+=1
print(feature[568])
print(data['target_names'][label[568]],label[568])
df_frt = pd.DataFrame(feature , columns = data['feature_names'])
df_lbl = pd.DataFrame(label , columns = ['label'])
df = pd.concat([df_frt, df_lbl], axis=1)
df = df.sample(frac = 1)
feature = df.values[ : , : 30]
label = df.values[ : ,30: ]
df
#500 Training
X_train = feature[:500]
y_train = label[:500]
#35 Validation
X_val = feature[500:535]
y_val = label[500:535]
#34 Testing
X_test = feature[535:]
y_test = label[535:]
model = Sequential()

model.add(Dense(32, activation = 'relu', input_dim = 30))
```

```
model.add(Dense(64, activation = 'relu'))
model.add(Dense(128, activation = 'relu'))
model.add(Dense(64, activation = 'relu'))
model.add(Dense(32, activation = 'relu'))
model.add(Dense(1, activation = 'sigmoid'))

model.compile( loss = 'binary_crossentropy' , optimizer = 'adam' , metrics = ['accuracy'])
model.fit( X_train , y_train, epochs = 10, batch_size = 5, validation_data = (X_val, y_val))
model.evaluate(X_test , y_test)
model.evaluate(X_val , y_val)
for i in range(30):
  sample = X_test[i]
  sample = np.reshape(sample, (1,30))

  if (model.predict(sample)[0][0] > 0.5):
    print("-Benign")
  else:
    print("-Malignant")

  if (y_test[i] == 1):
    print("*Banign")
  else:
    print("*Melignant")
  print("-----------")
t = 0
for i in y_val:
  if (i == 1):
    t += 1

print(t)
t = 0
for i in y_test:
  if (i == 1):
    t += 1
```

```
print(t)
X_test[0] * -.1
348/350
347/350
32/35
```

Output-:

```
_breast_cancer_dataset:

Breast cancer wisconsin (diagnostic) dataset
--------------------------------------------

*Data Set Characteristics:*

    :Number of Instances: 569

    :Number of Attributes: 30 numeric, predictive attributes and the class

    :Attribute Information:
        - radius (mean of distances from center to points on the perimeter)
        - texture (standard deviation of gray-scale values)
        - perimeter
        - area
        - smoothness (local variation in radius lengths)
        - compactness (perimeter^2 / area - 1.0)
        - concavity (severity of concave portions of the contour)
        - concave points (number of concave portions of the contour)
        - symmetry
        - fractal dimension ("coastline approximation" - 1)

        The mean, standard error, and "worst" or largest (mean of the three
        worst/largest values) of these features were computed for each image,
        resulting in 30 features.  For instance, field 0 is Mean Radius, field
        10 is Radius SE, field 20 is Worst Radius.

        - class:
                - WDBC-Malignant
                - WDBC-Benign

    :Summary Statistics:
```

	Min	Max
radius (mean):	6.981	28.11
texture (mean):	9.71	39.28
perimeter (mean):	43.79	188.5
area (mean):	143.5	2501.0
smoothness (mean):	0.053	0.163
compactness (mean):	0.019	0.345
concavity (mean):	0.0	0.427
concave points (mean):	0.0	0.201
symmetry (mean):	0.106	0.304
fractal dimension (mean):	0.05	0.097
radius (standard error):	0.112	2.873
texture (standard error):	0.36	4.885
perimeter (standard error):	0.757	21.98
area (standard error):	6.802	542.2
smoothness (standard error):	0.002	0.031
compactness (standard error):	0.002	0.135
concavity (standard error):	0.0	0.396
concave points (standard error):	0.0	0.053
symmetry (standard error):	0.008	0.079
fractal dimension (standard error):	0.001	0.03
radius (worst):	7.93	36.04
texture (worst):	12.02	49.54
perimeter (worst):	50.41	251.2
area (worst):	185.2	4254.0
smoothness (worst):	0.071	0.223
compactness (worst):	0.027	1.058
concavity (worst):	0.0	1.252
concave points (worst):	0.0	0.291
symmetry (worst):	0.156	0.664
fractal dimension (worst):	0.055	0.208

:Missing Attribute Values: None

:Class Distribution: 212 - Malignant, 357 - Benign

:Creator: Dr. William H. Wolberg, W. Nick Street, Olvi L. Mangasarian

:Donor: Nick Street

:Date: November, 1995

This is a copy of UCI ML Breast Cancer Wisconsin (Diagnostic) datasets.
https://goo.gl/U2Uwz2

Features are computed from a digitized image of a fine needle
aspirate (FNA) of a breast mass. They describe
characteristics of the cell nuclei present in the image.

Separating plane described above was obtained using
Multisurface Method-Tree (MSM-T) [K. P. Bennett, "Decision Tree
Construction Via Linear Programming." Proceedings of the 4th
Midwest Artificial Intelligence and Cognitive Science Society,
pp. 97-101, 1992], a classification method which uses linear
programming to construct a decision tree. Relevant features
were selected using an exhaustive search in the space of 1-4
features and 1-3 separating planes.

The actual linear program used to obtain the separating plane
in the 3-dimensional space is that described in:
[K. P. Bennett and O. L. Mangasarian: "Robust Linear
Programming Discrimination of Two Linearly Inseparable Sets",
Optimization Methods and Software 1, 1992, 23-34].

This database is also available through the UW CS ftp server:

ftp ftp.cs.wisc.edu
cd math-prog/cpo-dataset/machine-learn/WDBC/

.. topic:: References

- W.N. Street, W.H. Wolberg and O.L. Mangasarian. Nuclear feature extraction for breast tumor diagnosis. IS&T/SPIE 1993 International Symposium on Electronic Imaging: Science and Technology, volume 1905, pages 861-870, San Jose, CA, 1993.
- O.L. Mangasarian, W.N. Street and W.H. Wolberg. Breast cancer diagnosis and prognosis via linear programming. Operations Research, 43(4), pages 570-577, July-August 1995.
- W.H. Wolberg, W.N. Street, and O.L. Mangasarian. Machine learning techniques to diagnose breast cancer from fine-needle aspirates. Cancer Letters 77 (1994) 163-171.

mean radius : 17.99

mean texture : 10.38
mean perimeter : 122.8
mean area : 1001.0
mean smoothness : 0.1184
mean compactness : 0.2776
mean concavity : 0.3001
mean concave points : 0.1471
mean symmetry : 0.2419
mean fractal dimension : 0.07871
radius error : 1.095
texture error : 0.9053
perimeter error : 8.589
area error : 153.4
smoothness error : 0.006399
compactness error : 0.04904
concavity error : 0.05373
concave points error : 0.01587
symmetry error : 0.03003
fractal dimension error : 0.006193
worst radius : 25.38
worst texture : 17.33
worst perimeter : 184.6
worst area : 2019.0
worst smoothness : 0.1622
worst compactness : 0.6656
worst concavity : 0.7119
worst concave points : 0.2654
worst symmetry : 0.4601
worst fractal dimension : 0.1189
mean radius : 1.0970639814699807
mean texture : -2.0733350146975935
mean perimeter : 1.2699336881399383
mean area : 0.9843749048031144
mean smoothness : 1.568466329243428
mean compactness : 3.2835146709868264
mean concavity : 2.652873983743168
mean concave points : 2.532475216403245
mean symmetry : 2.2175150059646405
mean fractal dimension : 2.255746885296269
radius error : 2.4897339267376193
texture error : -0.5652650590684639
perimeter error : 2.833030865855184
area error : 2.4875775569611043

smoothness error : -0.21400164666895383
compactness error : 1.3168615683959484
concavity error : 0.72402615808036
concave points error : 0.6608199414286064
symmetry error : 1.1487566671861758
fractal dimension error : 0.9070830809973359
worst radius : 1.8866896251792757
worst texture : -1.3592934737640827
worst perimeter : 2.3036006236225606
worst area : 2.0012374893299207
worst smoothness : 1.3076862710715387
worst compactness : 2.616665023512603
worst concavity : 2.1095263465722556
worst concave points : 2.296076127561788
worst symmetry : 2.750622244124955
worst fractal dimension : 1.9370146123781782
[-1.80840125 1.22179204 -1.81438851 -1.34778924 -3.11208479 -1.15075248
 -1.11487284 -1.26181958 -0.8200699 -0.56103238 -0.07027874 0.3830925
 -0.15744905 -0.46615196 0.04934236 -1.16351619 -1.05750068 -1.91344745
  0.75282996 -0.382754   -1.41089258 0.76418957 -1.43273495 -1.07581292
 -1.85901852 -1.2075525 -1.30583065 -1.74506282 -0.04813821 -0.75120669]
benign 1
Epoch 1/10
100/100 [==============================] - 1s 5ms/step - loss: 0.2376 - accuracy: 0.9320 - val_loss: 0.0627 - val_accuracy: 0.9429
Epoch 2/10
100/100 [==============================] - 0s 3ms/step - loss: 0.0886 - accuracy: 0.9720 - val_loss: 0.0263 - val_accuracy: 1.0000
Epoch 3/10
100/100 [==============================] - 0s 3ms/step - loss: 0.0565 - accuracy: 0.9820 - val_loss: 0.0210 - val_accuracy: 1.0000
Epoch 4/10
100/100 [==============================] - 0s 3ms/step - loss: 0.0448 - accuracy: 0.9840 - val_loss: 0.0076 - val_accuracy: 1.0000
Epoch 5/10
100/100 [==============================] - 0s 3ms/step - loss: 0.0362 - accuracy: 0.9900 - val_loss: 0.0240 - val_accuracy: 1.0000
Epoch 6/10
100/100 [==============================] - 0s 3ms/step - loss: 0.0474 - accuracy: 0.9860 - val_loss: 0.0050 - val_accuracy: 1.0000
Epoch 7/10
100/100 [==============================] - 0s 3ms/step - loss: 0.0406 - accuracy: 0.9860 - val_loss: 0.0076 - val_accuracy: 1.0000

Epoch 8/10
100/100 [==============================] - 0s 2ms/step - loss: 0.0214 - accuracy: 0.9960 - val_loss: 0.0056 - val_accuracy: 1.0000
Epoch 9/10
100/100 [==============================] - 0s 3ms/step - loss: 0.0162 - accuracy: 0.9960 - val_loss: 0.0039 - val_accuracy: 1.0000
Epoch 10/10
100/100 [==============================] - 0s 3ms/step - loss: 0.0131 - accuracy: 0.9960 - val_loss: 0.0038 - val_accuracy: 1.0000
2/2 [==============================] - 0s 10ms/step - loss: 0.1259 - accuracy: 0.9412
2/2 [==============================] - 0s 8ms/step - loss: 0.0038 - accuracy: 1.0000
-Benign
*Banign

-Malignant
*Melignant

-Benign
*Banign

-Benign
*Melignant

-Benign
*Banign

-Benign
*Banign

-Malignant
*Melignant

-Malignant
*Melignant

-Malignant
*Melignant

-Malignant
*Melignant

-Benign
*Banign

-Malignant
*Melignant

-Benign
*Banign

-Benign
*Banign

-Benign
*Banign

-Benign
*Banign

-Malignant
*Banign

-Benign
*Banign

-Benign
*Banign

-Benign
*Banign

-Benign
*Banign

-Benign
*Banign

-Malignant
*Melignant

-Benign
*Banign

-Benign
*Banign

-Benign
*Banign

-Malignant
*Melignant

-Malignant
*Melignant

-Malignant
*Melignant

-Malignant
*Melignant

21
22
0.9142857142857143

CHAPTER-6 Result and Discussion

6.1. Experiment setting

The server we used is Google Colab. We used Scikit-learn machine learning package for the implementatin and evaluation of traditional models as well as preprocessing of data. TensorFlow Core v2.2.0 is used for developing the neural network model. For the visualization of results, we used Seaborm and Matplotlib visualization. We applied our models on the enhanced Breast Cancer Wisconsin dataset with a 60%-40% training-testing split (838/560) . For traditional methods, additional 70%-30% and 80%-20% training-testing splits were applied. For k-nearest neighbor model, the number of neighbors was set to be 5 and the ”Minkowski” distance function was used. For SVM model, linear kernel was used. For kernel SVM, the kernel function was set to be RBF. For decision tree model, the function to measure the quality of a split was set to be entropy. For the random forest model, the number of estimators was set to be 10 and the function to measure the quality of a split was set to be entropy. All these parameters were set as above based on [17]. With these parameters, the seven traditional models yield good results. The performance of the models are evaluated by performance metrics including accuracy, F score and confusion matrices. These metrics exhibit similar result, hence we will mainly talk about accuracy as the measure of model performance.

6.2. Results-

Model	Accuracy (60/40)	Accuracy (70/30)	Accuracy (80/20)	Confusion Matrix (60/40)		
Logistic Regression	95.53%	95%	93.57%	Benign	368	11
				Malignant	14	167
					Benign	Malignant
K-nearest Neighbor	94.82%	94.76%	95%	Benign	367	12
				Malignant	17	164
					Benign	Malignant
SVM	95.71%	95.48%	93.57%	Benign	366	13
				Malignant	11	170
					Benign	Malignant
Kernel SVM	95.89%	96.19%	95.36%	Benign	365	14
				Malignant	9	172
					Benign	Malignant
Naïve Bayes	95.54%	95.71%	95.71%	Benign	362	17
				Malignant	8	173
					Benign	Malignant
Decision Tree	95.54%	94.76%	96.07%	Benign	370	9
				Malignant	16	165
					Benign	Malignant
Random Forest	96.61%	96.19%	96.43%	Benign	366	13
				Malignant	6	175
					Benign	Malignant

Table 6: Traditional models' accuracy and confusion matrix

Fig.2 shows the accuracies of the seven traditional models with three different training-test splits (60%-40%, 70%-30% and 80%-20%). The confusion matrix of each model with a 60%-40% training-test split is also shown in Fig.2. As a result, all the seven models achieved a very high accuracy around 95%. With all three different training-test splits, the RF model achieved the highest accuracy of over 96%. As shown in the confusion matrix of random forest, the model predicted accurately on malignant cases. Only 6 cases were misclassified. For logisctic regression and SVM, the accuracy decreases with the increasing percentage of the training set. LR model did a relatively good work in classifying benign cases, but the number of misclassified malignant cases is relatively large. SVM model did a bit better than LR in classifying malignant cases, but did a bit worse in classifying benign cases. The change of training set ratio did not affect much on k-NN, Kernel SVM and NB models. NB and Kernel SVM models achieved a high accuracy in classifying malignant cases while k-NN did not. But the misclassify rates of benign cases of the two models were high while kNN model's was relatively low. For DT

model, the change of training set ratio had a slight effect on the accuracy. The highest acuuray of 96% was achieved with 80% training set, and the lowest accuracy of 94% was achieved with 60% training set. DT model has the second highest misclassify rate of malignant case, but it achieved the highest accuracy in classifying benign cases among all 7 models.

The feature importance of Decision Tree and Random Forest model are shown in Fig.3. We chose to drop the two features with minimum feature importance and trained the model again with remaining features. The accuracy of the model before and after feature selection is shown in Table 1. The accuracy of the DT model is increase by applying feature selection while the accuracy of RF model is decrease. From the feature importance map we can see the importance of the last two features of the DT model are significantly lower than the RF algorithm. The RF model distributes the importance over features more evenly, which makes dropping features have negative impact on the model accuracy.

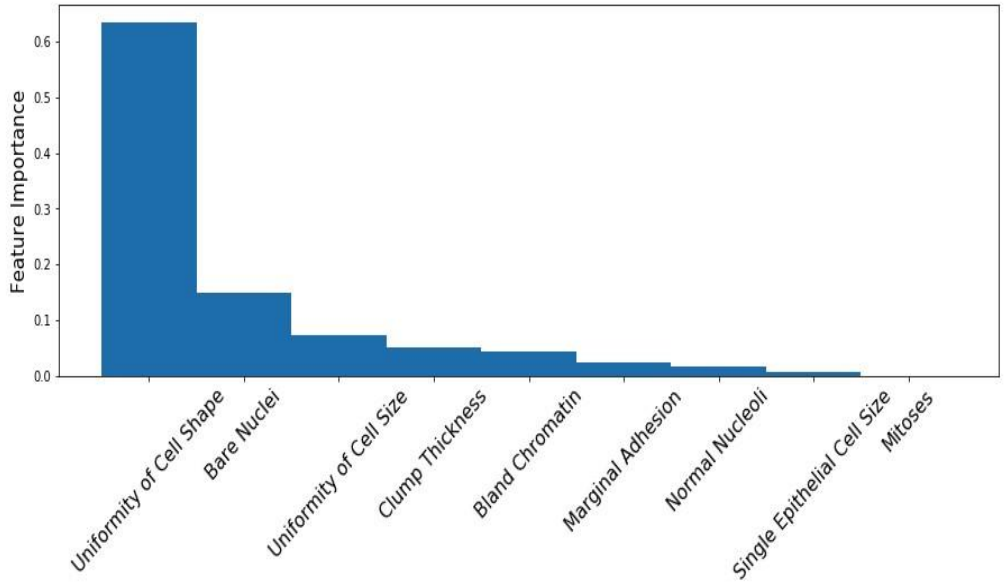
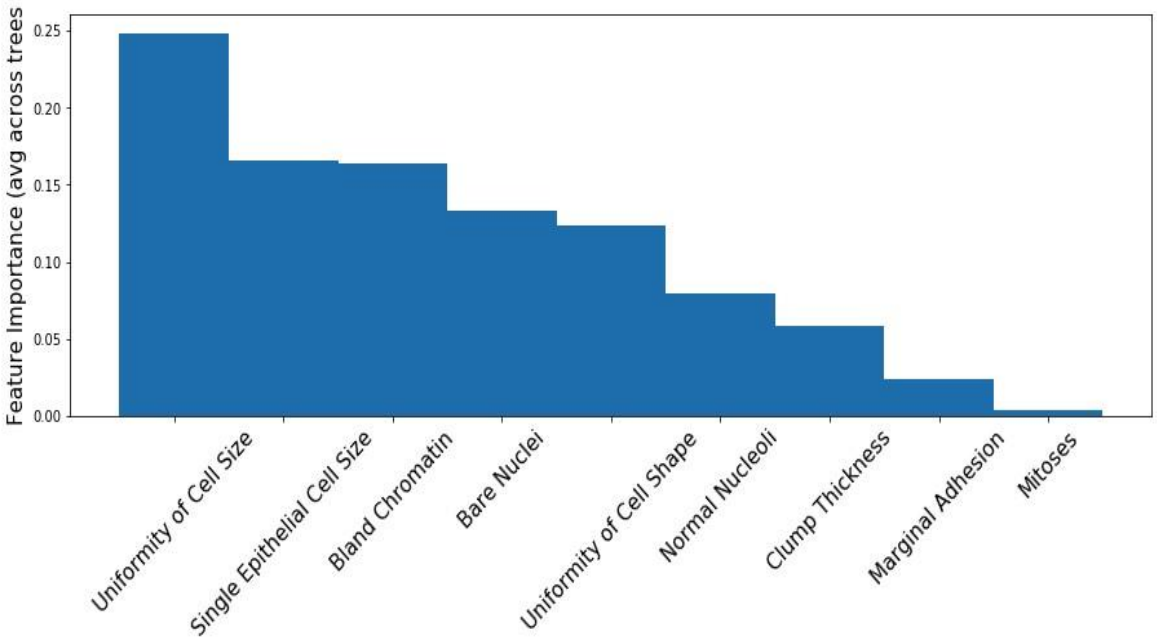


Fig 5.Decision Tree model



(a) Random Forest model

Fig. 6: Feature importances

	Before FS	After FS
Decision Tree	0.9660	0.9696
Random Forest	0.9678	0.9625

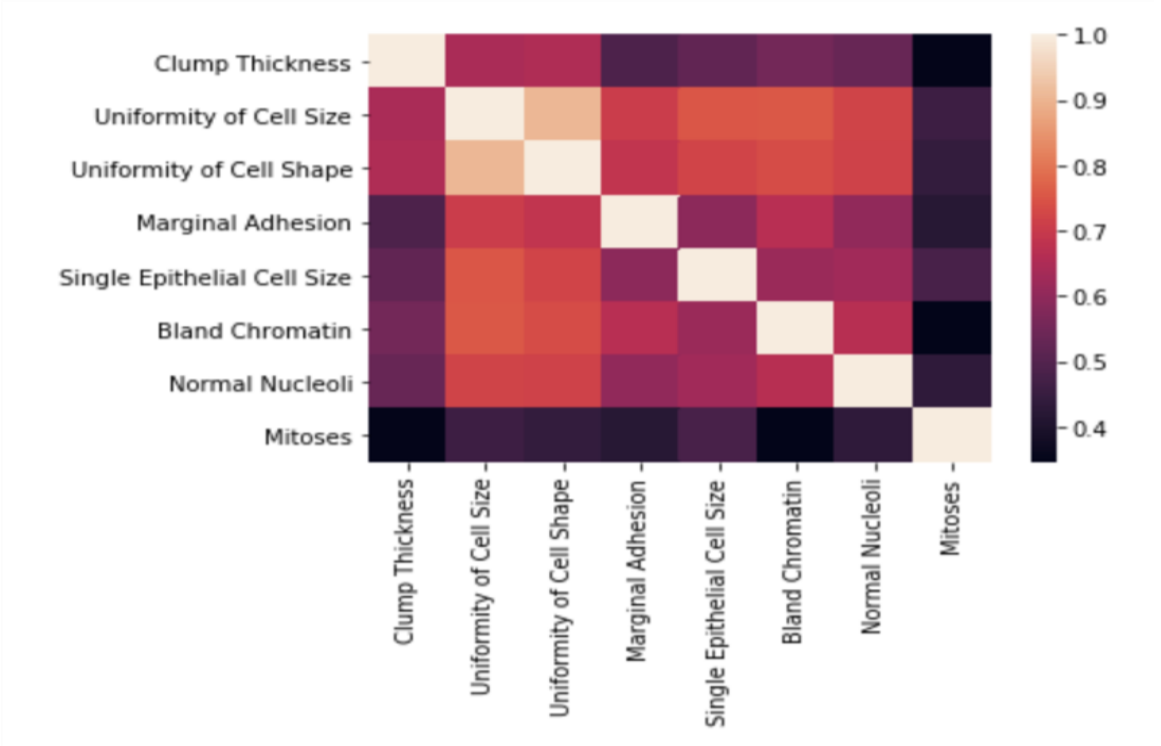
Table 7: Accuracy of models before and after feature selection by feature importance (FS: feature selection)

	Before FS	After FS
Linear Regression	0.9571	0.9589
k-NN	0.9500	0.9553

Table 8: Accuracy of models before and after feature selection by correlation matrix (FS: feature selection)

The heat map visualization is shown in Fig. 5. The features that have high correlations will have a lighter color in the corresponding grid. We replaced group of features that have high correlation with only one from the group. Table 2 shows

that applying feature selection have increased the accuracy of both the LR and the k-NN models.



Train/Validation data split	60/40	70/30	80/20	Confusion matrix (60/40)	Benign	Malignant
Accuracy	96.96%	97.14%	97.50 %	Benign	357	10
				Malignant	7	186

Fig. 7: Heat map visualization of correlation matrix

Fig. 8: Deep learning model's accuracy and confusion matrix

For the deep learning method, it has an accuracy of 96.96% under 60%-40% data split, which is slightly higher than traditional method. The accuracy increases with the number of training data. From the confusion matrix, we can see that the deep learning model missed 17 cases in all, as shown in Fig 5.

CHAPTER-7Conclusion and Future Scope

7.1 Conclusion-

Breast cancer is considered to be one of the significant causes of death in women. Early detection of breast cancer plays an essential role to save women's life. Breast cancer detection can be done with the help of modern machine learning algorithms. We proposed a system that can detect breast cancer and how machine learning algorithm (ML) can improve the early detection and diagnosis of breast cancer. According to many different research papers, support vector machine (SVM) is one of the most powerful machines learning (ML) algorithm that is able to model the human understanding of classifying data. It can find the relationship between data and segregates them accordingly. Here we try to propose the best (accuracy) results for diagnosis and classification in breast cancer.

7.2 Future Scope-

AI is set to change the medical industry in the coming decades — it wouldn't make sense for pathology to not be disrupted too.

Currently, ML models are still in the testing and experimentation phase for cancer prognoses. As datasets are getting larger and of higher quality, researchers are building increasingly accurate models.

Here's what a future cancer biopsy might look like:

You perform clinical tests, either at a clinic or at home. Data is inputted into a pathological ML system. A few minutes later, you receive an email with a detailed report that has an accurate prediction about the development of your cancer.

While you might not see AI doing the job of a pathologist today, you can expect ML to replace your local pathologist in the coming decades, and it's pretty exciting! ML models still have a long way to go, most models still lack sufficient data and suffer from bias. Yet, something we are certain of is that ML is the next step of pathology**, and** it will disrupt the industry.

In this project in python, we learned to build a breast cancer tumour predictor on the wisconsin dataset and created graphs and results for the same. It has been

observed that a good dataset provides better accuracy. Selection of appropriate algorithms with good home dataset will lead to the development of prediction systems. These systems can assist in proper treatment methods for a patient diagnosed with breast cancer. There are many treatments for a patient based on breast cancer stage; data mining and machine learning can be a very good help in deciding the line of treatment to be followed by extracting knowledge from such suitable database

References-

- https://link.springer.com/chapter/10.1007/978-981-15-7205-0_10
- https://www.ijert.org/breast-cancer-detection-using-machine-learning-techniques
- https://www.irjet.net/archives/V8/i2/IRJET-V8I2129.pdf
- https://www.sciencedirect.com/science/article/pii/S2405959520300801
- https://www.researchgate.net/publication/327974742_Early_Detection_of_Breast_Cancer_Using_Machine_Learning_Techniques
- A. C. Society, "How common is breast cancer?." https://www.cancer.org/cancer/breastcancer/about/how-common-is-breast-cancer.html, 2011.
- UCHealth, "How accurate are mammograms?." https://www.uchealth.org/today/how-accurate-aremammograms/, 2015.
- "Breast cancer misdiagnosis and mammography errors." https://hackmd.io/@shaochia/Hk6fwgI1r?type=view1Introduction, 2011.
- L. Wang, "Early diagnosis of breast cancer," *Sensors*, vol. 17, no. 7, p. 1572, 2017.
- K. R. Foster, R. Koprowski, and J. D. Skufca, "Machine learning, medical diagnosis, and biomedical engineering research-commentary," *Biomedical engineering online*, vol. 13, no. 1, p. 94, 2014.
- V. Chaurasia and S. Pal, "Data mining techniques: to predict and resolve breast cancer survivability," *International Journal of Computer Science and Mobile Computing IJCSMC*, vol. 3, no. 1, pp. 10–22, 2014.
- A. K. Dubey, U. Gupta, and S. Jain, "Analysis of kmeans clustering approach on the breast cancer wisconsin dataset," *International journal of computer assisted radiology and surgery*, vol. 11, no. 11, pp. 2033–2047, 2016.
- D. Lavanya and K. U. Rani, "Ensemble decision tree classifier for breast cancer data," *International Journal of Information Technology Convergence and Services*, vol. 2, no. 1, p. 17, 2012.

- X. Wang and O. Gotoh, "A robust gene selection method for microarray-based cancer classification," *Cancer informatics*, vol. 9, pp. CIN–S3794, 2010.
- R. Marshall, "Artificial neural networks in cancer management,"
- A. Purwar and S. K. Singh, "Hybrid prediction model with missing value imputation for medical data," *Expert Systems with Applications*, vol. 42, no. 13, pp. 5621– 5631, 2015.
- J. Dheeba, N. A. Singh, and S. T. Selvi, "Computeraided detection of breast cancer on mammograms: A swarm intelligence optimized wavelet neural network approach," *Journal of biomedical informatics*, vol. 49, pp. 45–52, 2014.
- D. Team, "Kernel functions-introduction to svm kernel examples." https://data-flair.training/blogs/svm-kernelfunctions/, 2018.
- S. B. Ranjit Panigrahi, "Classification and analysis of facebook metrics dataset using supervised classifiers," *Social Network Analytics*, 2019.
- A. Chakure, "Random forest classification and its implementation in python." https://towardsdatascience.com/random-forestclassification-and-its-implementation-d5d840dbead0, 2019.
- M. B. Kursa, W. R. Rudnicki, *et al.*, "Feature selection with the boruta package," *J Stat Softw*, vol. 36, no. 11, pp. 1–13, 2010.
- vishabh goel, "Building a simple machine learning model on breast cancer data." https://towardsdatascience.com/building-a-simplemachine-learning-model-on-breast-cancer-dataeca4b3b99fa3, 2018.