

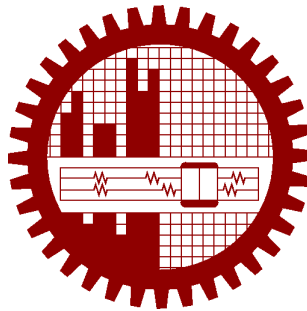
ANALYSIS OF SUPERVISED MACHINE LEARNING MODELS FOR BREAST CANCER PREDICTIONS

by

Md. Arman Hussain

ICT1015312007

MASTER OF ENGINEERING
IN
INFORMATION AND COMMUNICATION TECHNOLOGY



Institute of Information and Communication Technology
Bangladesh University of Engineering and Technology

Dhaka, Bangladesh

March 2022

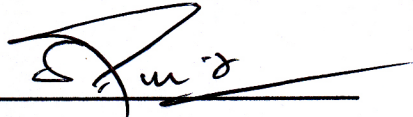
This Project titled, “**Analysis of Supervised Machine Learning Models For Breast Cancer Predictions**”, submitted by Md. Arman Hussain, Roll No.: ICT1015312007, Session: October 2015, has been accepted as satisfactory in partial fulfillment of the requirement for the degree of **MASTER OF ENGINEERING** in Information and Communication Technology on 27th March 2022.

BOARD OF EXAMINERS



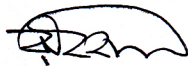
Dr. Hossen Asiful Mustafa
Associate Professor
IICT, BUET, Dhaka

Chairman
(Supervisor)



Dr. Md. Saiful Islam
Professor
IICT, BUET, Dhaka

Member



Dr. Md. Rubaiyat Hossain Mondal
Director and Professor
IICT, BUET, Dhaka

Member

Candidate's Declaration

This is to certify that the work presented in this Project entitled, "Analysis of Supervised Machine Learning Models For Breast Cancer Predictions", is the outcome of the research carried out by Md. Arman Hussain under the supervision of Dr. Hossen Asiful Mustafa, Associate Professor, Institute of Information and Communication Technology (IICT), Bangladesh University of Engineering and Technology (BUET), Dhaka-1000, Bangladesh.

It is also declared that neither this Project nor any part thereof has been submitted anywhere else for the award of any degree, diploma, or other qualifications.

Signature of the Candidate

Md. Arman Hussain

Md. Arman Hussain

ICT1015312007

Dedication

This Project work is dedicated to my mother, Shanaz Begum, who has been a constant source of support and encouragement during the challenges of postgraduate university and life. I am truly thankful for having you in my life. This work is also dedicated to my family and friends, who have always loved me unconditionally and whose good examples have taught me to work hard for the things that I aspire to achieve.

Contents

Certification	ii
Candidate’s Declaration	iii
Dedication	iv
List of Figures	viii
List of Tables	ix
Acknowledgement	x
Abstract	xi
1 Introduction	1
1.1 Motivation of the Project	2
1.2 Objectives of the Project	3
1.3 Outline of the Report	4
2 Literature Review	5
2.1 Summary	9
3 Background	10
3.1 Supervised Learning Algorithms	10
3.1.1 Support Vector Machine	10
3.1.2 K Nearest Neighbours (KNN)	12
3.1.3 Decision Tree	13
3.1.4 Random Forest	14
3.1.5 Logistic Regression	16
3.1.6 The Extra Trees	16
3.2 Feature Selection	17
3.3 Principal Component Analysis	18
3.4 Train-Test Split	19

3.5	Summary	19
4	Proposed System	21
4.1	Dataset	21
4.1.1	Dataset1	21
4.1.2	Dataset2	22
4.1.3	Dataset3	22
4.2	Data Pre-processing	23
4.3	Building Classification Models	25
4.3.1	SVM	25
4.3.2	KNN	25
4.3.3	LR	25
4.3.4	DT	25
4.3.5	RF	26
4.3.6	ET	26
4.4	Training and Testing Phases	26
4.5	Ensemble classification	27
4.5.1	Bagging based Ensemble Classifier	27
4.5.2	Boosting based Ensemble Classifier	28
4.5.3	Voting based Ensemble Classifier	28
4.5.4	Ensemble Model	28
4.6	Experimental Setup	29
4.7	Summary	29
5	Experimental Results	31
5.1	Performance Metrics	31
5.1.1	Confusion Matrix	31
5.2	Model Performance	33
5.2.1	Performance measure for Dataset 1	34
5.2.2	Performance measure for Dataset 2	35
5.2.3	Performance measure for Dataset 3	37
5.2.4	Performance Measure by Ensemble Classification Techniques	38
5.3	Performance Analysis	39
5.4	Summary	39
6	Conclusions	42
6.1	Conclusions	42
6.2	Future Prospects of Our Work	43

List of Figures

2.1	This figures shows the branches of machine learning.	7
3.1	Support Vector Machine	11
3.2	Decision Tree	14
3.3	Random Forest	15
3.4	Extremely Randomized Trees Classifier	17
4.1	Cumulative Variance of Data	24
5.1	Performance Measure for Wisconsin Breast Cancer Dataset	35
5.2	Performance Measure for Wisconsin Breast Cancer (Diagnostic) Dataset	36
5.3	Performance Measure for Breast Cancer Dataset of University Medical Centre,Institute of Oncology, Ljubljana, Yugoslavia)	38
5.4	This figures shows comparative analysis for the three datasets in terms of accuracy, precision, recall, F1 score and CV score	40

List of Tables

4.1	Attributes of Dataset-1	22
4.2	Attributes of Dataset-2	23
4.3	Attributes of Dataset-3	24
5.1	Confusion Matrix	32
5.2	Performance analysis for Wisconsin Breast Cancer Dataset	34
5.3	Performance Measure for Breast Cancer Wisconsin (Diagnostic) Dataset	35
5.4	Performance Measure for Breast Cancer Dataset of University Medical Centre	37
5.5	Performance Measure by using Ensemble Classifier	39
5.6	Comparison Study of Breast Cancer Prediction	41

Acknowledgement

First and foremost, I express my deepest gratitude to Almighty Allah for bestowing his blessings on me and giving me the ability to accomplish this work successfully.

I would like to express my deepest sense of thankfulness and gratitude to my Project supervisor **Dr. Hossen Asiful Mustafa**, Associate Professor, IICT, BUET for leading me into the research field of Machine Learning and deep learning. His scholarly guidance, constant and energetic supervision, and valuable advice made this work a successful one. He has been a continuous source of inspiration and a real motivating force throughout my research work.

Abstract

Carcinoma is one of the scariest and frequently occurring cancers nowadays among females. It affects nearly around ten percent of the females all over the world at some point of their lives. Although, the cure for this cancer is currently obtainable, the treatment is not effective enough if the disease is not identified at the early stages. Early detection of disease has become a crucial problem due to rapid population growth in medical research in recent times. With the rapid population growth, the risk of death incurred by breast cancer is rising exponentially. Breast cancer is the second most severe cancer among all of the cancers already unveiled. An automatic disease detection system aids medical staffs in disease diagnosis and offers reliable, effective, and rapid response as well as decreases the risk of death. Generally, some contemporary medical tests: roentgenogram, breast ultrasound, biopsy, etc., are used for identification of breast cancer. As an alternative, researchers are exploring machine learning techniques for classifying tumours at different stages, e.g., benign and malignant. Classification and data processing strategies can be an effective mechanism for prediction of cancer. Especially in medical field, these methods have been used to predict and to make decisions. In this project, we analyse six classification models: Decision Tree, K Nearest Neighbours, Random Forest, Logistic Regression, Extra Trees and Support Vector Machine on three different datasets from the UCI repository. With respect to the results of accuracy, precision, sensitivity, specificity and false positive rate the efficiency of each algorithm is measured and compared. These techniques are coded in python and executed in Spyder, the Scientific Python Development Environment. Experimental results show that Random Forest obtained the best accuracy, recall, CV score, and F1 score among the six classification techniques for all three datasets. After comparing the experimental results with alternative schemes that used with three different dataset, performance comparison shows that Random Forest outperformed the other five machine learning techniques with the best accuracy of 99.57%, 96.3% precision and 100% recall to predict the breast cancer.

Chapter 1

Introduction

In this chapter the basic impacts of breast cancer on women is briefly highlighted along with the steps of early prediction of cancer in the form of machine learning. The motivation of the project implies that what was the background and why we focus on this field. The project objective and the full outline of the project is also elaborated in this chapter.

Accurate identification of important information from medical data is challenging in bio-science. The diagnosing of the sickness could be a crucial task in bio-science. There is an enormous quantity of medical diagnosing information accessible that can be used for quick and correct diagnosis of various type of health issues. Manual identification of diseases is vulnerable to human errors, unwanted biases, and time waste. Such delay and errors can be fatal for cancer patients. Data suggests that the females are diagnosed more with breast cancer compared to all carcinoma [1]. Recent statistics within the United States reports that 282,000 females will be diagnosed with breast cancer and 43,000 ladies will die from breast cancer in a year. Breast cancer is an abnormal growth of some cells within any part of a breast. Several diagnostic process is available for proper identification of carcinoma. Mammogram has been proposed to diagnose carcinoma [2]. Ultrasound [3] is also a very efficient technique for the identification of carcinoma. In this process, the wave of sound is distributed within the specific area of body to observe the condition inside. Positron emission tomography (PET) [4] imaging illustrates F-fluorodeoxyglucose which allows doctors to get knowledge of the tumour's position within the human body. It is created specifically for the recognition of traces for radio-labelled cancer. Flexman et al. [5] used dynamic tomography with spread of cancer cells. Elastography [6] is a recent technique which supports imaging technology which can be used when carcinoma tissue supports substantial than the adjacent regular functional tissue. In recent years, neural network [7], different types of com-

putational intelligence techniques [8], predictive data mining [9], and support vector machine (SVM) and ensemble classification [10] technologies are designed in many medical predictions. Current machine learning methods to detect breast cancer uses different types of Naive Bayes, SVM, KNN, etc., and Xu et al. [11] reported the highest 98.53% accuracy on the University of Wisconsin Hospital dataset [12]. However, there is still room for improvement for the carcinoma detection performance.

In this paper, we have analysed six machine learning classifier models: (i) Decision Tree, (ii) K Nearest Neighbours (KNN), (iii) Random Forest, (iv) Logistic Regression, (v) Extra Trees and (vi) Support Vector Machine (SVM). Then we also performed analysis with Ensemble Technique. We applied these models on 3 datasets to compare performance of the classifier that is best suited to predict breast carcinoma at the very initial stage. We also compare our experimental results with alternative schemes that used a similar dataset.

1.1 Motivation of the Project

Many people are being affected from breast cancer. Causing of this disease depends on many factors and cannot be simply determined. In addition, the identification method that determines whether or not the cancer is benign or malignant additionally needs an excellent deal of effort from a doctors and physicians. Once many tests are concerned within the identification of breast cancer, like clump thickness, uniformity of cell size, uniformity of cell form, etc., the ultimate result could also be troublesome to get, even for doctors. This has given an increase the previous few years to the utilization of machine learning and computing generally as diagnostic tools. Robotics are taking part as a necessary role in operational rooms. Also, the skilled systems are conferred within the intensive treatment rooms. In turn, using another side of Artificial intelligence for breast cancer designation isn't unworthy. It's reported that breast cancer illness is that the second commonest cancer that affects girls, and was the rife cancer within the world by the year of 2002. This cancer may be a quite common sort of cancer among girls and therefore the second highest reason behind cancer death. With the uncontrolled division of one cell inside the breast leads to beginning to the breast cancer which results in a visible mass, called a tumour. The tumour can be either benign or malignant. The correct designation in determinant whether or not the tumour is benign or malignant may result in saving lives. Therefore, the necessity for precise classification within the clinic may be an explanation for specialists and doctors. This importance of artificial intelligence has been actuated for the last twenty five years, once scientists began to understand the

quality of taking bound selections to treat specific diseases. The employment of machine learning and data processing as tools in diagnosing becomes effective and one amongst the crucial diseases in medicines wherever the classification task plays a really essential role is that the diagnosis of breast cancer. Therefore, machine learning techniques will facilitate doctors to create correct identification for breast cancer and make the proper classification of being benign or malignant tumor. There is little question that analysis of information taken from the patient and selections of doctors and specialists are the foremost necessary factors within the identification; however knowledgeable systems and artificial intelligence techniques like machine learning for classification tasks, conjointly facilitate doctors and specialists in a great deal.

This project aimed to compare different classification algorithms significantly to predict a benign tumor from malignant cancer in breast cancer dataset. We aim to investigate different machine learning techniques and will use several algorithms and apply on breast cancer dataset. The focused machine learning techniques are Support Vector Machine, K-nearest neighbor, logistic regression, Decision Tree, Random Forest, Extra tree and Ensemble Method. After primary study on these mentioned techniques, their result will be analyzed.

1.2 Objectives of the Project

The objective of this project is to analyze different supervised machine learning techniques to predict breast cancer using publicly available datasets. To achieve this objective, we have identified the following specific aims:

- I Principal Component Analysis (PCA) will be used to identify which features are most helpful in predicting malignant or benign cancer and find general trends that may aid in model selection and hyper parameter selection
- II Select and implement a set of supervised classifiers for breast cancer prediction.
- III To compare and evaluate the selected classifiers to identify which classifier is best suited to predict breast cancer at the very initial stage in terms of different metrics.

1.3 Outline of the Report

The rest of the book is organized as follows:

Chapter 2 describes different machine learning techniques. This chapter gives a clear overview of all the processes and sections of machine learning. The Artificial Neural Network has also been described in this section.

Chapter 3 describes the previous contributions in this field. It describes different algorithms regarding the predictive models for breast cancer prediction. It also describes the most recent works in this field. The limitations of this field are also described in this chapter. The pseudocode and the mechanism of the machine learning techniques are also described briefly in this chapter.

Chapter 4 states the proposed model of our research; it affirms the data set we used in our research, the predictive models we selected, and how we generated results for both before and after applying Principal Component Analysis. The description of the three data sets is also there. The pre-processing method of the three data sets and the performance regarding all machine learning techniques we used are shown here. It described the system implementation. It talks about sub-sections such as Train Test split; the ratio in which the dataset used in our research was split into training and testing models, feature selection and gives a brief account of PCA.

Chapter 5 describes the experimental settings and results. A brief account of the performance metrics used in our research and the results have been described in this chapter. The performance matrix, model performance on the basis of three datasets and comparison is also described in this chapter.

Chapter 6 summarizes our research and also highlights the limitations of our research. A brief account of the future works or steps we intend to take to improve our models or research is also stated here.

Chapter 2

Literature Review

This chapter describes different machine learning techniques. It gives a clear overview of the processes and sections of machine learning. The Artificial Neural Network(ANN) has also been described in this section.

Machine learning is a part within artificial intelligence which belongs to the science and engineering of making intelligent machines. Automated knowledge acquisition focused by machine learning through the design and implementation of algorithms where empirical data is required by algorithms. Basically, the techniques for learning of a machine is taught by machine learning depending on the use of probability. Figure 2.1 shows different branches of machine learning.

Supervised learning: In supervised learning, starting with the datasets which contains training examples, an algorithm can identify data associated level. It does it by running data through a learning algorithm. The goal of supervised learning is, correctly identify the new data given to it through the supervised learning and using the previous data set and learning algorithms can learn the technique to identify the data. The algorithms operating below supervised learning takes the inputs that the output is already known for the reason in order that the algorithms will create the machine to find out by holding it compare the particular output with the already known output to test for to any extent further errors. The machine is then modeled consequently. The famous supervised learning algorithms include classification, gradient boosting, prediction and regression. Then the model is modified by it consequently. With such algorithms, a machine creates a use of supervised learning to try and do the prediction of label values on unlabeled information by exploitation appropriate patterns. Supervised learning finds the appliance in such areas wherever the longer term events are expected through the historical information.

Unsupervised learning: Unsupervised learning studies how systems will learn to represent specific input patterns in a manner that reflects the applied math structure of the assortment of input patterns. By contrast with supervised learning or reinforcement learning, there aren't any express target outputs or environmental evaluations related to every input; rather the unattended learner brings in contact previous biases on what aspects of the structure of the input ought to be captured within the output. A specific output is not having by unsupervised learning. Finding the structures and patterns in the data is aimed by the learning agent.

Semi-Supervised learning: Under this machine learning sort, the machine is formed capable of learning each labelled and untagged information for the coaching purpose. This particularly involves training the machine through a tiny low quantity of labelled data at the expense of an oversized quantity of untagged data. This can be for the rationale that untagged information are economical and straightforward to assemble. This sort of machine learning is employed oftentimes with the algorithms like classification, prediction and regression. Further, this sort of learning is employed within the field wherever the price of an associated labeling is splurging to create thanks to a totally labelled coaching method. The celebrated application of semi-supervised learning is face recognition through a digital camera.

Reinforcement learning: Under this machine learning sort, the machine learning algorithms run through the trial and error approach to form positive of the actions that offer the simplest results and it finds applications within the field of play, navigation, and artificial intelligence. Is usually used for artificial intelligence, gaming, and navigation. There are 3 elements that employment primarily below this machine learning sort - the agent, learner, the atmosphere with that the agent do the interaction and also the actions that the agent is meant to try and do. The entire objective of reinforcement learning is to form the agent choose actions which will facilitate to get maximized reward over the desired amount of time. Therefore the plan is evident that the reinforcement helps the machine learn the simplest policy to figure with to allow best results.

Collaborative learning: Recommendations generate through a technique which is known as collaborative filtering which is a primary type of recommender system. Among the large number of choices and based on comparison of preferences between users it helps the users to find item of relevance. Collaborative filtering is domain agnostic. It is an unsupervised learning

Clustering: Structure in collections of data where no specific structure previously existed is discovered by clustering algorithm, is an unsupervised learning. Through the examining different properties of the input data the clusters, naturally occur in data is

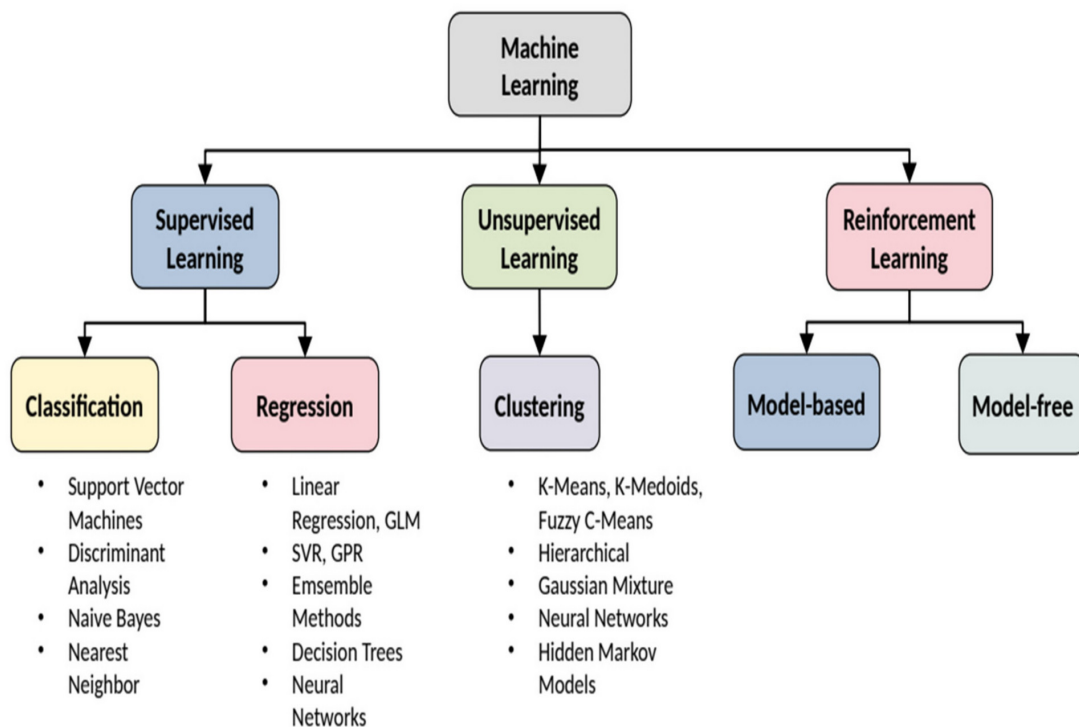


Figure 2.1: This figure shows the branches of machine learning.

discovered by clustering algorithm. Clustering is often used for dividing large amount of data into smaller group and tuning analysis for each group, which belongs to exploratory analysis.

Classifications: Classification belongs to supervised learning which requires training with data that has known labels. Application involving classification like by train using a set of spam and non-spam messages System will eventually learn to detect unwanted email. Through the training of previous records system will learn to identify the risk. Overall, the branches of machine learning can be identified from the mentioned picture.

Previously, research regarding classification and prediction of breast cancer has been carried out using several data mining techniques. Classification and agglomeration are 2 wide used ways in information mining [13]. Agglomeration or clustering ways aim to extract information from data set to get teams or clusters and describe the information set. Classification also known as supervised learning in machine learning, aims to classify unknown things supported learning existing patterns and classes from the information set and after predict future things. The training set, that is employed to build the classifying structure, and therefore the take a look at set, that tends to assess the classifier, are ordinarily mentioned in classification tasks [14].

Furthermore, essential progress has been carried out when it comes to breast cancer sur-

vivability prediction using labeled, unlabeled, and pseudo-labeled patient data. Prognostic studies of breast cancer survivability have been aided by machine learning algorithms, which can predict the survival of a particular patient based on historical patient data.

Neural networks and related techniques have a vast contribution when it comes to predicting breast cancer. Over the past few decades, Artificial Neural Networks have been employed increasingly by more and more researchers, and become an active research area [15]. ANNs have afforded numerous successes with great progress in Breast Cancer classification and diagnosis in the very early stages [16]. A typical ANN model is made up of a hierarchy of layers: input, hidden and output layers. Extensive research had been done with backpropagation artificial neural network (BP-ANN) method and its variations in breast cancer diagnosis [17]. The technique, however, has some limitations such as no guarantee to global optima, a lot of tuning parameters, and long training time. Single Hidden Layer Neural Networks (SFLN) was proposed by Huang and Babri [18] to tackle the mentioned problems with tree steps learning process that called extreme learning machine (ELM). Standard [19] and best parameterized [20] ELM model were proposed for breast cancer early prediction. Results showed that it generally gave better accuracy, specificity, and sensitivity compared to BP ANN. However, most existing works focus on prediction performance with limited attention with medical professional as end user and applicability aspect in real medical setting. With due respect to all related work referred above, this project compares the performance of the algorithms: Decision Tree, K Nearest Neighbours (KNN), Random Forest, Logistic Regression, Extra Trees and Support Vector Machine (SVM) using three different dataset in both diagnosis and analysis to make decisions. The goal is to achieve the most efficient algorithm to help us predict breast cancer at the very initial stages. To do so, we compare efficiency and effectiveness of those approaches in terms of certain criteria such as accuracy, precision, specificity confusion and normalized matrix, recall and f1-score.

2.1 Summary

The goal of supervised learning is to correctly identify the new data given to a machine. It does it by running data through a learning algorithm. The algorithms operating below supervised learning takes the input where the output is already known for the reason in order that the machine can hold it compare the particular output with the already known output. The celebrated application of semi-supervised learning is face recognition through a digital camera. Under this machine learning sort, the machine learning algorithms run through the trial and error approach. It finds applications within the field of play, navigation, and artificial intelligence. Neural networks and related techniques have a vast contribution when it comes to predicting breast cancer. Extensive research had been done with the backpropagation artificial neural network (BP-ANN) method.

Chapter 3

Background

This chapter describes different algorithms regarding the predictive models for breast cancer prediction. The pseudocode and the mechanism of the machine learning techniques are also described briefly in this chapter. In this section, we discuss the supervised machine learning algorithms which are analyzed in this paper. We also discuss principal Component Analysis (PCA) which is used for data processing.

3.1 Supervised Learning Algorithms

In supervised learning, known information is used to predict future unknown classes. Regression and classification are common ways in supervised learning category [5]. In this project, we evaluate the following six machine supervised learning algorithms.

3.1.1 Support Vector Machine

Support Vector Machine (SVM) is a supervised machine learning algorithmic rule which might be used for each classification or regression challenges. However, it's principally utilized in classification issues. In this algorithmic rule, we plot each data item as a point in n-dimensional space where n is number of features one has with the value of each feature being the value of a particular coordinate [21]. Often researchers tend to plot every knowledge item as some extent in n-dimensional area with the worth of every feature being the worth of a selected coordinate. Then, to perform classification by finding the hyper-plane that differentiate the two categories. It is a non probabilistic binary linear classifier, however are often manipulated during a manner that it will perform non-linear and probabilistic classification also, creating it versatile algorithmic

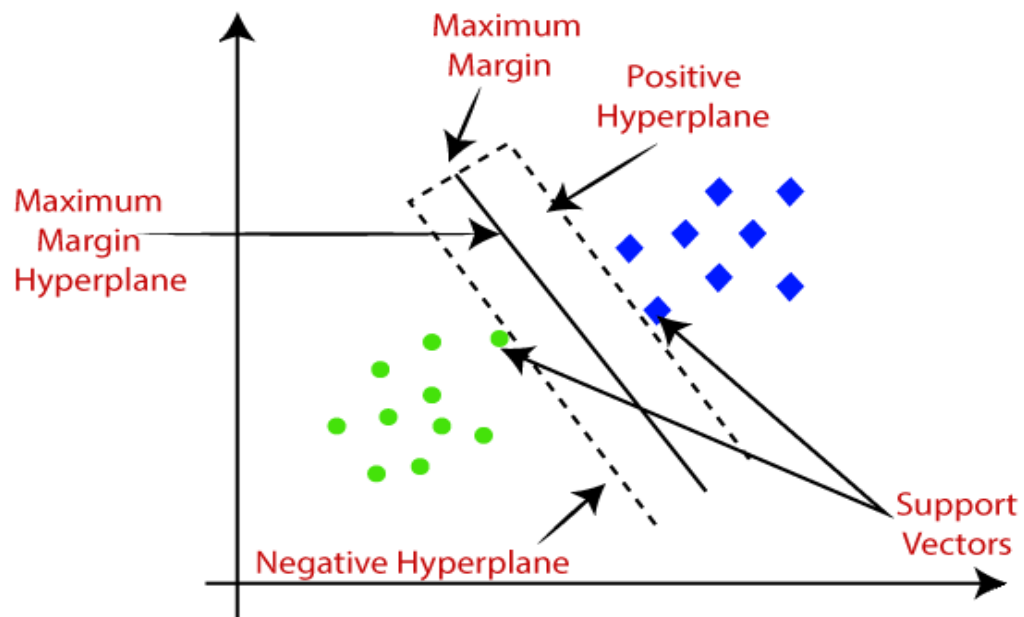


Figure 3.1: Support Vector Machine

program. An SVM model could be an illustration of the instances as points in area mapped, so they will be categorized and divided by a transparent gap. New instances are then mapped into the identical area and foreseen that within which class it would be supported which aspect of the gap they fall in. The advantage of SVM is that the indisputable fact that it is effective in high dimensional areas [22]. It is conjointly memory efficient since it uses a set of coaching points within the call operate. Then, we perform classification by finding the hyper-plane that differentiates the two classes as shown in the Figure 3.1.

3.1.2 K Nearest Neighbours (KNN)

The k-nearest neighbor's algorithmic program is one of the simplest machine learning algorithms. It has merely supported the concept that objects that are 'near' every alternative can additionally have similar characteristics [23]. So if it can recognize the characteristic options of one of the objects, it will be additionally predicted for its nearest neighbor. KNN is associate improvisation over the nearest neighbor technique. It is based mostly on the plan that any new instance will be classified by the majority vote of its 'k' neighbors, - wherever k is a positive number.

KNN is one amongst the foremost easy and simple data processing techniques. It is known as Memory-Based Classification as the coaching examples have to be in the memory at run-time. Once handling continuous attributes the distinction between the attributes is calculated using Euclidean distance. A serious drawback dealing with the Euclidean distance formula is that the big values frequency swamps the smaller ones.

When KNN is employed for classification, the output is calculated because the category with the very best frequency from the K-most similar instances. Every instance in essence votes for their class and therefore the class with the foremost votes is taken for the prediction.

Class probabilities is calculated because the normalized frequency of samples that belong to every class within the set of K most similar instances for a new data instance. For instance, during a binary classification problem (class is zero or 1):

$$(\text{class}=0) = \text{count}(\text{class}=0) / (\text{count}(\text{class}=0) + \text{count}(\text{class}=1))$$

If using K and having an even number of classes (e.g, 2), it is a good idea to choose a K value with an odd number to avoid a tie. And the inverse, use an even number for K when having an odd number of classes.

3.1.2.1 Pseudocode of K-Neighbors

1. Load the training and test data
2. Choose the value of K
3. For each point in test data:
 - *find the Euclidean distance to all training data points*
 - *store the Euclidean distances in a list and sort it*
 - *choose the first k points*
 - *assign a class to the test point based on the majority of classes present in the chosen points*

4. End

3.1.3 Decision Tree

Decision tree is a supervised learning rule that's used for classification and regression. It works by splitting the info into 2 or additional subsets supported the values of input variables. A value operate or cacophonous criterion is employed to see the most effective split among all the split points. The info is split recursively into teams till the leaves contain just one sample. During this model, associate degree optimized version of the CART rule is employed to implement the choice tree classifier. Call trees are straightforward to interpret and perceive, compared to different classification algorithms. Moreover, call trees need very little preprocessing as outliers don't have an effect on the performance. Moreover, they are not use the Euclidean distance. Hence, feature scaling isn't needed. Also, feature scaling may lead to wrong assumptions being tacit since the values would be modified. Call trees will handle each categorical and numerical variables as input; therefore it's acceptable for this model, since the info set contains each variable varieties. During this model, the link between the feature variable and target variable is complicated and high non-linear. Therefore, a call tree contains a larger likelihood of outperforming lin-ear models like provision regression. While call tree have many benefits, they even have some disadvantages. One is that, call trees will cause over fitting by creating a tree that's too complicated and thus doesn't predict well on new information. Finally, since call trees are greedy algorithms, the optimum tree isn't essentially came back.

3.1.3.1 Assumptions while creating Decision Tree

- The Figure 3.2 shows that the whole training set is considered as the root.
- Feature values are preferred to be categorical. If the values are continuous, then they are discretized prior to building the model.
- Records are distributed recursively on the basis of attribute values.
- Order to placing attributes as root or internal node of the tree is done by using some statistical approach.

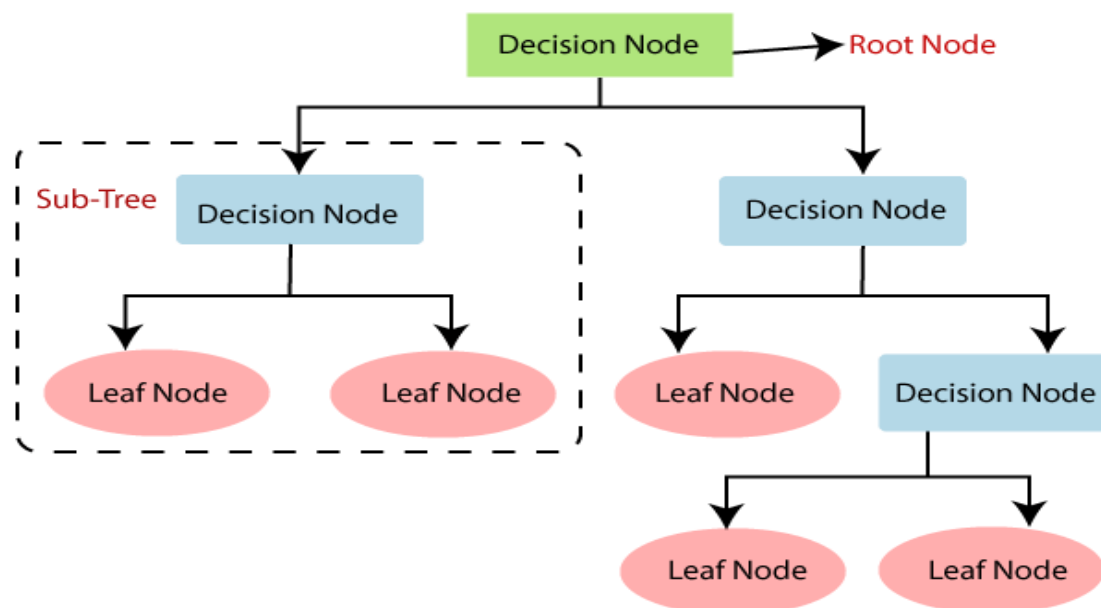


Figure 3.2: Decision Tree

3.1.4 Random Forest

Random Forest is a popular machine learning algorithm that belongs to the supervised learning technique [24]. It can be used for both Classification and Regression problems in ML. It is based on the concept of ensemble learning, which is a process of combining multiple classifiers to solve a complex problem and to improve the performance of the model. As the name suggests, Random Forest is a classifier that contains a number of decision trees on various subsets of the given dataset and takes the average to improve the predictive accuracy of that dataset. The Figure 3.3 shows that instead of relying on one decision tree, the random forest takes the prediction from each tree and based on the majority votes of predictions, and it predicts the final output.

3.1.4.1 Important Features of Random Forest

Diversity: Not all attributes/variables/features are considered while making an individual tree, each tree is different.

Immune to the curse of dimensionality: Since each tree does not consider all the features, the feature space is reduced.

Parallelization: Each tree is created independently out of different data and attributes. This means that we can make full use of the CPU to build random forests.

Train-Test split: In a random forest we don't have to segregate the data for train and test as there will always be 30% of the data which is not seen by the decision tree.

Stability: Stability arises because the result is based on majority voting/ averaging.

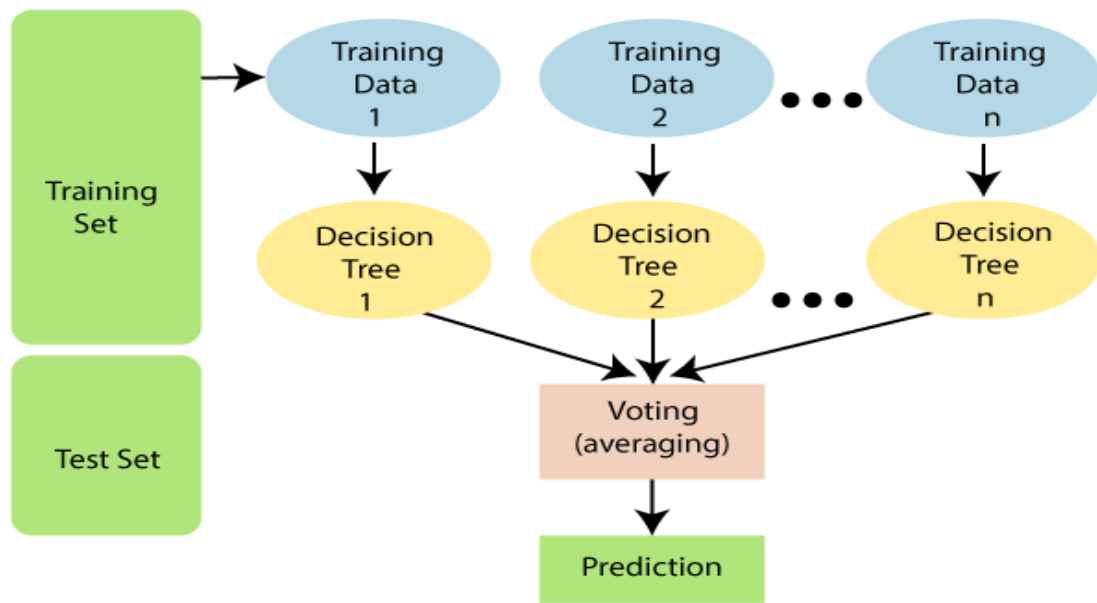


Figure 3.3: Random Forest

3.1.4.2 Assumptions for Random Forest

- There should be some actual values in the feature variable of the dataset so that the classifier can predict accurate results rather than a guessed result.
- The predictions from each tree must have very low correlations. Forests are non-parametric and can thus handle skewed and multi-modal data as well as categorical data that are ordinal or non-ordinal.

3.1.4.3 Pseudocode of Random Forest

1. In Random forest n number of random records are taken from the data set having k number of records.
2. Individual decision trees are constructed for each sample.
3. Each decision tree will generate an output.
4. Final output is considered based on Majority Voting or Averaging for Classification and regression respectively.

3.1.5 Logistic Regression

After linear regression, logistical regression is the most famous machine learning algorithm [25]. Linear regression and logistic regression are similar in many ways. But what they are used for is the biggest distinction. Algorithms for linear regression are used to predict values, but logistic regression is used for classification tasks. Logistic rule may be a supervised rule that trains the model by taking input variables and a target variable. In logistical rule the output or target variable may be a categorical variable, in contrast to regression toward the mean, and is therefore a binary classification rule that categorizes a knowledge purpose to one of the categories of information. The general equation of logistic regression is:

$$\text{logit}(p) = b^0 + b^1 X^1 + b^2 X^2 + b^k X^k$$

Logistic regression measures the link between the variable quantity, the output, and therefore the freelance variables, the input. It uses 1.2 penalty for regularization. Supply regression formula conjointly uses an equation with freelance predictors to predict a worth. The expected worth are often anyplace between negative eternity to positive eternity. The resultant chances are then born-again to binary values zero or one by the supply perform, conjointly referred to as the sigmoid function. The Sigmoid perform takes any real-valued variety and maps it into a worth between the vary 0-1 excluding the bounds themselves. Afterwards, a threshold classifier transforms the result to a binary worth. supply regression is that the input options ought to be freelance of every alternative. One variable ought to have very little or no co-linearity with the opposite variable.

3.1.6 The Extra Trees

Extremely Randomized Trees Classifier (Extra Trees Classifier) is a type of ensemble learning technique which aggregates the results of multiple de-correlated decision trees collected in a “forest” to output it’s classification result [26]. In concept, it is very similar to a Random Forest Classifier and only differs from it in the manner of construction of the decision trees in the forest. Figure 3.4 shows that each Decision Tree in the Extra Trees Forest is constructed from the original training sample. Then, at each test node, each tree is provided with a random sample of k features from the feature-set from which each decision tree must select the best feature to split the data based on some mathematical criteria (typically the Gini Index). This random sample of features

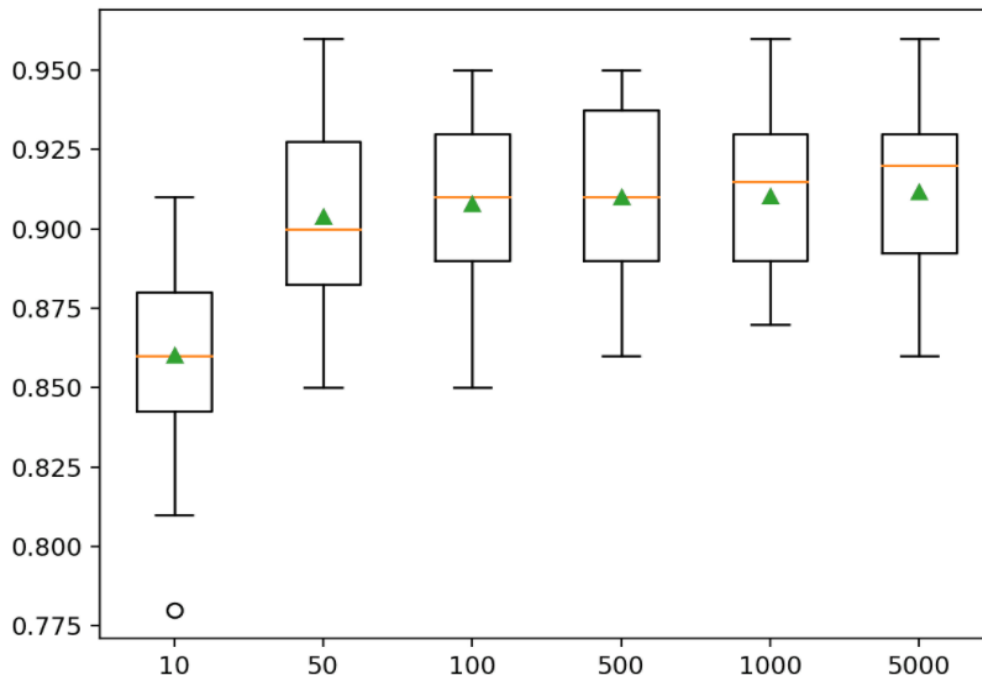


Figure 3.4: Extremely Randomized Trees Classifier

leads to the creation of multiple de-correlated decision trees. To perform feature selection using the above forest structure, during the construction of the forest, for each feature, the normalized total reduction in the mathematical criteria used in the decision of feature of split (Gini Index if the Gini Index is used in the construction of the forest) is computed. This value is called the Gini Importance of the feature. To perform feature selection, each feature is ordered in descending order according to the Gini Importance of each feature and the user selects the top k features according to choice.

3.2 Feature Selection

Within the fields of machine learning high dimensional data analysis could be a challenge for re-researchers and engineers. Solving drawback by removing immaterial and redundant data through an efficient way provided by feature selection [27], which might cut back the computation time, improve learning accuracy, and facilitate a higher understanding for the learning model or data. During this study, we have a tendency to discuss many frequently used analysis measures for feature choice, and so survey supervised, unsupervised, and semi-supervised feature selection strategies, that are wide applied in machine learning issues, like classification and clustering. Variable selection

or attribute selection is known as feature selection. Automatic selection of attributes in the data that are most relevant to the predictive modeling problem. Dimensionality reduction is completely different from feature selection. Each strategies request to scale back the quantity of attributes within the dataset, however a dimensionality reduction methodology do thus by making new combination of attributes, wherever as feature selection strategies embrace and exclude attributes present within the data while not ever changing them. An accurate predictive model is created by feature selection methods. Identifying and removing unneeded can be done by using the feature selection method. There are three general classes of feature selection algorithms: filter methods, wrapper methods and embedded methods.

Filter method: Statistical measure to assign evaluation to each feature applied by the filter feature selection methods. The features are hierarchic by the score and either selected to be unbroken or off from the dataset. The methods are typically univariate and take into account the feature severally, or with reference to the variable quantity.

Wrapper method: Wrapper ways think about the selection of a group of options as a search drawback, wherever completely different features are ready, evaluated and compared to different mixtures. A predictive model us accustomed valuate a mixture of combinations and assign a score supported model accuracy. The search method is also organized like a best-first search, it should random like a random hill-climbing formula, or it should use heuristics, like forward and backward passes to feature and take away options.

Embedded method: Embedded strategies learn that options best contribute to the accuracy of the model whereas the model is being created. the foremost common kind of embedded feature choice methods are regularization methods. Additional constraints into the optimization of a predictive algorithm is introduced by Regularization methods are also called penalization methods. That bias the model toward lower complexity.

3.3 Principal Component Analysis

The main idea of principal component analysis (PCA) is to cut back the dimensionality of a data set consisting of the many variables related with one another, either heavily or gently, whereas holding the variation present within the data set, up to the utmost extent [28] The identical is finished by remodeling the variables to a replacement set

of variables, that are referred to as the principal elements (or merely, the PCs) and are orthogonal, ordered specified the retention of variation present within the original variables decreases as we tend to move down within the order. So, during this method, the first principal element retains most variation that was gift within the original elements. The principal elements are the Manfred Eigen vectors of a co variance matrix, and therefore they're orthogonal. Importantly, the dataset on that PCA technique is to be used should be scaled. The results are sensitive to the relative scaling. As a layman, it's a technique of summarizing information. Imagine some wine bottles on a board, every wine is delineate by its attributes like color, strength, age, etc. However, redundancy can arise as a result of several of them can live connected properties. Thus, what PCA can neutralize this case is summarize every wine within the stock with less characteristics. Intuitively, PCA will provide the user with a lower-dimensional image, a projection or "shadow" of this object once viewed from its most informative viewpoint.

3.4 Train-Test Split

Data, in machine learning, in most scenarios are split into training data and testing data (and sometimes to three: train, validate and test), and fit the model on the train data, in order to make predictions on the test data. Training dataset is a part of the actual dataset that we use to train the model. The model sees and learns from this data. Test data, on the other hand, is the sample of data used to provide an unbiased analysis of a final model fit on the training dataset. The Test dataset provides the ideal standard used to evaluate the model. It is used once the model is completely trained [29].

Splitting the dataset into training, validation testing sets can be determined on two categories. Firstly, it depends on how much the total number of samples in the data and second, on the actual model the user is training. Some models need efficient or large data to train upon, so in that case one could optimize for the larger training sets. Models with very few hyper parameters are estimated to be easy to validate and tune, so one can possibly reduce the size of your validation set. However, given the model has many hyper parameters, the user would want to have a large validation set as well.

3.5 Summary

Support Vector Machine (SVM) is a supervised machine learning algorithmic rule which might be used for each classification or regression challenge. An SVM model could be

an illustration of the instances as points in the area mapped, so they will be categorized and divided by a transparent gap. The k-nearest neighbor algorithm is based mostly on the plan that any new instance will be classified by the majority vote of its 'k' neighbors. It is known as Memory-Based Classification as the coaching examples have to be in the memory at run-time. The Decision tree may be a supervised learning rule that's used for classification and regression. Call trees are straightforward to interpret and perceive, compared to different classification algorithms. Random Forest is a classifier that contains a number of decision trees on various subsets of the given dataset. Logistic regression measures the link between the variable quantity, the output, and the free-lance variables, the input. The supply regression formula conjointly uses an equation with freelance predictors to predict a worth. The first assumption of supply regression is that the input options ought to be freelance. Extra Trees Classifier is a type of ensemble learning technique that aggregates the results of multiple de-correlated decision trees collected in a "forest". To perform feature selection, each feature is ordered in descending order according to the Gini Importance of each feature and the user selects the top k features according to choice. High-dimensional data analysis could be a challenge for researchers and engineers. Identifying and removing unneeded can be done by using the feature selection method. There are three general classes of feature selection algorithms: filter methods, wrapper methods, and embedded methods. Principal component analysis (PCA) cuts back the dimensionality of a data set. The principal elements are the Manfred Eigenvectors of a covariance matrix, and therefore orthogonal. Importantly, the dataset on that PCA technique is to be used should be scaled. Data, in machine learning, is split into training data and testing data. The training dataset is a part of the actual dataset that we use to train the model. Test data is the sample of data used to provide an unbiased analysis of a final model fit on the training dataset.

Chapter 4

Proposed System

This chapter states the proposed model of our research; it affirms the dataset we used in our research, the predictive models we selected, and how we generated results for both before and after applying Principal Component Analysis. The description of the three datasets is also there. The pre-processing method of the three datasets and the performance regarding all machine learning techniques we used is shown here.

4.1 Dataset

In this work, we used 3 datasets for performance analysis of the model. The datasets are publicly available. The datasets are (i) Dataset1: Wisconsin Breast Cancer Dataset [12], (ii) Dataset2: Breast Cancer Wisconsin (Diagnostic) Dataset [30], (iii) Dataset3: Breast Cancer Dataset of University Medical Centre, Institute of Oncology, Ljubljana, Yugoslavia [31].

4.1.1 Dataset1

Dataset1 [12] used for this work is publicly available and was created by Dr. WilliamH. Wolberg, physician at the University Of Wisconsin Hospital at Madison, Wisconsin, USA. The samples were analyzed based on a digital scan. This dataset contains 699 instances where the cases are either non-cancerous or infectious. Among all instances; 65.50% are from benign, and 34.50% are from malignant class. The features within the Dataset are shown in Table 4.1 where attribute, and domain are shown. The benign cases are set as a positive category and the malignant cases are set as a negative category in our analysis. The Dataset contains 11 columns, with the first column being the ID number,

the last column being the Class (benign or malignant). The nine attributes detailed in Table 4.1 are graded on an interval scale from a normal state of 1–10, with 10 being the most abnormal state.

Table 4.1: Attributes of Dataset-1

Attributes	Domain
Clump thickness	1-10
Uniformity of cell size	1-10
Uniformity of cell shape	1-10
Marginal adhesion	1-10
Single epithelial cell size	1-10
Bare nuclei	1-10
Bland chromatin	1-10
Normal nuclei	1-10
Mitoses	1-10

4.1.2 Dataset2

The Dataset2 [30] used for this work is publicly available and was created by Dr. William H. Wolberg, physician at the University Of Wisconsin Hospital at Madison, Wisconsin, USA. To create the dataset, Dr. Wolberg used fluid samples, taken from patients with solid breast masses and an easy-to-use graphical computer program called Xcyt, which is capable of performing the analysis of cytological features based on a digital scan. The program uses a curve-fitting algorithm to compute ten features from each one of the cells in the sample; then it calculates the mean value, extreme value and standard error of each feature for the image, returning a 30 real-valuated vector. The dataset contains 357 cases of benign breast cancer and 212 cases of malignant breast cancer. The dataset contains 32 columns, with the first column being the ID number, the second column being the diagnosis result (benign or malignant), and followed by the mean, standard deviation and the mean of the worst measurements of ten features. There were no missing values in the dataset. The thirty attributes detailed in Table 4.2 are graded on an interval scale from a normal state of 1–32, with 32 being the most abnormal state.

4.1.3 Dataset3

Dataset3 [31] consists of 286 instances with ten attributes. In this dataset, 29.72% instances are defined as malignant and the other 70.28% are defined as benign. This dataset contains categorical data. To fit this dataset in the proposed model, One-Hot

Table 4.2: Attributes of Dataset-2

Attributes	Domain
Radius mean	1-32
Texture mean	1-32
Perimeter mean	1-32
Area mean	1-32
Smoothness mean	1-32
Compactness mean	1-32
Concavity mean	1-32
Concave points mean	1-32
Symmetry mean	1-32
Fractal dimension mean	1-32
Radius Se	1-32
Texture Se	1-32
Perimeter Se	1-32
Area Se	1-32
Smoothness Se	1-32
Compactness Se	1-32
Concavity Se	1-32
Concave points Se	1-32
Symmetry Se	1-32
Fractal dimension Se	1-32
Radius worst	1-32
Perimeter worst	1-32
Area worst	1-32
Smoothness worst	1-32
Compactness worst	1-32
Concave pints worst	1-32
Symmetry worst	1-32
Fractal dimension worst	1-32

Encoding method is used. The class attribute of this dataset is defined as recurrence-events and no-recurrence-events. The nine attributes detailed in Table 4.3 are graded on an interval scale from a normal state of 1–10, with 10 being the most abnormal state.

4.2 Data Pre-processing

Data preprocessing is a process of preparing the raw data and making it suitable for a machine learning model. It is the first and crucial step while creating a machine learning model. When creating a machine learning project, it is not always a case that we come across the clean and formatted data. And while doing any operation with

Table 4.3: Attributes of Dataset-3

Attributes	Domain
Age	1-10
Menopause	1-10
Tumor size	1-10
Inv nodes	1-10
Node caps	1-10
Deg malig	1-10
Breast	1-10
Breast quad	1-10
Irradiat	1-10

data, it is mandatory to clean it and put in a formatted way. So for this, we use data preprocessing task. It is necessary to handle missing values, to process outliers, and to solve self-contradiction. We use mean of attribute to process absent data for a category. In addition, random choice of dataset is utilized to verify correct circulation of data. The number of variances of the first data that is calculated as the ratio between the variance of the residual data for the parts from one to nine; and therefore the variance of the initial data. Figure 4.1 shows the variance vs cumulative variance among 10 elements. PCA is applied to reduce the dimensionality of the feature columns. We got nine features within the data for dataset-1 [12]; therefore we needed to reduce the quantity of feature columns whereas maintaining the variance in data. We applied the variance as 0.95.

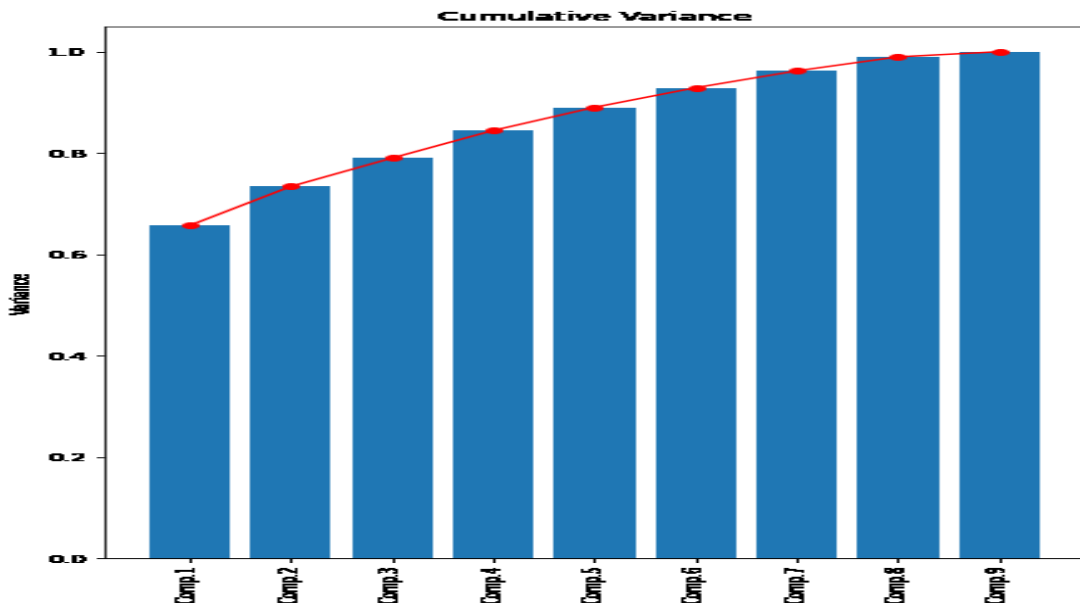


Figure 4.1: Cumulative Variance of Data

By applying PCA, we transform the present set of features into new set of reduced features which contains nine clump thickness, uniformity of cell size, uniformity of cell

shape, marginal adhesion, single epithelial cell size and bare nuclei.

4.3 Building Classification Models

With our aim being to predict whether the tumor is Benign(non-cancerous) or Malignant (cancerous), we have outlined a simple model to come with the most accurate predictions. The first objective was to attain a dataset of numerical values of various instances. Upon finalizing our dataset, we split the train-test ratio in order to train and test the six algorithms: Support Vector machine(SVM), K-Neighbors, Logistic Regression, Decision Tree(DT), Random Forest(RF) and Extra Trees. Feature selection in the form of PCA is used to reduced imensionality of the dataset. The models are trained again by means of training and testing after PCA is applied.

4.3.1 SVM

Parameter selection for kernel functions is important to the robust classification performance of SVM. It must be one of 'linear', 'poly', 'rbf', or 'sigmoid'. If a callable is given, it is used to pre-compute the kernel matrix from data matrices. It's important to start with the intuition for SVM with the special linearly separable classification case.

4.3.2 KNN

Here, K is the number of nearest neighbors which is the core deciding factor. In our mechanism, we found that K=20 is the best choice.

4.3.3 LR

LR does not really have any critical hyper parameters to tune. In our mechanism, we used L-BFGS as solver. Regularization (penalty) can be helpful for better performance; we have used penalty of 12.

4.3.4 DT

Decision tree complexity has a crucial effect on its accuracy. In our mechanism, the tree complexity is measured by one of the following metrics: the total number of nodes,

total number of leaves, tree depth and number of attributes used.

4.3.5 RF

For RF, have used bootstrap = true, criterion = gini, maximum depth = 10, maximum features = sqrt. Another important parameter for random forest is the number of trees (n-estimators). In addition, this should be increased until no further improvement is seen in the model. In our mechanism, n-estimators value is 10.

4.3.6 ET

Extra Trees implements a meta estimator that fits a number of randomized decision trees. The number of trees can be set via the n-estimators and we used 100. In our mechanism, parameter selection for ET are bootstrap =true, criterion = entropy, maximum depth = 40, and maximum features = sqrt.

4.4 Training and Testing Phases

The most critical factor affecting the success of machine learning is the training and testing process. An effective training process improves the quality of the developed system. Researchers divide datasets into two parts for training and testing. However, the separation process is done according to specific rules. The amount of training and test is the most critical factor in the success rate. If there is a high correlation between the features and the label, the Training-Test set is divided by 50%–50%. This means that 50% of all the data will be used for training and 50% for the test. However, if there is a fear of success falling, the rate of training can be increased. The training-testing ratio used in the literature varies according to the data structure. Less than 50% of the training data is not preferred because the test results will be negatively affected. After the machine learning model is trained according to the training data, it is also tested using the training data. The purpose of this is to determine how much data is learned. Performance evaluation procedures are performed according to specific criteria. These criteria vary according to the structure of the data. Once the training process is completed, the machine learning model tested with test data has never been seen before. The researcher evaluates the test performance according to the performance evaluation criteria. The research can be repeated by changing the training and test data in the training and testing process to avoid the situation of unstable data. In this case, the researcher

uses the average of performance values. We applied our model in 3 different datasets. For each, the whole dataset is divided into training and test data. The model is built based on training data. Test data is used to analyse the trained model. We used k fold cross-validation for our analysis. Cross-validation is a technique used to minimize over-fitting. In order to get more stable results and use all valuable data for training, a data set can be repeatedly split into several training one validation datasets. This is known as cross-validation. To validate the model performance, an additional test dataset held out from cross-validation, is normally used. In our study, we have used a $k=10$ to partition the data.

4.5 Ensemble classification

Ensemble methods consist of combining multiple techniques to solve the same task [32]. This approach was designed to overcome the weaknesses of single techniques and consolidate their strengths. Ensemble methods are now widely used to carry out prediction tasks (e.g, classification and regression) in several fields, including that of bio-informatics. Researchers have particularly begun to employ ensemble techniques to improve research into breast cancer, as this is the most frequent type of cancer and accounts for most of the deaths among women. This work exhibited an ensemble classification to predict the breast cancer using several machine learning techniques which are Support Vector Machine, Logistic Regression , Decision Tree , K-Nearest Neighbour, Random Forest , Extra Trees , Gauss NB and Neural Network. Results shows that ensemble framework is more accurate in contrast of proposed single classification system.

4.5.1 Bagging based Ensemble Classifier

Bagging is one of the Ensemble Construction techniques which is also known as Bootstrap Aggregation. Bootstrap establishes the foundation of Bagging technique. Bootstrap is a sampling technique in which we select “n” observations out of a population of “n” observations. But the selection is completely random, i.e, each observation can be selected from the original population so that each observation is equally likely to be selected in each iteration of the Bootstrapping process. After the Bootstrapped samples are formed, separate models are trained with the Bootstrapped samples. In the experiment the Bootstrapped samples are drawn from the training set and the sub-models are tested using the testing set. The final output prediction is combined across the predic-

tions of all of the sub-models. For the experimental purpose, a Decision Tree based classifier model is chosen.

4.5.2 Boosting based Ensemble Classifier

Boosting is a form of sequential learning technique. The algorithm works by training a model with the entire training set and subsequent models are constructed by fitting the residual error values of the initial model. In this way, Boosting attempts to give higher weight to those observations that were poorly estimated by the previous model. Once the sequence of the models is created the predictions made by models are weighted by their accuracy scores and the results are combined to create a final estimation. Models that are typically used in Boosting technique are XGBoost, GBM, ADABOOST, etc. ADABOOST is used for the experimental purpose.

4.5.3 Voting based Ensemble Classifier

Voting is one of the simplest ways in ensemble learning technique where we combine two or more algorithms to increase accuracy of the prediction model. It works by first creating two or more individual models from training dataset. A voting classifier mainly wrap the models into one model and average the predictions of the sub-models to make predictions for new data. By using voting classifier class we can create a voting ensemble model. The predictions of the individual model can be weighted, but specifying the weights for classifiers automatically or even heuristically is difficult. More advanced methods can learn how to best weight the predictions from sub-models.

In this proposed work, it is aimed to predict breast cancer using ensemble model of machine learning techniques. Here, ensemble model is using feature extraction techniques and voting technique to get the improved prediction. The performances of the ensemble and standalone models were evaluated using Accuracy, Precision, Recall, F-score and 10-fold Cross validation.

4.5.4 Ensemble Model

The proposed approach is a voting classifier which is one of the ensemble approach. Voting classification is a good strategy when one classifier algorithm defects can be advantage for another classifier. Voting combines the prediction outputs of the classifiers. The dataset is filtered by preprocessing the dataset. With the help of ranker algorithm

attributes having low ranks are omitted without reaching global minimum. The filtered dataset is used for each classifiers and combination of different classifiers to attain highest accuracy rate. The prediction outputs of each classifiers are combined and extremely predicted classes are chosen as class variables of test instances. All the individual classifiers are applied initially then voting of different classifiers are combined to improve prediction rate. Finally we analyze the results by using evaluation criteria and conclude which vote ensemble technique has the high accuracy rate.

By combining different combination of classifiers we built a new ensemble classifier by using voting strategy. In this approach we combine five classifiers out of six classifiers making in to six combinations. Finally we combine all six algorithms outputs to achieve high accuracy rate. Voting uses majority voting as combination rule which applies on these classifiers that increase percentage of accuracy. The data used in this study are obtained from the University of Wisconsin Hospitals [12], Breast Cancer Wisconsin (Diagnostic) Dataset [30] and Breast Cancer Dataset of University Medical Centre, Institute of Oncology, Ljubljana, Yugoslavia [31].

4.6 Experimental Setup

In this Project, we have split our dataset into 70%-30% ratio for training and test respectively (the first 400 instances for training while the next 169 instances for testing the model). Keeping in mind that training the model, making the machine learn, is vital, we have slotted 70% of the dataset to training. Out of the 70% dataset for training, we are keeping 63 percent for training and 7 percent for cross validation test. A round of cross-validation comprises separating a section of data into complementary subsets, performing the analysis on one sub set (the training set), and validating the analysis on the other subset (called the validation set or testing set). To reduce variability, in most methods multiple rounds of cross-validation are performed using different partitions, and the validation results are combined (e.g. averaged) over the rounds to give an estimate of the model's predictive performance.

4.7 Summary

In this work, we used 3 datasets for performance analysis of the model. The datasets are Wisconsin Breast Cancer Dataset, Breast Cancer Wisconsin (Diagnostic) Data Set, Breast Cancer Dataset of University Medical Centre, Institute of Oncology, Ljubljana,

Yugoslavia. Dataset1 [12] contains 699 instances where the cases are either non-cancerous or infectious. Dataset2 [30] contains 357 cases of benign breast cancer and 212 cases of malignant breast cancer. Each case is graded on an interval scale from a normal state of 1–10, with 10 being the most abnormal state. Dataset3 [31] consists of 286 instances with ten attributes. 29.72% instances are defined as malignant and the other 70.28% as benign. To fit this dataset in the proposed model, the One-Hot Encoding method is used. There were no missing values in the Dataset. Data preprocessing is the first and crucial step while creating a machine learning model. PCA is applied to reduce the dimensionality of the feature columns. We got nine features within the data for dataset-1; therefore we reduced the number of feature columns while maintaining the variance in data. An effective training process improves the quality of the developed system. Researchers divide datasets into two parts for training and testing. The training-testing ratio used in the literature varies according to the data structure. Ensemble methods consist of combining more than one single technique to solve the same task. This approach was designed to overcome the weaknesses of single techniques and consolidate their strengths. Bootstrapping is a sampling technique in which we select "n" observations. Bootstrap establishes the foundation of the Bagging technique. Boosting attempts to give higher weight to those observations that were poorly estimated by the previous model. The final output prediction is combined across the predictions of all of the sub-models.

Chapter 5

Experimental Results

In this chapter, the experimental settings and results are described. A brief account of the performance metrics used in our research and the results have been described in this chapter. The performance matrix, model performance on the basis of three datasets and comparison are also described in this chapter.

The next step after implementing machine learning models is to seek out how effective is that the model, i.e, how the models performed on the datasets. This is carried out by running the models on the test dataset which was set earlier. The test dataset comprised of 30% of the dataset for Breast Cancer prediction. 10-fold cross-validation was also used for Breast cancer prediction. In order to determine and compare the performances of the different algorithms, several metrics have been used.

5.1 Performance Metrics

Several performance metrics have been used to figure out the performance of the Machine Learning algorithms in our Project. For breast cancer prediction, if the target variable is 1(malignant), then it is a positive instance, meaning the patient has Breast cancer. And, if the target variable is 0 (benign), then it is a negative instance, stating that the patient does not have the cancer.

5.1.1 Confusion Matrix

Summarizing the performance of a classification algorithm is based on a technique which is known as confusion matrix [33]. It is arguably the easiest way to regulate the performance of a classification model by comparing how many positive instances

Table 5.1: Confusion Matrix

	Predicted Negative	Predicted Positive
Actual Negative	TN	FP
Actual Positive	FN	TP

are correctly/incorrectly classified and how many negative instances are correctly/incorrectly classified. In a confusion matrix, as shown in the table 5.1, the rows represent the actual labels while the columns represent the predicted level.

True Positives (TP): These are the occurrences where both the predictive and actual class are true (1), i.e., when the patient has complications (breast cancer in this case) and is also classified by the model to have complications.

True Negatives (TN): True negatives are the occurrences where both the predicted class and actual class are false (0), i.e., when a patient does not have complications and is also classified by the model as not having complications.

False Negatives (FN): These are occurrences where the predicted class is false (0) but actual class is true (1), i.e., case of a patient being classified by the model as not having complications even though in reality, they do.

False Positive (FP): False positives are the occurrences where the predicted class is true (1) while the actual class is false (0), i.e., when a patient is classified by the model as having complications even though in reality, they do not.

Normalized Matrix: Normalized Confusion Matrix represents results in a more efficient way. The results are similar to that of the confusion matrix. The values are distributed within the range of 0-1. An even distribution of data makes prediction easier.

Accuracy: Evaluation of classification models is done by one of the metrics called accuracy. Accuracy is the fraction of prediction. It determines the number of correct predictions over the total number of predictions made by the model. The formula of accuracy is:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (5.1)$$

Recall: It is a measure of the proportion of patients that were predicted to have the complications among those patients that actually have the complications. Recall can be calculated as follows:

$$Recall = \frac{TP}{TP + FN} \quad (5.2)$$

Precision: It is described as a measure of proportion of patients that actually have complications among those classified to have complications by the model. The formula for Precision is as follows:

$$Precision = \frac{TP}{TP + FP} \quad (5.3)$$

Specificity: Classifier's performance to spot negative results is related by Specificity. It is exactly the negative of Recall. It is a measure of the number of patients who are classified as not having complications among those who actually did not have the complications. Specificity is calculated as follows:

$$Specificity = \frac{TN}{TN + FP} \quad (5.4)$$

F1 Score: Weighted average of precision and recall is known as F1 score. Therefore false positives and false negatives are taken by this score into the consideration. Intuitively it is not as simple to grasp as accuracy; however F1 is typically additional helpful than accuracy. It is calculated as follows:

$$F1Score = \frac{P * R}{P + R} * 2 \quad (5.5)$$

Cross Validation: Cross Validation (CV) score is the score for a model for the desired k-fold cross validation. It is a statistical method used to estimate the skill of machine learning models. It is commonly used in applied machine learning to compare and select a model for a given predictive modeling problem because it is easy to understand, easy to implement, and results in skill estimates that generally have a lower bias than other methods.

5.2 Model Performance

A total of six classification algorithms are used - Decision Tree (DT), Random Forest (RF), Logistic Regression (LR), Extra Tree (ET) and Support Vector Machine (SVM) and K Neighbors(KN) Classifier have been applied on three different datasets. The algorithms have been applied after Principal Component Analysis (PCA). For each ex-

periment, the performance of the algorithms are measured using Accuracy, Precision, Recall, F1 Score and CV Score. The following demonstrates the results of different metrics for the algorithms to predict Breast Cancer with Principal Component Analysis on three individual dataset.

5.2.1 Performance measure for Dataset 1

Table 5.2 shows the performance of six algorithm for Wisconsin Breast Cancer Dataset which are measure by using Accuracy, Precision, Recall, F1 Score and CV Score [12]. It summarizes the performance metrics for all six models. For Dataset1 [12], we used 90% (629 instances) of the overall data to train all six models and the rest 10% (70 instances) for testing. The experimental results in shown in Table 5.2 ; it depicts that random forest performed best among the 6 machine learning techniques with the best accuracy of 99.57%. Among the others, accuracy of decision tree is 99.1% and SVM, LR, and Extra tree achieved 98% accuracy.

Table 5.2: Performance analysis for Wisconsin Breast Cancer Dataset

Method	Accuracy (%)	Precision (%)	Recall(%)	F1 Score	CV Score(%)
SVM	98	97	97	97	95.9
KNN	97	97	95	96.3	97
Logistic Regression	98	99	96	97	96.1
Decision Tree	99.1	96	100	98	95.2
Random Forest	99.57	96.3	100	98.2	96.3
Extra Tree	98	96	97	97	96.3

KNN got the lowest accuracy of 97% in our analysis. The precision for Random Forest is comparatively low but it achieved the highest recall, F1 score and CV score. Even though SVM and KNN achieved best precision of 97%, they have relatively lower recall; we see 97% recall for SVM, and 95% for KNN while the recall for random forest is 100%. Additionally, highest CV Score of 97 is achieved by KNN whereas random forest obtained 96.3% which is relatively higher. Overall, we can see that random forest performs best among these 6 models for Dataset1.

Figure 5.1 shows performance measure for Wisconsin Breast Cancer Dataset of six algorithm which are measure by using Accuracy, Precision, Recall, F1 Score and CV Score where random forest carried out best among the 6 machine learning techniques

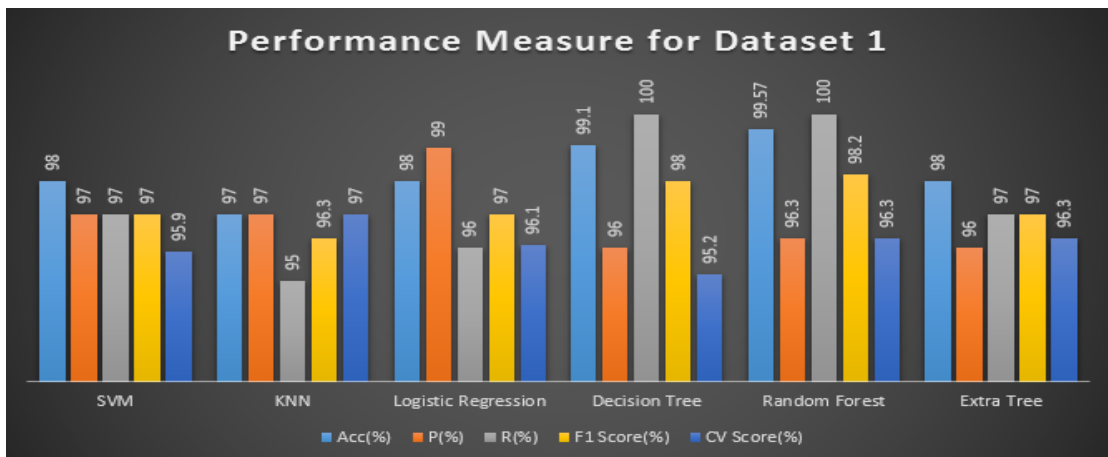


Figure 5.1: Performance Measure for Wisconsin Breast Cancer Dataset

with the best accuracy of 99.57%. Among the others, accuracy of decision tree is 99.1% and SVM, LR, and Extra tree achieved 98% accuracy. KNN got the lowest accuracy of 97% in our analysis.

5.2.2 Performance measure for Dataset 2

Table 5.3 shows the overall performance of six algorithm for Wisconsin Breast Cancer(Diagnostic) Dataset through the use of Accuracy, Precision, Recall, F1 Score and CV Score. For Dataset2 [30], we used 90% (512 instances) of the overall records to train all six models and the rest 10% (57 instances) for testing. The experimental outcomes in shown in Table 5.3; it depicts that Extra Tree performed exceptional some of the 6 machine learning strategies with the accuracy of 96.27%. Among the others, accuracy of KNN, LR, DT reached 95%,95.17%,94% and SVM reached second highest accuracy of 96.1%.In addition, RF carried out 95.25% accuracy.

Table 5.3: Performance Measure for Breast Cancer Wisconsin (Diagnostic) Dataset

Method	Accuracy (%)	Precision (%)	Recall(%)	F1 Score	CV Score(%)
SVM	96.1	96.54	96.23	97	98.13
KNN	95	95.41	97.39	96.47	97
Logistic Regression	95.17	97	95.03	96	97
Decision Tree	94	94.14	96.1	95	94.18
Random Forest	95.25	95.73	96	95	96
Extra Tree	96.27	95	98.13	96	96.17

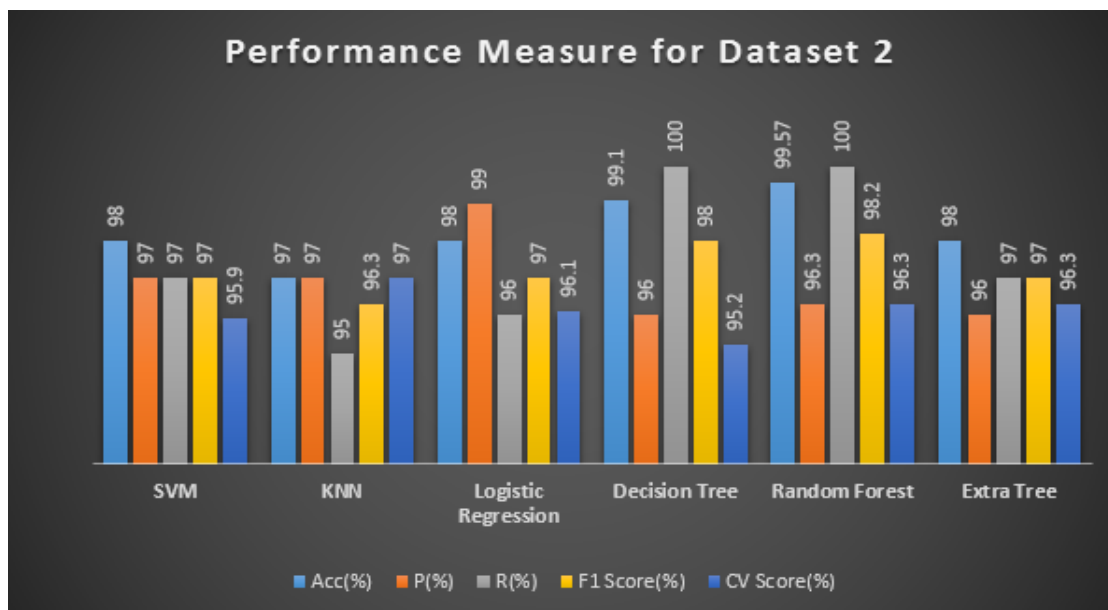


Figure 5.2: Performance Measure for Wisconsin Breast Cancer (Diagnostic) Dataset

The Precision of 95% for ET is comparatively low but it achieved the highest Recall of 98.13%. In addition, F1 Score of 96% and CV Score of 96.17%. Overall, we can see in Figure 5.2 that ET performs best among these 6 models for Dataset2 [30].

5.2.3 Performance measure for Dataset 3

Table 5.4 shows the overall performance of six algorithm for Breast Cancer Dataset of University Medical Centre through the use of Accuracy, Precision, Recall, F1 Score and CV Score. For Dataset3 [31], we used 90% (190 instances) of the overall records to train all six models and the rest 10% (21 instances) for testing. Dataset3 [31] contains categorical data. To fit this, dataset3 in the proposed model, One-Hot Encoding method is used. Machine learning models require all input and output variables to be numeric. This means that if the data contains categorical data, it must be encoded it to numbers before it can fit and evaluate a model. The two most popular techniques are an Ordinal Encoding and a One-Hot Encoding. For categorical variables where no ordinal relationship exists, the integer encoding may not be enough, at best, or misleading to the model at worst. Forcing an ordinal relationship via an ordinal encoding and allowing the model to assume a natural ordering between categories may result in poor performance or unexpected results (predictions halfway between categories). In this case, a one-hot encoding can be applied to the ordinal representation. This is where the integer encoded variable is removed and one new binary variable is added for each unique integer value in the variable. The experimental outcomes in proven in Table 5.4; it depicts that KNN performed best some of the 6 machine learning strategies with the accuracy of 85.19%. Among the others, SVM reached 84.27% second highest accuracy and LR, DT, RF, ET carried out 83.57%, 81.67%, 84% also 81.27% accuracy.

Table 5.4: Performance Measure for Breast Cancer Dataset of University Medical Centre

Method	Accuracy (%)	Precision (%)	Recall(%)	F1 Score	CV Score(%)
SVM	84.27	84	83.13	84.19	85
KNN	85.19	83	85	84	86.31
Logistic Regression	83.57	84	83.1	82	83
Decision Tree	81.67	80.46	81.42	81	82
Random Forest	84	83.61	83.61	83	84.57
Extra Tree	81.27	82	81.05	82.16	83

The Precision of 77% for ET is comparatively low but it achieved the highest. Overall, we can see that ET performs best among these 6 models for Dataset3 [31]. The Figure 5.3 shows that for considering the Recall, LR is performed 84% of the highest Recall among all other methods.

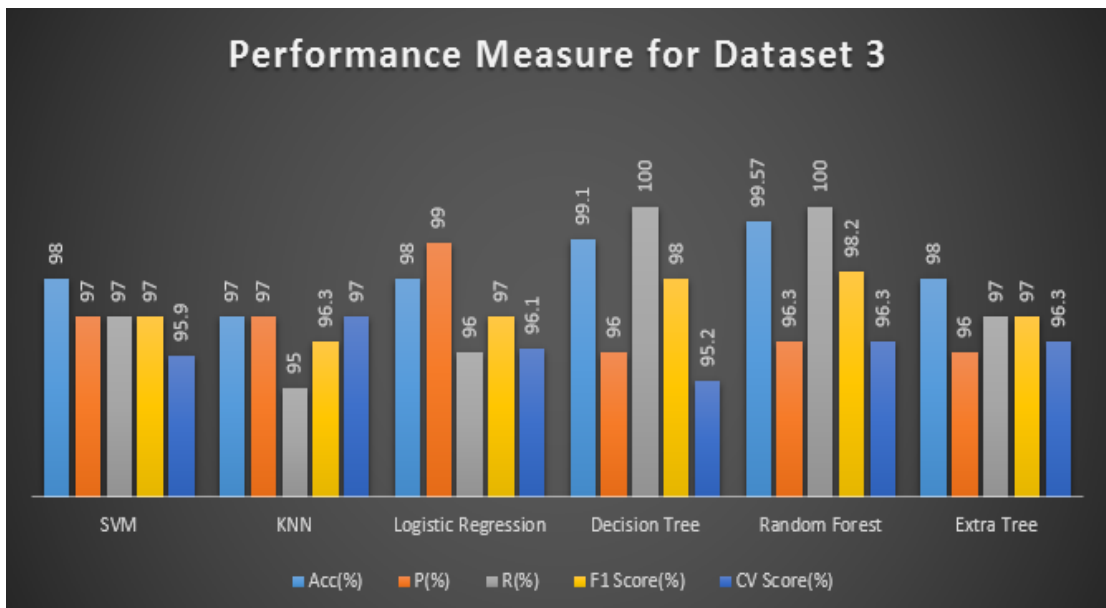


Figure 5.3: Performance Measure for Breast Cancer Dataset of University Medical Centre, Institute of Oncology, Ljubljana, Yugoslavia)

5.2.4 Performance Measure by Ensemble Classification Techniques

This technique performed in Python programming language. To predict breast cancer several machine learning techniques are applied. Results are assembled via voting technique. Evaluation of this ensemble approach is performed by some performance parameter such as Accuracy, F1-score, Recall and precision. Table 5.5 represents the results of ensemble model in contrast to individual ML technique. Result shows that ensemble technique is more accurate at each performance parameter.

The dataset is trained and tested for every individual classification algorithm. Cross validation is used for accurate prediction and it is also called rotation estimation. We are using 10-fold cross-validation to limit problems like over fitting and this method uses over repeated random subsampling where all observations are used for both training and validation, and each observation is used for validation exactly once. In our proposed method we first train and test the dataset with the individual algorithms. Later we have prioritized the data by giving the ranks and removing the attributes contains lowest rank. Then finally by using vote strategy, we combines all predicted outputs to achieve greater accuracy.

Simulation results in Table 5.5 shows that Wisconsin Breast Cancer Dataset [12] having 99.61% has the highest accuracy when comparing with the other Dataset. Wisconsin Breast Cancer Dataset [12] is the best among all individual Dataset as it has the highest Precision 98%, Recall 100% also F1 Score 99% of all for predicting breast cancer.

Table 5.5: Performance Measure by using Ensemble Classifier

Dataset	Precision (%)	Recall(%)	F1 Score	Accuracy (%)
Wisconsin Breast Cancer Dataset [12]	98	100	99	99.61
Breast Cancer Wisconsin (Diagnostic) Dataset [30]	97	97.5	98	97.1
Breast Cancer Dataset of University Medical Centre [31]	75	98	89	82

5.3 Performance Analysis

A comparative study for breast cancer prediction of existing works which also used Dataset1 [12] is illustrated in Table 5.6. Among these, the accuracy of Kernel-based orthogonal transform [11] was best (98.53%). Azar et al. [34] studied that the performance of different decision tree models to make prediction of breast cancer and got the best accuracy of 97.07% for boosted decision tree. Local linear wavelet neural network (LLWNN) [35] secured associate accuracy level of 97.2%. On the other hand, Azar et al. analyzed different types support vector machine models and got the best accuracy of 97.1429%, by Linear Programming SVM (LPSVM) [36]. The proposed system in [37] included Naive Bayes, SVM and J48 maltreatment as classifier methodology to realize accuracy of 97.13%. Latchoumi et al. [38] also used a weighted smooth SVM and got 98.42% accuracy. Sakri et al. [39] reported 81.3%, 80% and 75%, accuracy for Fast Decision Tree Learner (RepTree), NB and KNNs using particle swarm optimization feature selection. In [40], with the assistance of gradient boosting, 91.7% accuracy is achieved by BBN, and BAN and 94.11% gained for TAN. Chaurasia et al. [41] reported accuracy of 97.36% using Naive Bayes.

In our analysis the Figure 5.4 shows that for Dataset1 [12], the Random Forest model performs relatively higher than the other techniques with 99.57% accuracy, 96.3% precision, 100% recall and 98.2 F1 score. If we compare existing techniques, Random Forest and Decision Tree outperform all of these in terms of accuracy.

5.4 Summary

For Breast Cancer prediction, if the target variable is 1(malignant), then it is a positive instance, meaning the patient has Breast cancer. Several performance metrics have been used to figure out the performance of the machine learning algorithms in this project.

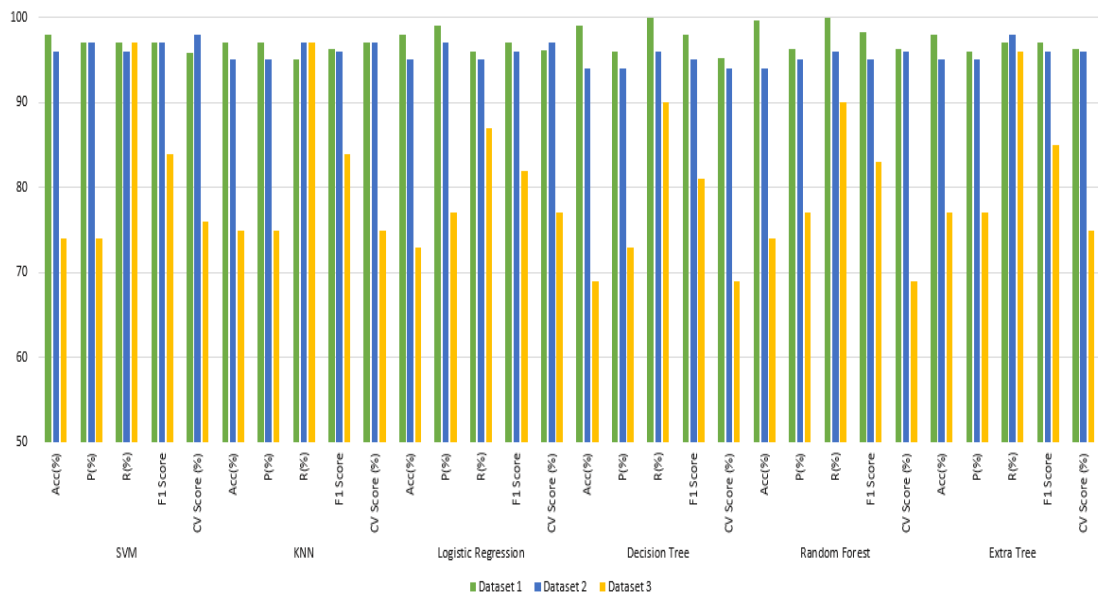


Figure 5.4: This figures shows comparative analysis for the three datasets in terms of accuracy, precision, recall, F1 score and CV score

Table 5.6 below demonstrates the results of different metrics for the algorithms to predict breast cancer on three individual datasets. Random forest performed best among the 6 machine learning techniques with the best accuracy of 99.57%. Among the others, the accuracy of the decision tree is 99.1%, and SVM, LR, and Extra tree achieved 98% accuracy.

Table 5.6: Comparison Study of Breast Cancer Prediction

Methods	Accuracy
Kernel orthogonal transform [11]	98.53
Single decision tree [34]	95.75
Boosted decision tree [34]	97.07
Decision tree forest [34]	95.51
Local Linear Wavelet Neural Network [35]	97.2
Linear Programming SVM [36]	97.14
Lagrangian SVM [36]	95.42
Smooth SVM [36]	96.57
Proximal SVM [36]	96
Lagrangian SVM [36]	96.57
Standard SVM [36]	94.86
Weighted-Particle Swarm Optimization Smooth SVM [38]	98.42
SVM-Naïve Bayse-J48 [37]	97.13
Naïve Bayes [39]	81.3
Fast Decision Tree Learner [39]	80
K-Nearest Neighbor [39]	75
Bayes Belief Network [42]	91.7
Boosted Augmented Naïve Bayes [40]	91.7
Tree Augmented Naïve Bayes [40]	94.11
Naïve Bayes [41]	97.36
Proposed Random Forest	99.57
Proposed Decision Tree	99.1

Chapter 6

Conclusions

This chapter summarizes our research and also highlights the limitations of our research. A brief account of the future works or steps we intend to take to improve our models or research is also stated here.

6.1 Conclusions

We present a generic mechanism for feature selection and model building for the prediction of breast cancer. The proposed mechanism has been used to generate six different machine learning models and 3 different datasets are used for comparative analysis. Among all these techniques, the Random Forest came out with the very best accuracy which is 99.57% for the UC Irvine breast cancer dataset; KNN had the lowest accurate (97%) one among the six techniques. In addition, the cross-validation score for the random forest is 96.3% and 95.2% for decision trees. The other two datasets are used to compare the accuracy and find out the model consistency. Experimental results show that the model works well for both numeric and categorical data. This project shows that the machine learning technique could be highly effective for the early detection of breast cancer which is crucial for the survival of a patient.

6.2 Future Prospects of Our Work

Despite attaining accurate results and accuracies with the six algorithms we have used, we wish to confirm the results we obtained are not biased thanks to the scale of our dataset. We would like to search out an even bigger dataset and perform a similar analysis and see if the results are identical. Additionally, besides the models we have tried, we would conjointly wish to attempt other algorithms such as Adaboost in order to compare results and continue our search for the best model for prediction. The idea of applying other feature selection on the currently used models is also under consideration, such as the Recursive Feature Elimination and the Correlation Heat Map. Overall, we believe that if the quality of studies continues to improve, it is likely that the use of machine learning classifiers will become much more commonplace in many clinical and hospital settings.

References

- [1] C. E. DeSantis, J. Ma, M. M. Gaudet, L. A. Newman, K. D. Miller, A. Goding Sauer, A. Jemal, and R. L. Siegel, “Breast cancer statistics, 2019,” *CA: a cancer journal for clinicians*, vol. 69, no. 6, pp. 438–451, 2019.
- [2] T. L. Nguyen, D. F. Schmidt, E. Makalic, G. Maskarinec, S. Li, G. S. Dite, Y. K. Aung, C. F. Evans, H. N. Trinh, L. Baglietto *et al.*, “Novel mammogram-based measures improve breast cancer risk prediction beyond an established mammographic density measure,” *International Journal of Cancer*, vol. 148, no. 9, pp. 2193–2202, 2021.
- [3] N. Ozmen, R. Dapp, M. Zapf, H. Gemmeke, N. V. Ruiter, and K. W. van Dongen, “Comparing different ultrasound imaging methods for breast cancer detection,” *IEEE transactions on ultrasonics, ferroelectrics, and frequency control*, vol. 62, no. 4, pp. 637–646, 2015.
- [4] R. Kumar and A. Alavi, “Fluorodeoxyglucose-pet in the management of breast cancer,” *Radiologic Clinics*, vol. 42, no. 6, pp. 1113–1122, 2004.
- [5] M. L. Flexman, M. A. Khalil, H. K. Kim, C. J. Fong, A. H. Hielscher, R. Al Abdi, R. L. Barbour, E. Desperito, and D. L. Hershman, “Digital optical tomography system for dynamic breast imaging,” *Journal of biomedical optics*, vol. 16, no. 7, p. 076014, 2011.
- [6] S. H. Park and K. Han, “Methodologic guide for evaluating clinical performance and effect of artificial intelligence technology for medical diagnosis and prediction,” *Radiology*, vol. 286, no. 3, pp. 800–809, 2018.
- [7] F. F. Ting, Y. J. Tan, and K. S. Sim, “Convolutional neural network improvement for breast cancer classification,” *Expert Systems with Applications*, vol. 120, pp. 103–115, 2019.
- [8] D. P. Acharjya and C. L. Chowdhary, “Breast cancer detection using hybrid computational intelligence techniques,” in *Handbook of Research on Emerging Per-*

- spectives on Healthcare Information Systems and Informatics*. IGI Global, 2018, pp. 251–280.
- [9] R. Ghorbani and R. Ghousi, “Predictive data mining approaches in medical diagnosis: A review of some diseases prediction,” *International Journal of Data and Network Science*, vol. 3, no. 2, pp. 47–70, 2019.
- [10] M.-W. Huang, C.-W. Chen, W.-C. Lin, S.-W. Ke, and C.-F. Tsai, “Svm and svm ensembles in breast cancer prediction,” *PloS one*, vol. 12, no. 1, p. e0161501, 2017.
- [11] Y. Xu, Q. Zhu, and J. Wang, “Breast cancer diagnosis based on a kernel orthogonal transform,” *Neural Computing and Applications*, vol. 21, no. 8, pp. 1865–1870, 2012.
- [12] W. Wolberg, O. Mangasarian, and D. Aha, “Uci machine learning repository: Breast cancer wisconsin (original) data set,” *UCI Machine Learning Repository: Breast Cancer Wisconsin (Original) Data Set. University of Wisconsin Hospitals Madison, Wisconsin, USA, nd Web*, 2015.
- [13] C. Kahlenborn, F. Modugno, D. M. Potter, and W. B. Severs, “Oral contraceptives and breast cancer,” in *Mayo Clinic Proceedings*, vol. 83, no. 7. Elsevier, 2008, pp. 849–850.
- [14] E. B. C. T. C. Group *et al.*, “Effect of radiotherapy after breast-conserving surgery on 10-year recurrence and 15-year breast cancer death: meta-analysis of individual patient data for 10 801 women in 17 randomised trials,” *The Lancet*, vol. 378, no. 9804, pp. 1707–1716, 2011.
- [15] C. E. Floyd Jr, J. Y. Lo, A. J. Yun, D. C. Sullivan, and P. J. Kornguth, “Prediction of breast cancer malignancy using an artificial neural network,” *Cancer: Interdisciplinary International Journal of the American Cancer Society*, vol. 74, no. 11, pp. 2944–2948, 1994.
- [16] Y. Wu, M. L. Giger, K. Doi, C. J. Vyborny, R. A. Schmidt, and C. E. Metz, “Artificial neural networks in mammography: application to decision making in the diagnosis of breast cancer.” *Radiology*, vol. 187, no. 1, pp. 81–87, 1993.
- [17] H. A. Abbass, “An evolutionary artificial neural networks approach for breast cancer diagnosis,” *Artificial intelligence in Medicine*, vol. 25, no. 3, pp. 265–281, 2002.

- [18] G.-B. Huang, Q.-Y. Zhu, and C.-K. Siew, "Extreme learning machine: theory and applications," *Neurocomputing*, vol. 70, no. 1-3, pp. 489–501, 2006.
- [19] N. Aksha and N. Alapati, "Classification of breast cancer data using enhanced supervised machine learning algorithm optimized svm."
- [20] C. P. Utomo, P. S. Pratiwi, A. Kardiana, I. Budi, and H. Suhartanto, "Best-parameterized sigmoid elm for benign and malignant breast cancer detection," in *International Conference on Artificial Intelligence and Pattern Recognition, AIPR 2014*, 2014.
- [21] T. A. Assegie, "An optimized k-nearest neighbor based breast cancer detection," *Journal of Robotics and Control (JRC)*, vol. 2, no. 3, pp. 115–118, 2021.
- [22] T. Joachims, "Making large-scale support vector machine learning practical, advances in kernel methods," *Support vector learning*, 1999.
- [23] R. Gandhi, "Nearest neighbor," *Understanding Machine Learning*, 2018.
- [24] M. Pal, "Random forest classifier for remote sensing classification," *International journal of remote sensing*, vol. 26, no. 1, pp. 217–222, 2005.
- [25] R. E. Wright, "Logistic regression." 1995.
- [26] C. Désir, C. Petitjean, L. Heutte, M. Salaun, and L. Thiberville, "Classification of endomicroscopic images of the lung based on random subwindows and extra-trees," *IEEE transactions on biomedical engineering*, vol. 59, no. 9, pp. 2677–2683, 2012.
- [27] K. Kira and L. A. Rendell, "A practical approach to feature selection," in *Machine learning proceedings 1992*. Elsevier, 1992, pp. 249–256.
- [28] G. H. Dunteman, *Principal components analysis*. Sage, 1989, no. 69.
- [29] A. Bronshtein, "Train/test split and cross validation in python," *Understanding Machine Learning*, 2017.
- [30] M. Lichman, "Uc irvine machine learning repository," *University of California "http://archive.ics.uci.edu/ml"*, Irvine, School of Information and Computer Sciences, 2013.
- [31] 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.

- [32] T. Hastie, R. Tibshirani, J. H. Friedman, and J. H. Friedman, *The elements of statistical learning: data mining, inference, and prediction*. Springer, 2009, vol. 2.
- [33] S. Visa, B. Ramsay, A. L. Ralescu, and E. Van Der Knaap, "Confusion matrix-based feature selection." *MAICS*, vol. 710, pp. 120–127, 2011.
- [34] A. T. Azar and S. M. El-Metwally, "Decision tree classifiers for automated medical diagnosis," *Neural Computing and Applications*, vol. 23, no. 7, pp. 2387–2403, 2013.
- [35] M. R. Senapati and P. K. Dash, "Local linear wavelet neural network based breast tumor classification using firefly algorithm," *Neural Computing and Applications*, vol. 22, no. 7, pp. 1591–1598, 2013.
- [36] A. T. Azar and S. A. El-Said, "Performance analysis of support vector machines classifiers in breast cancer mammography recognition," *Neural Computing and Applications*, vol. 24, no. 5, pp. 1163–1177, 2014.
- [37] U. K. Kumar, M. S. Nikhil, and K. Sumangali, "Prediction of breast cancer using voting classifier technique," in *2017 IEEE international conference on smart technologies and management for computing, communication, controls, energy and materials (ICSTM)*. IEEE, 2017, pp. 108–114.
- [38] T. Latchoumi and L. Parthiban, "Abnormality detection using weighed particle swarm optimization and smooth support vector machine," *Biomedical Research*, vol. 28, no. 11, pp. 4749–4751, 2017.
- [39] S. B. Sakri, N. B. A. Rashid, and Z. M. Zain, "Particle swarm optimization feature selection for breast cancer recurrence prediction," *IEEE Access*, vol. 6, pp. 29 637–29 647, 2018.
- [40] A. Bazila Banu and P. Thirumalaikolundusubramanian, "Comparison of bayes classifiers for breast cancer classification," *Asian Pacific journal of cancer prevention: APJCP*, vol. 19, no. 10, p. 2917, 2018.
- [41] V. Chaurasia, S. Pal, and B. Tiwari, "Prediction of benign and malignant breast cancer using data mining techniques," *Journal of Algorithms & Computational Technology*, vol. 12, no. 2, pp. 119–126, 2018.
- [42] M. R. Senapati, G. Panda, and P. K. Dash, "Hybrid approach using kpso and rls for rbfn design for breast cancer detection," *Neural Computing and Applications*, vol. 24, no. 3, pp. 745–753, 2014.