BACHELOR OF SCIENCE IN
COMPUTER SCIENCE AND ENGINEERING

Inspiring Excellence

Early Detection of Breast Cancer Using Machine Learning

Authors
Wasi Mohammad Fuad

Supervisor
Dr. Md. Ashraful Alam
Assistant Professor
Department of CSE

A thesis submitted to the Department of CSE
in partial fulfillment of the requirements for the degree of
B.Sc. Engineering in CSE

Department of Computer Science and Engineering
BRAC University, Dhaka - 1212, Bangladesh

December 2018

I would like to dedicate this thesis to my father, who taught me that the best kind of knowledge to have is that which is learned for its own sake. It is also dedicated to my mother, who taught me that even the largest task can be accomplished through perseverance and dedication.

Declaration

I hereby declare that this thesis is my original work based on the results I have calculated. The materials or work found by other researchers and sources have been properly acknowledged by giving credit where due, through appropriate references. This thesis report or any part of it has not been submitted elsewhere for the award of any Degree or Diploma.

Authors:

Wasi Mohammad Fuad
Student ID: 14201029

Supervisor:

Dr. Md. Ashraful Alam
Assistant Professor, Department of Computer Science and Engineering
BRAC University

December 2018

The thesis titled Early Detection of Breast Cancer Using Machine Learning
Submitted by:
Wasi Mohammad Fuad Student ID: 14201029
of Academic Year 2018 has been found as satisfactory and accepted as partial fulfillment of the requirement for the Degree of Computer Science and Engineering

1.

Dr. Md. Ashraful Alam
Assistant Professor
BRAC University

2.

Dr. Md. Abdul Mottalib
Professor and Chairperson
BRAC University

Acknowledgements

I would like to thank the Almighty Allah for giving me the strength and support to complete this research work.

Also, I would like to express my sincere gratitude to my supervisor Dr. Md. Ashraful Alam for his valuable and constructive suggestions throughout the process. His willingness to give his time so generously has been very much appreciated.

Besides my advisor, I would like to thank both my parents for their constant support and for believing in me.

Last but not the least, I would like to thank everyone I learnt something from throughout my life.

Abstract

Breast cancer is the most common cancer among women but in can occur in both the genders. It is accountable for an appalling number of deaths worldwide. In a particularly low-resource developing country like Bangladesh, there is a lack of awareness and facilities mostly in rural areas and high rate of instances of breast cancer that is diagnosed in the last stages. However, the early detection of breast cancer can lead to help increase the odds of survival. Nowadays, with the increasing number of patients, manual analysis of medical images becomes tedious, time consuming and unfeasible. With the advancement in the field of machine learning, it is now possible to create an automated and accurate Computer Aided Diagnosis (CAD) system in order to make the entire process of detecting a malignant tumor more resource efficient and time saving through proper utilization. This paper presents the comparative analysis of different machine learning algorithms and their results in predicting cancerous tumors. The proposed model uses supervised machine learning algorithms such as Random Forest, Support Vector Machine, K-Nearest Neighbors, Naïve Bayes and Logistic Regression with and without PCA on a dataset with 30 features extracted from a digitized image of a fine needle aspirate (FNA) of a breast mass. Deep learning models like Artificial Neural Network and Convolutional Neural Network are used and their performances are compared. From the comparative analysis, it is observed that the deep learning models outperform all other classifiers and achieves impressive scores across multiple performance metrics such as Accuracy of 98.83%, Precision of 98.44% and Recall of 100%.

Keywords: Computer Aided Diagnosis, Breast Cancer Detection, Random Forest, Support Vector Machine, K-Nearest Neighbors, Naïve Bayes, Logistic Regression, PCA, FNA, Artificial Neural Network, Convolutional Neural Network

Table of contents

List of figures

List of tables

Chapter 1

Introduction

Breast cancer develops from breast tissues with abnormal cells growing, changing and multiplying out of control. It is the most common type of cancer among women in both developed and less developed nations with an estimated death of 508,000 women in the year 2011 alone [24] and accounted for 25% of all cancer cases and 15% of all cancer deaths among females in the estimated cancer case in 2012 [21]. Cancer constitutes an enormous burden on society in more and less economically developed countries alike. Cancer cases are becoming more common due to the growth and aging of the population, as well as a widespread rise of established risk factors such as smoking, overweight, physical inactivity. Early detection of cancer significantly increases the probability of recovering through successful treatment. Delays in diagnosis results in late-stage presentation with consequences of lower likelihood of survival, higher costs of treatments and even death.

The most common techniques used for cancer detection are X-ray mammography and magnetic resonance imaging (MRI). However, these present innovations have a few downsides as they are very costly, extensive in size and are only affordable in large hospital facilities. The mentioned methods also may have some side effects and false positives. However, with the volume of data generating extremely fast in the field of biomedical and advancement of technology, machine learning techniques offer promising results. Machine learning helps to extract information and knowledge from the basis of past experiences and detect hard-to-perceive pattern from large and noisy dataset to give accurate results within a short period of time. Application of machine learning in the medical domain is growing rapidly due to the effectiveness of its approach in prediction and classification, especially in medical diagnosis to predict breast cancer, now it is widely applied to biomedical research.

1.1 Motivation

In Bangladesh, breast cancer is the most common cancer among women. It accounts for 69% of death related to cancer among women [6]. Evidently, National Institute of Cancer Research Hospital has been established in Bangladesh due to the vital public health concern of the Bangladesh government. The incidence rate of breast cancer was about 22.5 per 100000 in females [22]; breast cancer has been reported as the highest prevalence rate (19.3 per 100,000) among Bangladeshi women between 15 and 44 years of age when compared to other types of cancer. A study carried out in Khulna, northern division of Bangladesh, showed that 87% of the new instances of breast cancer were diagnosed as stage III+, where cancer had spread to other parts of the body. The fatal state meant that the scope of treatment was restricted and extremely costly in a particularly low-resource developing country like Bangladesh. The primary reason is the absence of awareness in the early detection of cancer which portrays the real circumstance in rural areas of Bangladesh. There is a scope to introduce suitable computer aided diagnosis systems that can accurately detects breast cancer in the early stages leading to help prevent thousands, if not millions, of avoidable deaths in our beloved developing nation and also the entire world. This influenced the research work done in this paper.

1.2 Objective

The objective of this paper is to describe a predictive model built using Machine Learning algorithms for early detection of breast cancer in order to improve the prognosis and chances of survival through timely clinical treatment to patients. Most influential machine learning algorithms [26] have been used for comparison in terms of accuracy, sensitivity, specificity and precision. Data mining approaches, for instance, applied to medical science topics rise rapidly due to their high performance in predicting outcomes, reducing costs of medicine, promoting patients' health, improving healthcare value and quality and in making real time decision to save people's lives. Therefore, the correct diagnosis of breast cancer and classification of patients into malignant and benign groups is the center of a lot of research.

1.3 Thesis Orientation

The next chapter of the paper discusses about the similar works done before in the same field by different researchers and the fundamentals of machine learning. The proposed model is described in the third chapter and its subsections explain the implementation including

the visualization of the dataset, data preprocessing, Principal Component Analysis (PCA), train-test split and the short background of the algorithms used for this model. The following section consists of the performance metrics and the experimental results. Chapter four ends with the discussion and comparison of the results of the different models. Lastly, the final chapter of this paper includes few concluding remarks and future prospects.

Chapter 2

Literature Review

The evolution and integration of technology in the medical field is ever increasing. Numerous modern techniques have been introduced in order to help identify/diagnose diseases, provide personalized treatment, drug discovery and manufacturing, clinical trial research, radiology and radiotherapy, smart electronic health records, epidemic outbreak predictions, etc. Computer-Aided Diagnosis (CAD) systems have been used for detection and characterization of multiple types of cancer including a number of CAD systems designed and used for breast cancer. CAD is becoming an increasingly important tool in the mammographic interpretation process to assist radiologists to come to definite conclusions. Present CAD systems in clinical use serve as a second reader for breast cancer detection. CAD systems for classification of malignant and benign lesions are under development by a number of research groups.

2.1 Related Works

There are numerous modern techniques have been evolved with the evolution of technology for the prediction of breast cancer. The work related to this field is outlined shortly as follows.

Some of the studies [29], [16], and [5] displayed work associated to prediction and diagnosis of diseases using machine learning techniques like decision tree for detection of cancer. KNN algorithm is well known for simplicity and versatility in implementation which makes it one of the most frequently used classification algorithm in machine learning according to Jin [17].

Liu Lei [15] proposed a model that uses machine learning for cancer detection. In this research, Logistic Regression algorithm of Sklearn machine learning library has been used to classify the data sets of breast cancer. Two features of maximum texture and minimum perimeter was selected and the classification accuracy stood at 96.5%.

Zemouri, Omri, Devalland, Arnould, Morello, Zerhouni and Fnaiech [28] proposed a

model that uses a Breast Cancer Computer Aided Diagnosis (BC-CAD) based on joint variable selection and a Constructive Deep Neural Network "ConstDeepNet". Wisconsin Breast Cancer Dataset (WBCD) and real data from the north hospital of Belfort (France) were used to predict the recurrence score of the Oncotype DX. They applied a method to lower the number of inputs for training a deep learning neural network. Accordingly, performance of the use of the Deep Learning architecture alone was exceeded by the use of joint variable algorithm with ConstDeepNet.

Bellaachia and Guven (2006) [4] looked into the use of Naïve Bayes, the back-propagated neural network and the C4.5 decision tree algorithms on SEER dataset which contained 16 attributes and 482,052 records. The dataset is considered to be ideal due to large amount of patient and a moderate number of attributes. From their experiment, C4.5 algorithm outperformed the rest with an accuracy of 86.7%.

In [27], they demonstrated a new strategy for breast cancer diagnosis by integrating a deep learning based unsupervised feature extraction algorithm, stacked auto-encoders with greedy layer-wise pre-training algorithm to extract important features and information, with a support vector machine model to identify samples with new features into benign and malignant tumors. The proposed method of deep learning based unsupervised feature extraction was tested on the Wisconsin Diagnostic Breast Cancer data set and it significantly improved the performance of classification and provided a promising approach to breast cancer diagnosis.

There have been researches in the recent past and there are on-going researches which aims to observe the features that are most helpful in predicting malignant or benign cancer and to see general trends that might help us in selecting particular models and hyper parameter selections. The aim of almost all researches have been to reach the highest accuracy possible in the shortest time.

2.2 Machine Learning

Machine learning (ML) is one of the fields of Artificial Intelligence (AI) where statistical techniques are used to provide the computer systems with the capability to "learn" and improve by itself progressively without being explicitly programmed. Machine learning explores the study and construction of algorithms that can learn from and make predictions on data and datasets [14] based on different teaching mechanisms. The term 'machine learning' was initially coined by Arthur Samuel in 1959 [20]. Three important categories of machine learning can be described as follows:

1. Supervised learning: In this form of learning, the machine uses data that is labeled

and some of the data is already tagged with the correct answers to learn the mapping function from the input to the output. The goal is to approximate the mapping function so well that when you have new input data (x) that you can predict the output variables (Y) for that data.

2. Unsupervised learning: The machine is trained using data that is neither classified nor labeled. This allows the algorithm to perform calculations on it's own accord without guidance. The task of the machine is to group unsorted data according to patterns and differences in order to find the hidden structure by itself.

3. Reinforcement learning: The machine or the agent learn how to behave in an environment by performing actions and seeing the results based on the action allowing it to dictate the ideal action in a specific circumstance.

In the modern times, the vast amount of data available is not feasible for human being to keep up with and analyze them. Machine learning, which is a subset of computer science and an important branch of artificial intelligence, primarily focuses on the development and building of algorithms to over this problem. The very recent advancement in this field has opened up vast and almost limitless applications in fields ranging from financial industries, data security to medical fields. But there is still much space to make progress by means of using Machine Learning in social media services, disease prediction and identification, virtual assistants, search engine refining, fraud detection, manufacturing, etc. It is only going to improve and integrate in our daily lives making it easier and more convenient in the future.

Chapter 3

Proposed Model

In this paper, various top algorithms were applied on Wisconsin Breast Cancer (Diagnostic) Data Set (WBCD) [16] consisting of 569 patients for early detection of breast cancer and the results were compared against each other using multiple performance metrics. The workflow of the comparative analysis study is shown in figure 3.1 for this model.

3.1 Dataset

The dataset chosen for this research is the Wisconsin Breast Cancer (Diagnostic) Data Set (WBCD). The dataset is publicly available on the reputed Machine Learning Repository that is UCI-Repository. WBCD was made by Dr. William H. Wolberg, doctor at the University Of Wisconsin Hospital at Madison, Wisconsin, USA. Dr. Wolfberg used Xcyt to analyze fluids samples taken from patients with solid breast masses [25]. Xcyt is an easy-to-use graphical computer program which is equipped to perform the investigation of cytological features based on digital scans. The dataset comprises of 569 samples and 32 attributes of visually measured atomic features computed from a digitized image of a fine needle aspirate (FNA) of a breast mass. FNA is a thin needle which is injected into the region of abnormal-appearing body fluid or tissues and collects a sample to make a diagnosis or predicting disease such as cancer. Among the 569 samples, the class distribution are 212 cancerous tumors (malignant) and other 357 non-cancerous tumors (benign). Ten features are computed from each one of the cells in the sample which are as follows:

1. radius (mean of distances from center to points on the perimeter)
2. texture (standard deviation of gray-scale values)
3. perimeter

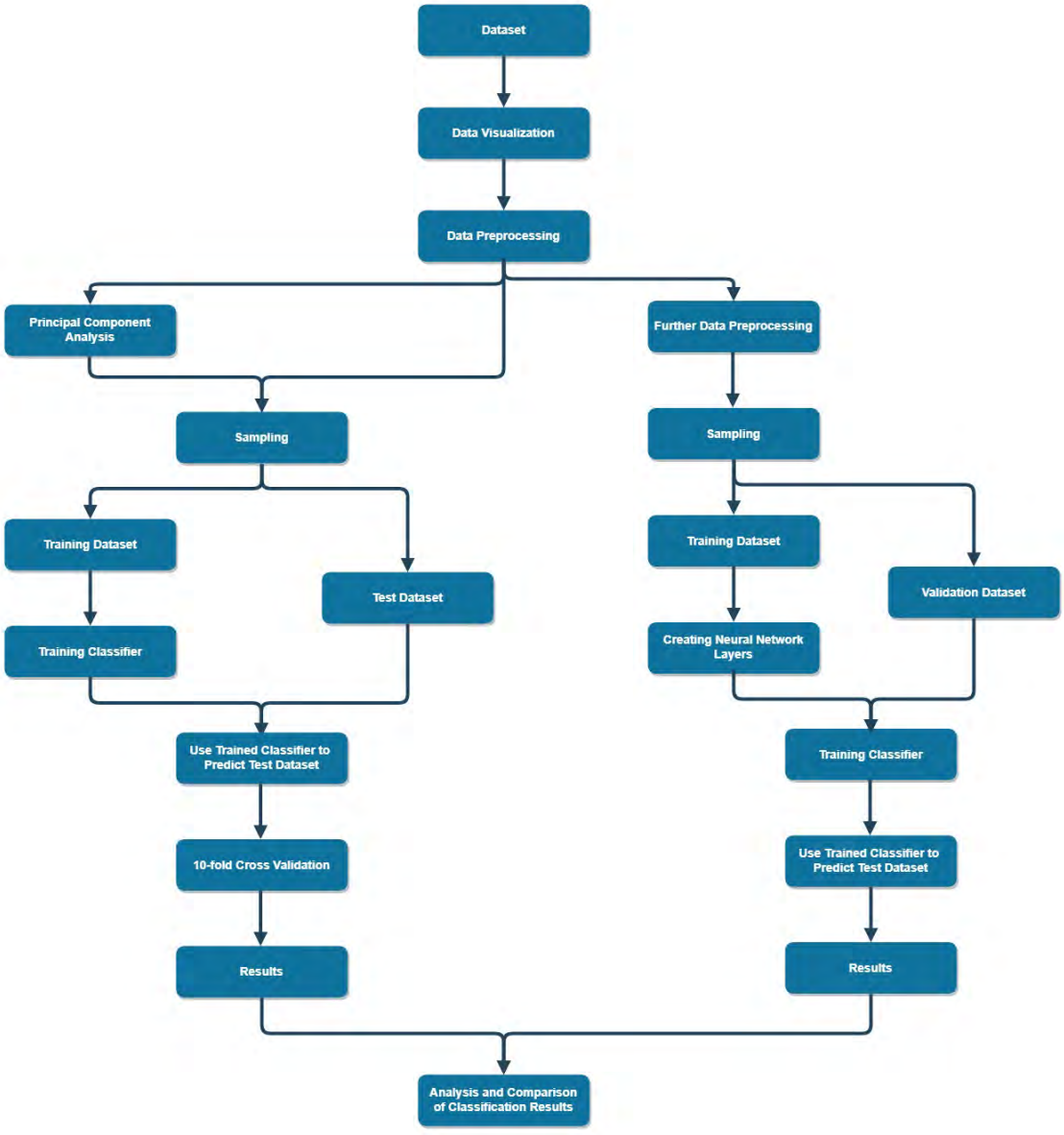


Fig. 3.1 Workflow of the Proposed System.

4. area
5. smoothness (local variation in radius lengths)
6. compactness (perimeter² / area—1.0)
7. concavity (severity of concave portions of the contour)
8. concave points (number of concave portions of the contour)
9. symmetry
10. fractal dimension (“coastline approximation”—1)

The mean, standard error and "worst" or largest (mean of the three largest values) of these features were computed for each image, resulting in 30 features. For instance, field 3 is Mean Radius, field 13 is Radius SE and field 23 is Worst Radius. All feature values are recoded with four significant digits. There are no missing values in this dataset. There is presence of both numerical and categorical features in the dataset. ‘Diagnosis’ is the only column with categorical feature, which we are going to predict, which says if the cancer is M = malignant or B = benign. The rest of the features are numerical.

3.2 Data Visualization

3.2.1 Histogram

A histogram is a graphical representation of data or information using bars of different heights where each bar groups numbers into ranges. Higher bars show that more data falls in that range. The shape and spread of continuous sample data can be shown using a histogram. Figure 3.2 shows the class distribution of diagnosed malignant (M) and benign (B) tumors. There are 212 malignant tumors which is approximately 38% and other 357 benign tumors making up the rest of the 62% of the predictive class.

The nucleus features can be plotted against diagnosis as seen on figure 3.3 from which we can observe that mean values of cell radius, perimeter, area, compactness, concavity and concave points can be used in classification of the cancer. Larger values of these parameters tends to show a correlation with malignant tumors. The mean values of texture, smoothness, symmetry or fractal dimension does not show a particular preference of one diagnosis over the other. In any of the histograms there are no distinguishable large outliers that require a further cleanup.

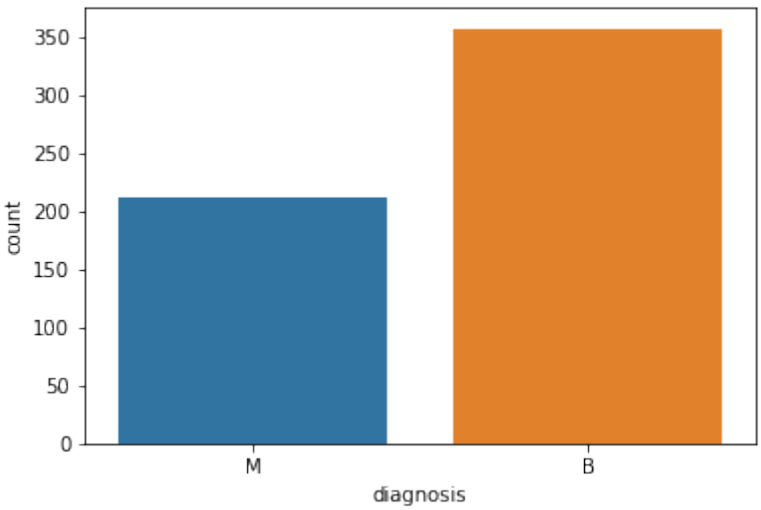


Fig. 3.2 Class Distribution.

3.2.2 Heatmap

A heatmap is a two-dimensional representation with the help of colors for visualization of both simple and complex information. Heatmap is an extremely useful way to see which intersections of the values have higher concentration of the data compared to the others. Figure 3.4 represents a correlation matrix using a heatmap. It is used to show the correlation among all 30 features in this dataset.

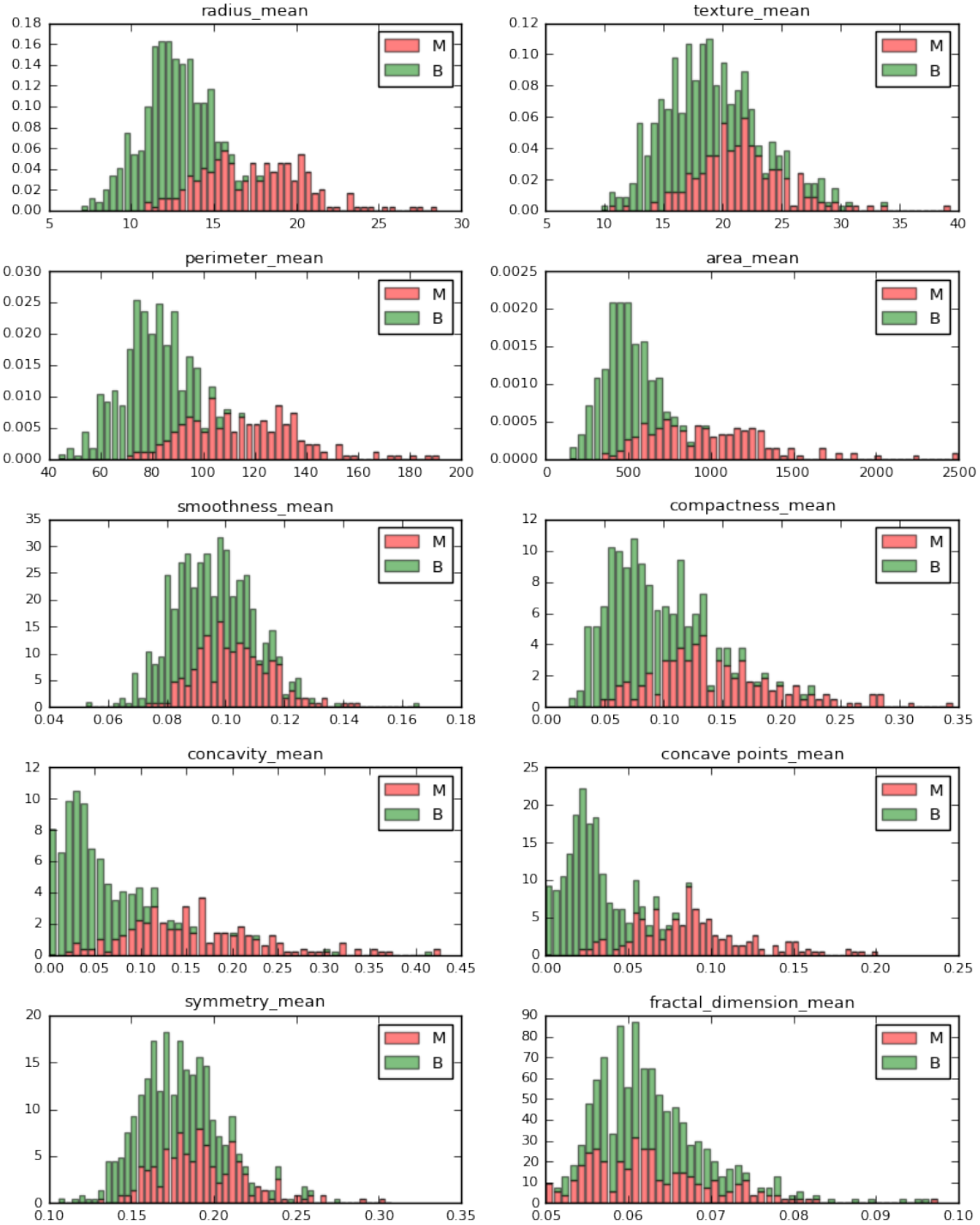


Fig. 3.3 Nucleus Features vs Diagnosis.

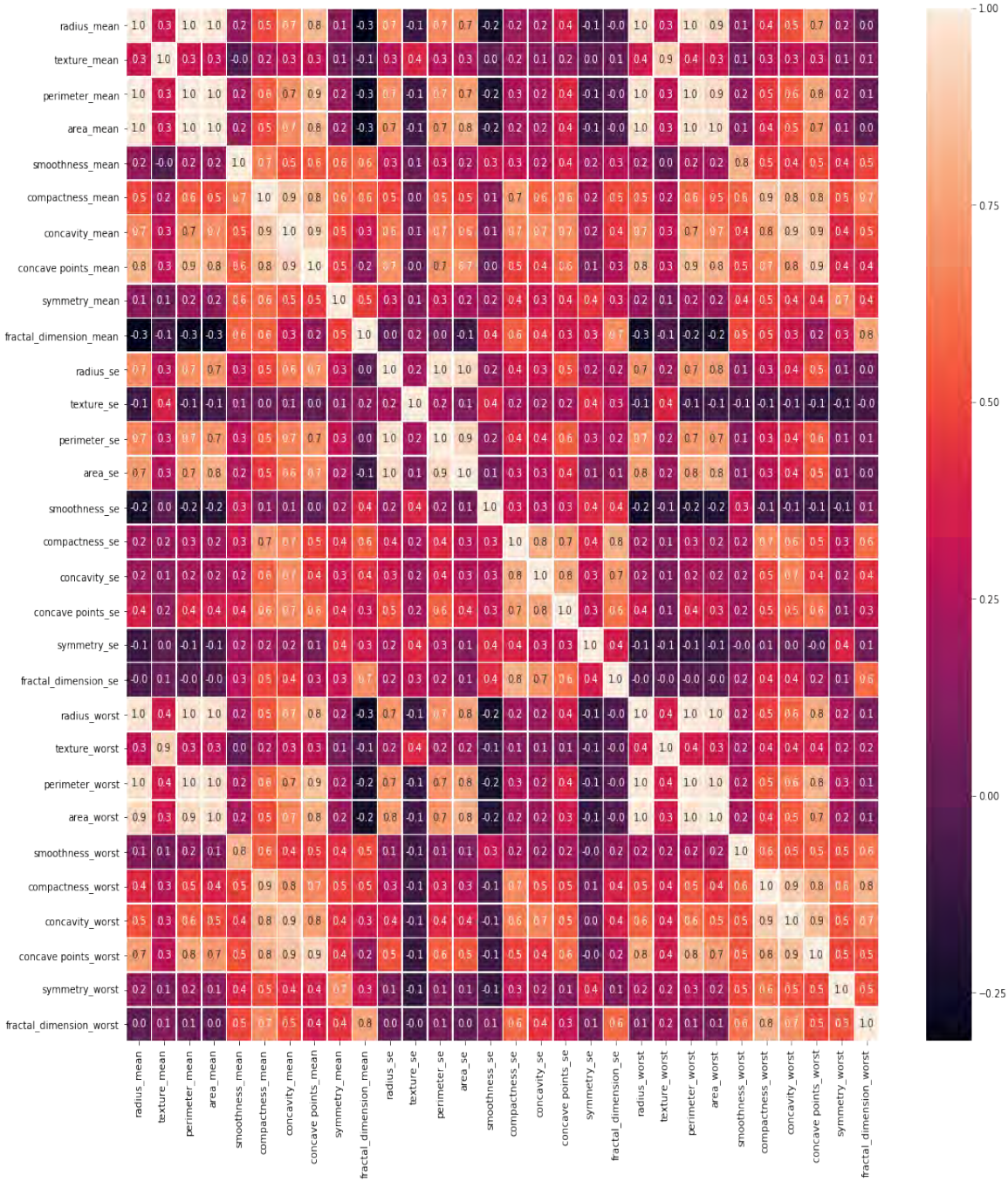


Fig. 3.4 Correlation between all variables

3.3 Data Preprocessing

3.3.1 Categorical Variable Conversion

The dataset included both numerical and categorical features. Among which the 'Diagnosis' column had categorical feature, which says if the cancer is M = malignant or B = benign. The rest of the features are numerical. Most of the algorithms produce better result with numerical variable. In python, library "sklearn" requires features in numerical arrays and categorical variables cannot be fitted into a regression equation in their raw form. Hence, Label Encoder was used to transform non-numerical to numerical labels.

3.3.2 Feature Scaling

The range of values of the attributes in the dataset varies widely and feature scaling is used to bring it to a standardized range. It is also known as data normalization. This is done because some algorithms will not function properly without it and data should be standardized before applying PCA as variables with higher and lower variance are going to be treated differently. In this paper, scikit-learn module sklearn.preprocessing.StandardScaler is used to implement standardization in python. The standard score of a sample x is calculated as:

$$z = \frac{x - u}{s}$$

3.3.3 Principal Component Analysis (PCA)

Principal Component Analysis is a method of dimension-reduction which reduces a large set of variables to a small set that still contains most of the information in the large set. PCA is basically a mathematical procedure that transforms a number of correlated variables into a smaller number of uncorrelated linear variables called principal components using uses an orthogonal transformation [1]. After standardizing the data, PCA was applied for Random Forest, SVM, K-Nearest Neighbors, Logistic Regression and Gaussian Naïve Bayes classifiers but not for the neural networks as both Artificial Neural Network and Convolutional Neural Network because the neural network can approximate any nonlinear mapping through learning and is free from the constraints of a non-linear model. After applying PCA, the dataset is reduced to 8 principal components from previous 30 attributes which represents each observation.

3.3.4 Data Reshaping

Further data processing of is needed in the form of data reshaping for the input of Convolutional Neural Network. The dataset was initially in the shape (569, 30) in 2 dimensional form. Using NumPy library, the data is reshaped to (569, 10, 3) in 3 dimensional form for CNN.

3.3.5 Train-Test Split

The data is normally split into two subsets: training data and testing data (and sometimes to three: train, validate and test). The training dataset is the actual dataset that is used to train the model (weights and biases in the case of Neural Network). The model sees and learns from this data. The test dataset is the sample of data used to provide an unbiased evaluation of a final model fit on the training dataset. While validation dataset is the sample of data used to provide an unbiased evaluation of a model fit on the training dataset while tuning model hyperparameters, but this is for frequent evaluation. The splitting ratio depends on the total number of samples and the actual training model. The train/test split was implemented using the train_test_split class of scikit-learn's model_selection package into a 70:30 ratio with 70% going to the training set and the rest to the test set which is considered to be ideal [8]. The Accuracy, Precision, Recall and F1 score for RF, SVM, KNN, LR and GNB was evaluated by using cross_val_score from sklearn.cross$_validationpackageusing10foldsonthetestdataset$.

3.3.6 Neural Network Layers

Every network has a single input layer and a single output layer. The number of neurons in the input layer equals the number of input variables in the data being processed. The number of neurons in the output layer equals the number of outputs associated with each input. But the challenge is deciding on the number of hidden layers and their neurons. The input layer just passes on the information to the hidden nodes and no computation is performed. Whereas, in the hidden layer computations take place and information is transferred from the input nodes to the output nodes. While a network will only have a single input layer and a single output layer, it can have multiple hidden layers. The output layer is responsible for computations and transfer of information from the network to the outside world. The hidden layers of a Convolutional Neural Network is usually composed of convolutional layers, pooling layers, fully connected layers and normalization layers. The layers created for CNN and ANN model in this paper consists of multiple hidden layers with activation functions namely ReLU, Sigmoid and Softmax.

3.4 Algorithms

The model deals with binary classification of labeled data and the algorithms were chosen based on that fact. But deep learning models like Artificial Neural Network and Convolutional Neural Network were also tested on this as there is no perfectly tailored single solution or one approach that fits all. The machine learning algorithms chosen for this problem are Random Forest (RF). Support Vector Machine (SVM), K-Nearest Neighbors (KNN), Logistic Regression (LR), Gaussian Naïve Bayes, Artificial Neural Network (ANN) and Convolutional Neural Network (CNN). The results of the algorithms were compared to determine the best classifier for the problem.

3.4.1 Random Forest

Random Forest is one of the most popular and powerful machine learning algorithm. It consists of many decision trees and outputs the class that is the mode of the class's output by means of individual trees. After building decision trees on the sample sets, many trees are generated, thus creating a forest. It is good for classification problems like this one and other tasks like regression which functions as explained above by creating a multitude of trees at training and outputting the classes or mean predictions of each specific tree [9][3]. The numerous deep decision trees are trained on separate groups of the same dataset and averaged with the target of decreasing the variance [7].

3.4.2 Support Vector Machine

Support Vector Machine (SVM) is a supervised machine learning technique that is broadly used in pattern recognition and classification problems, especially when the dataset has exactly two classes. SVMs are used to find the ideal hyperplane which separates the classes. An input pattern, called feature vector, is taken by the classifier and determines to which class it belongs to. It classifies linearly separable data, but in general, the feature vectors might not be linearly separable. The kernel trick is used to fix this [11]. SVM uses kernel methods to map input data to higher dimensional data and provides a fast training algorithm. It is used for pattern classification and regression. The performance of an SVM classifier matters on the choice of the kernel function. Different kernel functions are used for different classification tasks. In this project, SVC class from scikit-learn was used to implement SVM. However, SVM can be memory-intensive and complicated to interpret and tune.

3.4.3 K-Nearest Neighbors

K-Nearest Neighbors algorithm is a very simple yet functional non-parametric method used for classification and regression [2]. The model that is created are results from the training sample, associated with a distance function and the choice function of the class. Before a new element is classified, it is compare it to other elements using a similarity measure. The k-nearest neighbors of the elements are then considered, and the class that is most common among the neighbors, is assigned to the element to be classified. Using the distance, the neighbors are weighted. KNN can demand a lot of memory or space to store all of the data but only performs a calculation when a prediction is needed.

3.4.4 Logistic Regression

Logistic regression is another supervised learning technique borrowed by machine learning from the field of statistics. In Logistic Regression the output or target variable is a categorical variable, unlike Linear Regression, and is thus a binary classification algorithm that categorizes a data point to one of the classes of the data [10]. The general equation of Logistic Regression is:

$$log(\frac{p(X)}{1-p(X)}) = \beta_0 + \beta_1 X$$

Where, p(X) is the dependent variable, X is the independent variable, β_0 is the intercept and β_1 is the slope co-efficient.

Input values (X) are combined linearly using coefficient values to determine an output (Y). It significantly differs from linear regression because the output value being modeled is a binary value (0 or 1) rather than a numeric value. The value of the score lies between positive and negative infinity but needs to be between [0, 1]. Therefore, the logistic function, also known as the sigmoid function is used to make the conversion. It's a curve that is shaped like an S, which takes any real number and maps it into a value between 0 and 1. It is widely used for binary classification problems and works better when attributes are reduced that are correlated to each other which explains the results in the future section of result analysis. It's a fast model to learn and effective on binary classification problems.

3.4.5 Naïve Bayes

Naïve Bayes algorithm is one of the top algorithms in machine learning for binary classification. It is a simple probabilistic classifier founded on using the Bayes' theorem with the strong assumption of that the effect of the value of an attribute on a class is independent of the other values of attributes. The formula of Bayes' theorem is:

$$P(A \mid B) = \frac{P(B \mid A)P(A)}{P(B)}$$

Where P(A|B) is posterior probability, P(B|A) is likelihood, P(A) is class prior probability and P(B) is predictor prior probability.

An advantage of the Naive Bayes classifier is that it not dependent on the amount of training data. It can use a small amount of training data to estimate the parameters necessary for classification. It performs better in complicated platforms such as Spam Classification, Medical Diagnosis and Weather forecasting. It is suited when dimensionality of input is high [13]. There are several different types of Naïve Bayes classifiers, among them, the Gaussian Naïve Bayes classifier was used in this model. A typical assumption while handling continuous data is that the continuous values corresponding with each class are distributed according to a Gaussian distribution. Naïve Bayes can compete with more advanced methods like SVM and others with proper pre-processing of the data [19]. Naive Bayes is also a good choice when CPU and memory resources are a limiting factor.

3.4.6 Artificial Neural Network (ANN)

Artificial Neural Network are biological neural networks inspired computing systems which are intended to replicate the way that we humans learn [23]. It is a branch of computational intelligence that uses a variety of optimization tools to acquire information and learn from past experiences and use that learning to classify new data, identify new patterns or predict outcomes [12]. Neural networks consist of input and output layers, as well as a single or multiple hidden layers consisting of nodes called artificial neurons as seen in figure 3.5. The nodes transform the input into something that the output layer can use and the output of each artificial neuron is computed by some non-linear function of the sum of its inputs. The drawback of the neural networks are black boxes, in which the user feeds in data and receives answers. The answers can be fine-tuned, but there is no access to the exact decision making process hence the black box title. One of the bigger challenges is the amount of time it takes to train networks, which can require a considerable amount of compute power as well.

Despite the fact that our dataset was comparatively small for a neural network, it performed well and gives more space to improve with a large dataset in the future.

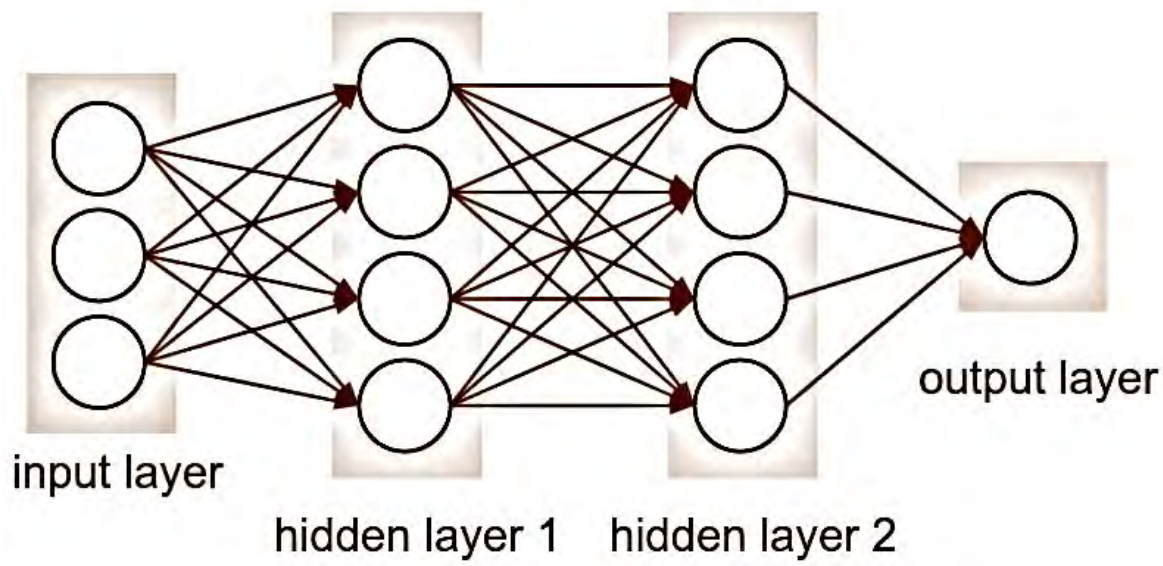


Fig. 3.5 Schematic representation of an Artificial Neutral Network.

3.4.7 Convolutional Neural Network (CNN)

Convolutional Neural Network is a deep learning model which is mainly used for image and video recognition, recommender systems, image classification, medical image analysis, and natural language processing. CNNs require comparatively little pre-processing than other image classification algorithms. CNNs are slightly different from other neural networks as seen in figure 3.6. Firstly, the layers are arranged in 3 dimensions: width, height and depth. Moreover, the neurons in one layer is only partially connected to the neurons in the next. These layers execute operations in order to modify the data with the intent of learning features specific to the data and the most prevalent layers among them are convolutional, activation or ReLU, and pooling. Finally, the last output will be reduced to a probability for the object on the image being what the algorithm predicts it is. In this paper, I have reshaped the dataset into 3 dimensions to feed it as the input for the CNN model. The limited dataset created some doubts but the results showed something quite different.

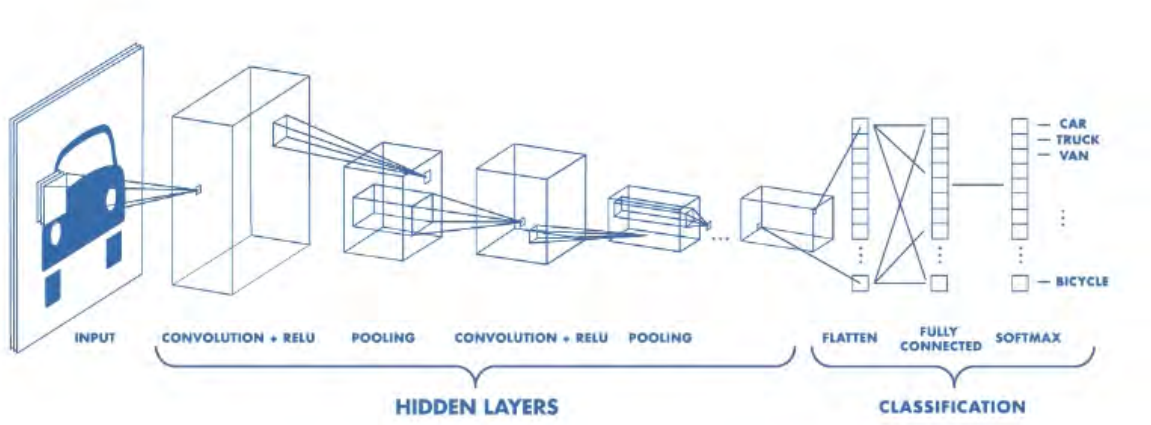


Fig. 3.6 Schematic representation of a Convolutional Neutral Network. (Source: "Introduction to Deep Learning: What Are Convolutional Neural Networks?")

Chapter 4

Result Analysis

The performance of the algorithms differed with and without principal component analysis implementation on the dataset. Analysis and comparison of the performance of different models implemented on the test portion of the dataset was evaluated across various performance metrics.

4.1 Performance Metrics

This paper deals with classification problem and therefore the chosen performance metrics primarily focus on classification. For the detection of breast cancer, if the target variable is 1 then it is a positive instance, meaning the patient has a malignant tumor and therefore cancer. And if the target variable is 0, then it a negative instance, meaning the tumor is benign and the patient does not have cancer.

4.1.1 Confusion matrix

The Confusion matrix is one of the most intuitive and easiest metrics used for finding the correctness and accuracy of the model. It is used for classification problems where the output can be of two or more types of classes which makes it perfect for this paper. The table layout or the matrix layout helps to visualize the performance of an algorithm. Each row of the matrix in table 4.1 represents the instances in an actual class while each column represents the instances in a predicted class or vice versa [18].

Table 4.1 Confusion Matrix

	Predictive Negative	Predictive Positive
Actual Negative	True Negative (TN)	False Positive (FP)
Actual Positive	False Negative (FN)	True Positive (TP)

Terms associated with Confusion matrix:

1. True Positives (TP): True positives are the cases when the actual class of the data point was True(1) and the predicted is also True(1) Ex: The case where a person is actually having malignant (1) tumor and the model classifying his case as malignant (1) comes under True Positive.

2. True Negatives (TN): True negatives are the cases when the actual class of the data point was False (0) and the predicted is also False (0). Ex: The case where a person having benign (0) tumor and the model classifying his case as benign (0) comes under True Negatives.

3. False Positives (FP): False positives are the cases when the actual class of the data point was False (0) and the predicted is True (1). False is because the model has predicted incorrectly and positive because the class predicted was a positive one (1). Ex: A person having a benign (0) tumor and the model classifying his case as malignant (1) comes under False Positives.

4. False Negatives (FN): False negatives are the cases when the actual class of the data point was True (1) and the predicted is False (0). False is because the model has predicted incorrectly and negative because the class predicted was a negative one (0). Ex: A person having malignant (1) tumor and the model classifying his case as benign (0) tumor comes under False Negatives.

The ideal scenario for the model would be when it gives 0 False Positives and 0 False Negatives.

4.1.2 Accuracy

Accuracy in classification problems is the number of correct predictions made by the model over the summation of all different types of predictions made. Accuracy is a good measure when the target variable classes in the data are nearly balanced. Ex: 60% of the data are benign and 40% are malignant.

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN}$$

4.1.3 Precision

Precision is the ratio of True Positives to the summation of True Positives and False Positives. Ex: Precision is a measure of proportion of patients that has been diagnosed as having malignant tumor, actually had malignant tumor. The predicted positives (People predicted as having malignant tumor are TP and FP) and the people actually having a malignant tumor are TP.

$$Precision = \frac{TP}{TP + FP}$$

4.1.4 Recall or Sensitivity

Recall is a measure that shows the proportion of patients that actually had malignant tumor was diagnosed by the algorithm as having malignant tumor. The actual positives (People having malignant tumor are TP and FN) and the people diagnosed by the model having a malignant tumor are TP. Therefore, if we want to focus more on minimizing False Negatives, we would want our Recall to be as close to 100% as possible.

$$Recall = \frac{TP}{TP + FN}$$

4.1.5 F1 Score

F1 Score is the Harmonic Mean between precision and recall. The range for F1 Score is [0, 1]. It shows how precise the classifier is and how robust it is at the same time.

$$F1 = \frac{2 * Precision * Recall}{Precision + Recall}$$

4.1.6 Receiver Operating Characteristics (ROC) Curve

The ROC curve plots the true positive rate (Sensitivity) against the false positive rate (100-Specificity) as its discrimination threshold is varied. All the points on the ROC curve can be regarded as a sensitivity/specificity pair corresponding to a particular decision threshold.

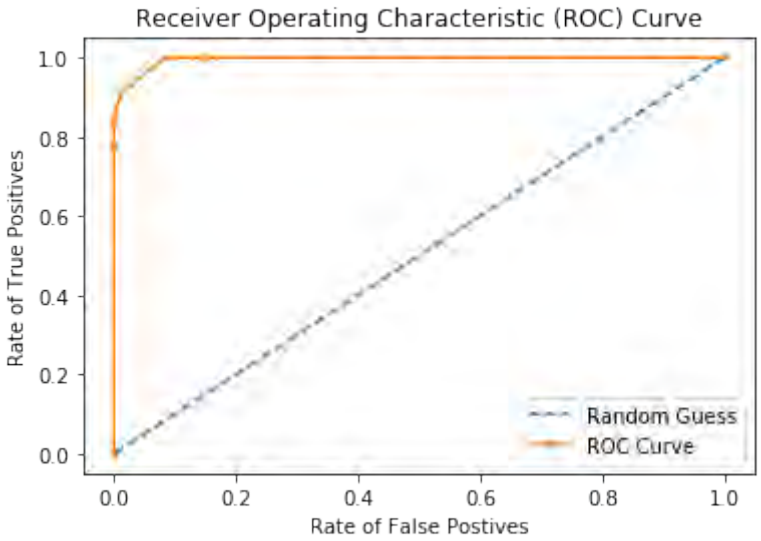


Fig. 4.1 Receiver Operating Characteristics (ROC) Curve.

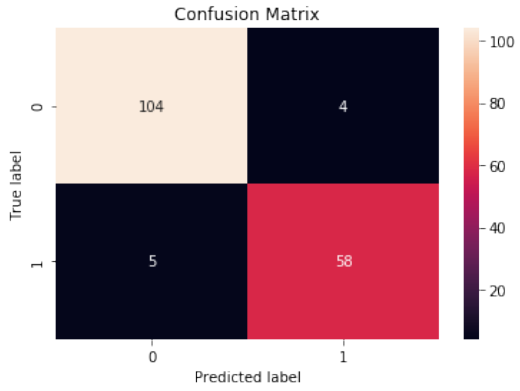
4.1.7 Area under the ROC Curve (AUC)

AUC is the total area under the ROC curve which gives an aggregate measure of the performance across all possible classification thresholds. The value lies between 0 and 1. It is one of the most widely used evaluation metrics for binary classification problems.

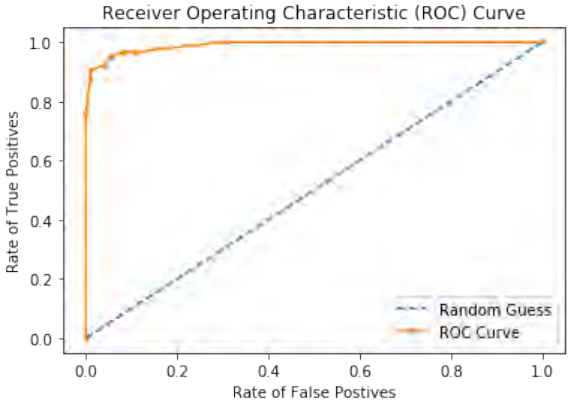
4.2 Model Performances

4.2.1 Random Forest (RF)

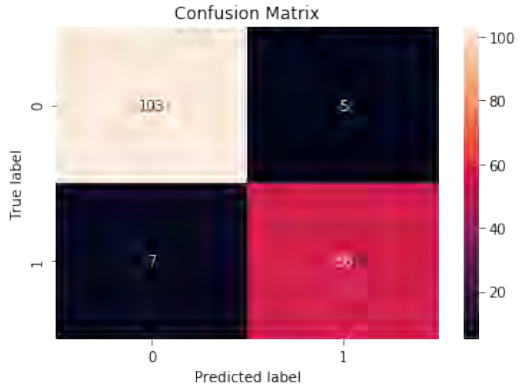
The Accuracy of Random Forest model after applying 10-folds cross validation was 94.11% whereas after applying PCA and standardization, the Accuracy dropped sharply to 89.32%. Along with Accuracy, all of the performance metrics Precision, Recall and F1 score falls after the introduction of PCA which can be seen from figure 4.2a representing the confusion matrix of the model before PCA and figure 4.2c representing after the introduction of PCA. The ROC curve in figure 4.2b has a higher AUC than figure 4.2d.



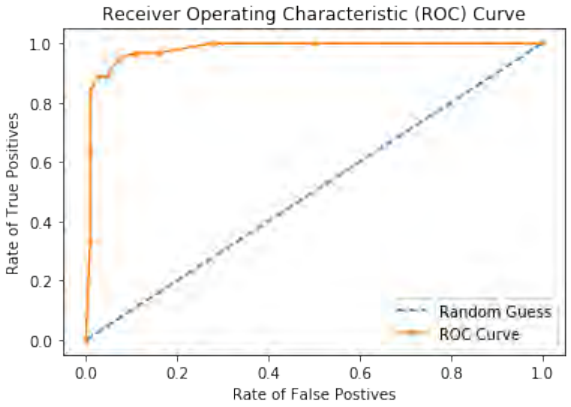
(a) Confusion Matrix Without PCA



(b) ROC Curve Without PCA



(c) Confusion Matrix With PCA

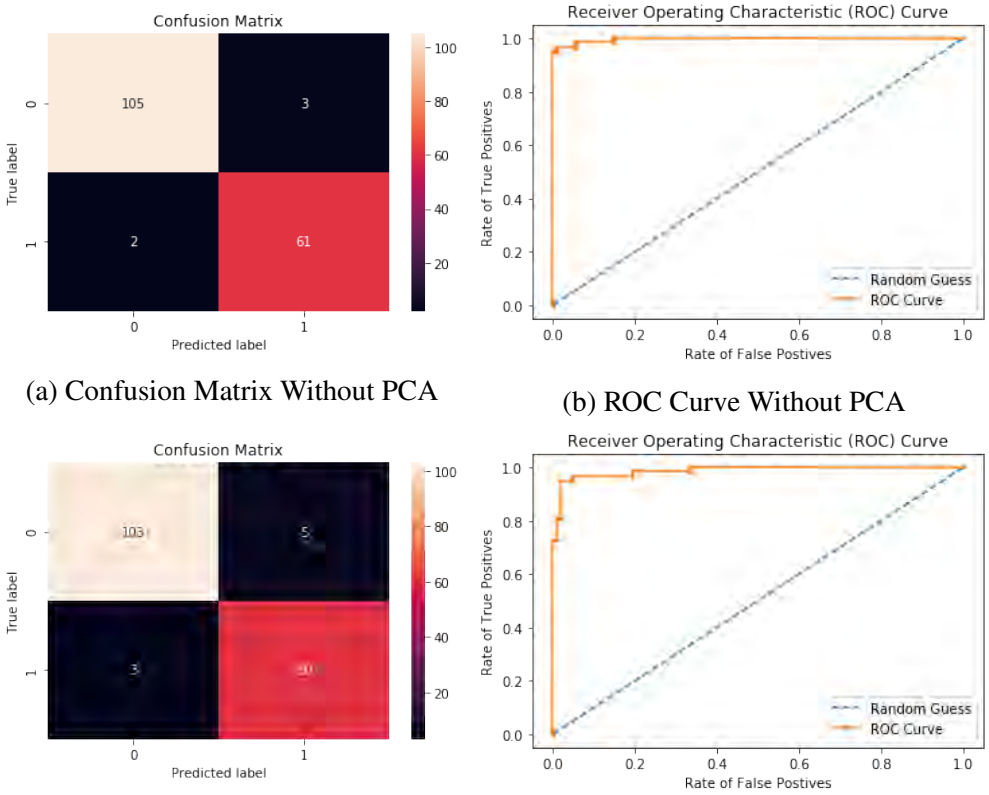


(d) ROC Curve With PCA

Fig. 4.2 Performance Comparison of Random Forest

4.2.2 Support Vector Machine (SVM)

The Accuracy of Support Vector Machine model after applying 10-folds cross validation was 97.05% whereas after applying PCA, the Accuracy dropped to 93.48%. Precision, Recall and F1 scores fell from 97.34% to 94.17%, 97.05% to 93.48% and 97.01% to 93.55% respectively after the introduction of PCA which can be seen from figure 4.3a representing the confusion matrix of the model before PCA and figure 4.3c representing after the introduction of PCA for the Support Vector Machine model. The AUC in figure 4.3b is 99.70% which is higher than in figure 4.3d at 98.80%. We can safely say that the SVM model is better before the introduction of PCA.



(a) Confusion Matrix Without PCA

(b) ROC Curve Without PCA

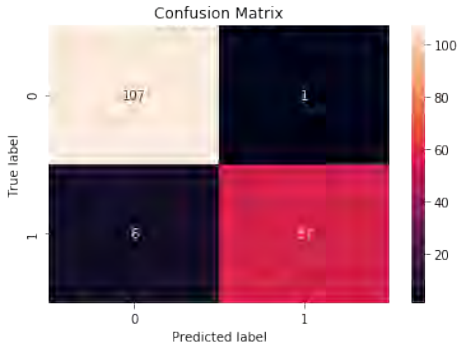
(c) Confusion Matrix With PCA

(d) ROC Curve With PCA

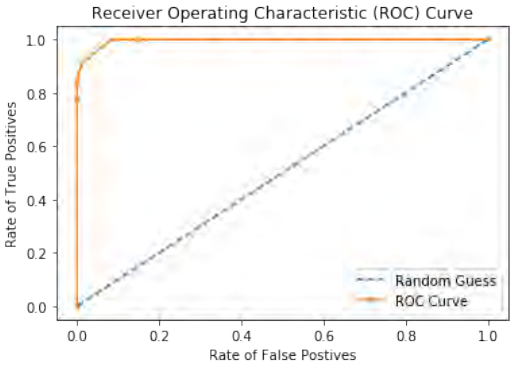
Fig. 4.3 Performance Comparison of Support Vector Machine

4.2.3 K-Nearest Neighbors (KNN)

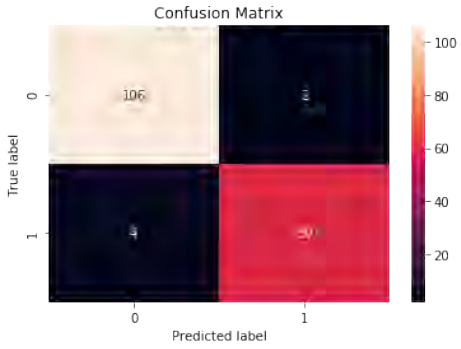
K-Nearest Neighbors model, after applying 10-folds cross validation, gave an impressive Accuracy of 97.12% which is the highest among the supervised learning models in this paper. However, the Accuracy dropped after applying PCA and standardization by almost 2%. Along with Accuracy, the performance metrics Precision and F1 score fell after the introduction of PCA which can be seen from figure 4.4a representing the confusion matrix of the model before PCA and figure 4.4c representing after the introduction of PCA. The AUC in figure 4.4b is 99.5% which is higher than in figure 4.4d at 98.80%. K-Nearest Neighbors algorithm has better numbers for this model without the use of PCA.



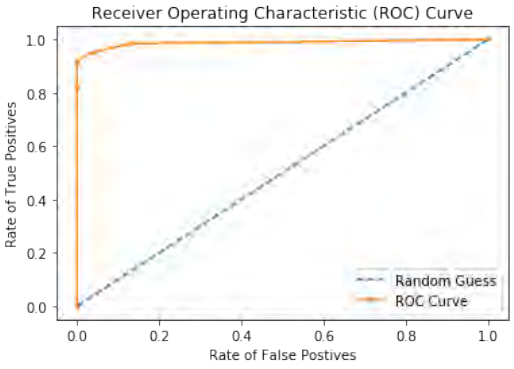
(a) Confusion Matrix Without PCA



(b) ROC Curve Without PCA



(c) Confusion Matrix With PCA

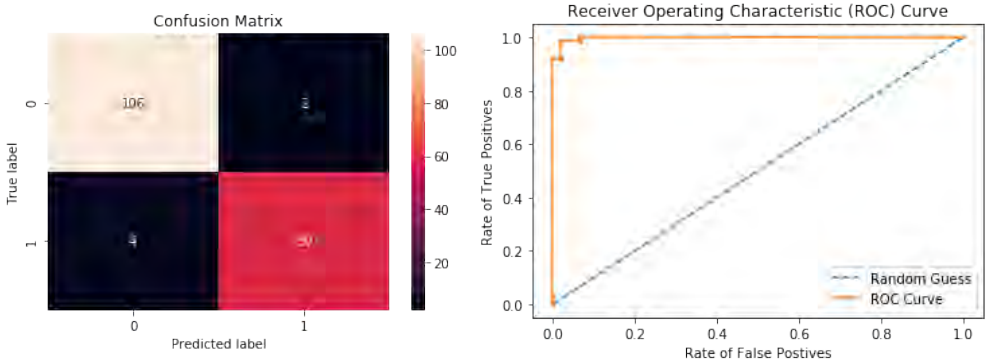


(d) ROC Curve With PCA

Fig. 4.4 Performance Comparison of K-Nearest Neighbors

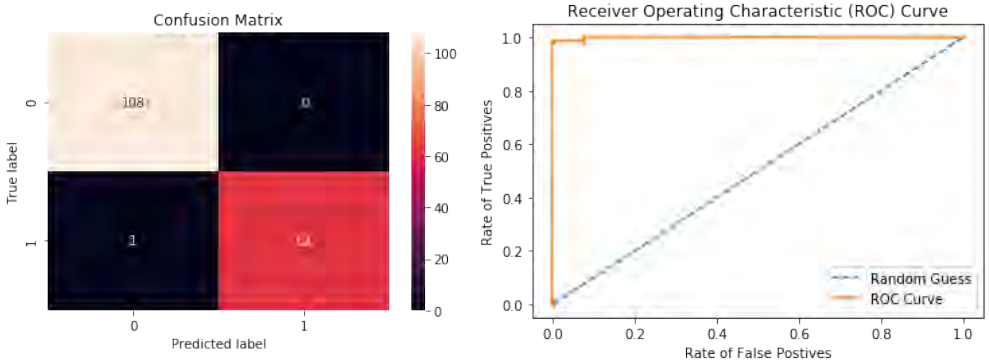
4.2.4 Logistic Regression (LR)

Figure 4.5a and figure 4.5c illustrates the confusion matrix of the Logistic Regression model without and with PCA applied on the dataset. Initially, the Accuracy of the model after 10-fold cross validation is 96.54% with an AUC score of 99.80%. After applying PCA to the dataset, unlike the rest of the models, the Accuracy increases to an impressive 97.68% with AUC score of 99.90% as seen in figure 4.5b which illustrates the ROC curve of the Logistic Regression model while figure 4.5d shows the ROC curve of the Logistic Regression model with PCA. The results show how Logistic Regression performs well for this problem with a Precision of 97.02% and Recall score of 96.54%. Introduction of PCA leads to an increase in Precision and Recall to 97.94% and 97.68% respectively. Logistic Regression with PCA is the better model among the two.



(a) Confusion Matrix Without PCA

(b) ROC Curve Without PCA



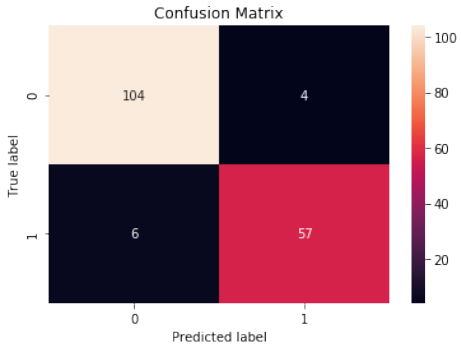
(c) Confusion Matrix With PCA

(d) ROC Curve With PCA

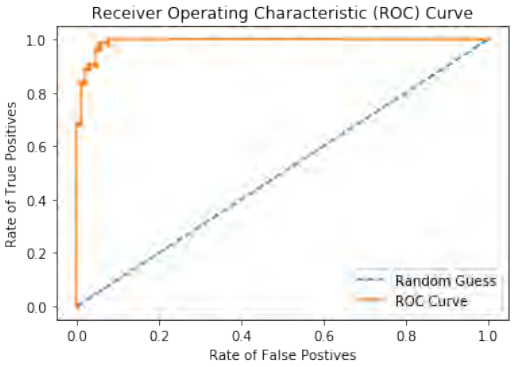
Fig. 4.5 Performance Comparison of Logistic Regression

4.2.5 Naïve Bayes

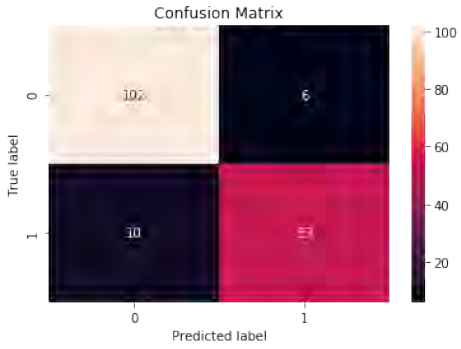
From the figure 4.6a, the confusion matrix of the Gaussian Naïve Bayes model and Figure 4.6c, the confusion matrix of the model after PCA is applied on the dataset, we can deduce that there is not much difference between their performances. The Accuracy of the first is 93.44% and the Accuracy of the latter is 92.89%. The Precision and F1 scores drops slightly from 94.02% to 93.48% and 93.30% to 92.72% respectively. The Recall, which is more important than Precision in disease detection, also fell slightly from 93.45% to 92.89%. The AUC from figure 4.6b is 99.20% and figure 4.6d is 96.70% which fell the most compared to the other metrics for this model.



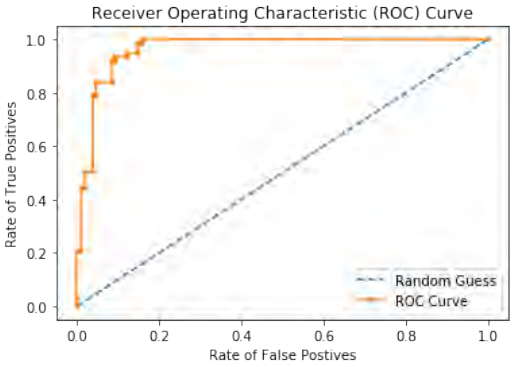
(a) Confusion Matrix Without PCA



(b) ROC Curve Without PCA



(c) Confusion Matrix With PCA

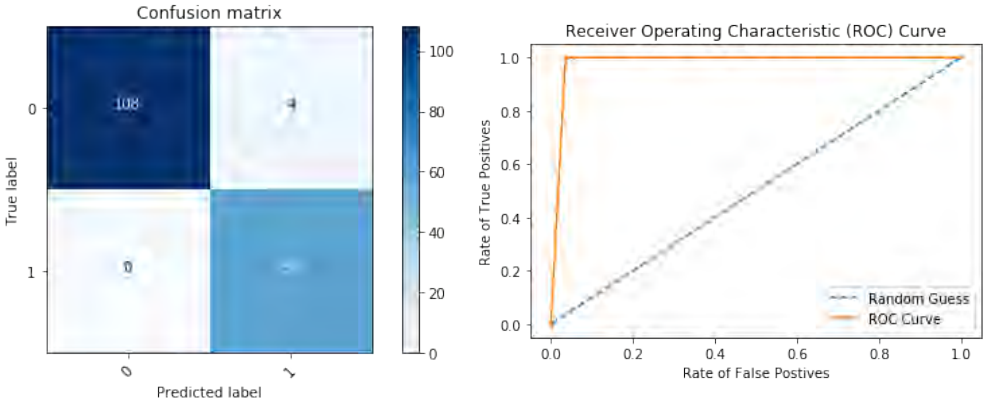


(d) ROC Curve With PCA

Fig. 4.6 Performance Comparison of Gaussian Naive Bayes

4.2.6 Artifical Neural Network (ANN)

Artificial Neural Network performs significantly well with the comparatively small size of the dataset for a neural network giving an Accuracy of 97.66% with a high precision of 93.65% and a F1 score of 96.72%. Recall is more important than Precision in disease detection, because the cost of missing a positive is more problematic than the cost of including a negative and this model gives a perfect Recall score of 100%. Figure 4.7a shows that there were no false negatives for this model. Figure 4.7b shows the ROC curve and the area under the curve (AUC) is 98.2%.



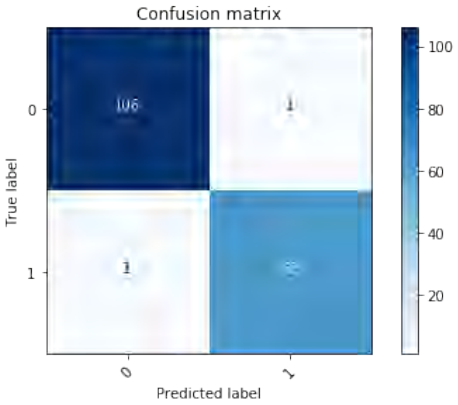
(a) Confusion Matrix Without PCA

(b) ROC Curve Without PCA

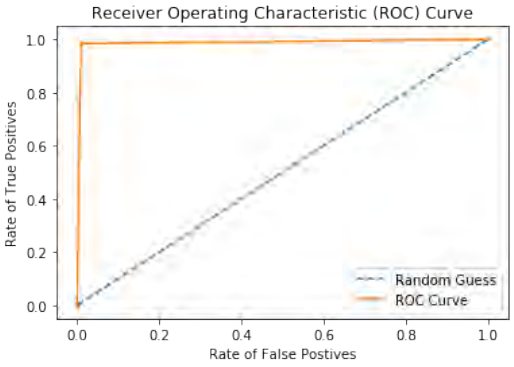
Fig. 4.7 Performance Comparison of Artificial Neural Network

4.2.7 Convolutional Neural Network (CNN)

The main advantage of Neural Network lies in their ability to outperform nearly every other Machine Learning algorithms, but this goes along with some disadvantages like the need of large dataset and high computing power. Despite the small size of the dataset, Convolutional Neural Network performed outstandingly well as you can see from the figure 4.8a which shows the confusion matrix for the CNN model. Accuracy of 98.83% with a high precision of 98.44% and a F1 score of 98.44% is achieved from this model. Recall is more important than Precision in disease detection and this model gives a Recall score of 98.44%. Figure 4.8b shows the ROC curve and the area under the curve (AUC) is 98.8%.



(a) Confusion Matrix Without PCA



(b) ROC Curve Without PCA

Fig. 4.8 Performance Comparison of Convolutional Neural Network

4.3 Discussion

After completing the implementation of all seven algorithms for detecting breast cancer from the dataset, the results can be compared from the table 4.2 and table 4.3 using the performance metrics discussed previously.

Table 4.2 Comparison of Scores of Various Models without PCA

	Accuracy	Precision	Recall	F1 Score	AUC
Random Forest	94.11%	94.97%	94.11%	94.04%	99.0%
Support Vector Machine	97.05%	97.34%	97.05%	97.01%	99.7%
K-Nearest Neighbors	97.12%	97.50%	97.12%	97.06%	99.5%
Logistic Regression	96.54%	97.02%	96.54%	96.51%	99.8%
Gaussian Naïve Bayes	93.44%	94.02%	93.45%	93.30%	99.2%
Convolutional Neural Network	98.83%	98.44%	98.44%	98.44%	98.8%
Artificial Neural Network	97.66%	93.65%	100%	96.72%	98.2%

From table 4.2, it is clear that the deep learning models outperform Gaussian Naïve Bayes and Random Forest by a clear margin whereas Support Vector Machine, K-Nearest Neighbors and Logistic Regression performs close to Artificial Neural Network's performance. Although ANN has worse Precision score, it has a perfect Recall score which is more important in terms of cancer detection. K-Nearest Neighbors high performance scores of 97.12% Accuracy, 97.50% Precision and 97.12% Recall comes in at third best without implementation of PCA. Gaussian Naïve Bayes has the lowest performance score in all categories among the seven algorithm for this dataset with 93.44% Accuracy, 94.02% Precision, 93.45% Recall and 93.30% F1 Score. Among the top 2 performers, CNN has the highest Accuracy of 98.83% and Precision of 98.44% but although ANN has a lower Accuracy of 97.66%, it has a perfect Recall Score of 100% meaning it does not give any false negatives.

After the implementation of PCA, we can see from table 4.3 that the performance of Random Forest and SVM falls significantly across all metrics. Whereas, the Accuracy, Precision and Recall scores of K-Nearest Neighbors decreases to 95.32%, 95.74%, 95.32%. Gaussian Naïve Bayes performs worse with PCA but is replaced by Random Forest as the worst performer among the group with an Accuracy of 89.32%, Precision of 90.38% and Recall 89.32%. However, Logistic Regression performs outstandingly well with PCA and achieves a higher Accuracy of 97.68% but still has a lower Precision of 97.68% compared to ANN without PCA.

Table 4.3 Comparison of Scores of Various Models with PCA

	Accuracy	Precision	Recall	F1 Score	AUC
Random Forest	89.32%	90.38%	89.32%	89.14%	98.1%
Support Vector Machine	93.48%	94.17%	93.48%	93.55%	98.8%
K-Nearest Neighbors	95.32%	95.74%	95.32%	95.26%	98.8%
Logistic Regression	97.68%	97.94%	97.68%	97.67%	99.9%
Gaussian Naïve Bayes	92.89%	93.48%	92.89%	92.72%	96.7%

After taking all the results into account from both table 4.2 and table 4.3, it can be concluded that CNN and ANN achieved higher performance scores in all categories with the limited amount of data considering they are deep learning models but after applying PCA, Logistic Regression outperforms ANN in a couple of performance metrics.

Chapter 5

Conclusion

Breast cancer, the most common cancer among women, being responsible for 69% of death related to cancer among the same gender gives a glimpse of the magnitude of the problem. The early detection of breast cancer can lead to chances of survival of a large number of these people by receiving clinical treatments on time. This paper shows the comparative analysis of different machine learning algorithms in detecting breast cancer from a digitized image of a fine needle aspirate (FNA) of a breast mass. The simple, safe, accurate, and inexpensive procedure of FNA combined with the predictive model in this paper can be used for prognosis, diagnosis and assist doctors in making the final decision more accurately in shorter time span with less human and monetary resource. The performance of the deep learning models show promising results for a near perfect detection system. However, the lack of instances of the data and the conversion of data in order to use it for CNN proved to be a challenging part.

In the future, the model can be perfected with the increase in availability of data and most importantly the growth of data. Deep learning models perform proportionately well to the amount of data which means there is space to improve with the availability of a large dataset.

As written in the holy book of Islam, the Quran: “Whoever saves one life, it is written as if he has saved all humanity.”, this paper explores the field of machine learning in order to integrate it in the medical field as a means for early detection of breast cancer which can ultimately result in a complete clinical system helping save lives.

References

[1] Abdi, H. and Williams, L. J. (2009). Principal component analysis. In *Encyclopedia of Biometrics*.

[2] Altman, N. S. (1992). An introduction to kernel and nearest-neighbor nonparametric regression. *The American Statistician*, 46(3):175–185.

[3] Barandiaran, I. (1998). The random subspace method for constructing decision forests. *IEEE transactions on pattern analysis and machine intelligence*, 20(8).

[4] Bellaachia, A. and Guven, E. (2006). Predicting breast cancer survivability using data mining techniques. *Age*, 58(13):10–110.

[5] Delen, D., Walker, G., and Kadam, A. (2005). Predicting breast cancer survivability: a comparison of three data mining methods. *Artificial intelligence in medicine*, 34(2):113–127.

[6] Ferlay, J., Shin, H., Bray, F., Forman, D., Mathers, C., and Parkin, D. (2011). Globocan 2008, cancer incidence and mortality worldwide: Iarc cancerbase no. 10. lyon, france: International agency for research on cancer; 2010. *Disponible en: URL: http://globocan. iarc. fr*.

[7] Friedman, J., Hastie, T., and Tibshirani, R. (2001). *The elements of statistical learning*, volume 1. Springer series in statistics New York, NY, USA:.

[8] Gholami, V., Chau, K., Fadaee, F., Torkaman, J., and Ghaffari, A. (2015). Modeling of groundwater level fluctuations using dendrochronology in alluvial aquifers. *Journal of hydrology*, 529:1060–1069.

[9] Ho, T. K. (1995). Random decision forests. In *Document analysis and recognition, 1995., proceedings of the third international conference on*, volume 1, pages 278–282. IEEE.

[10] Hosmer Jr, D. W., Lemeshow, S., and Sturdivant, R. X. (2013). *Applied logistic regression*, volume 398. John Wiley & Sons.

[11] Hussain, M., Wajid, S. K., Elzaart, A., and Berbar, M. (2011). A comparison of svm kernel functions for breast cancer detection. In *2011 Eighth International Conference Computer Graphics, Imaging and Visualization*, pages 145–150. IEEE.

[12] Janghel, R., Shukla, A., Tiwari, R., and Kala, R. (2010). Breast cancer diagnosis using artificial neural network models. In *Information Sciences and Interaction Sciences (ICIS), 2010 3rd International Conference on*, pages 89–94. IEEE.

[13] Kharya, S. and Soni, S. (2016). Weighted naive bayes classifier: A predictive model for breast cancer detection. *International Journal of Computer Applications*, 133(9):32–7.

[14] Kohavi, R. (1998). Glossary of terms. *Special issue on applications of machine learning and the knowledge discovery process*, 30(271):127–132.

[15] Liu, L. (2018). Research on logistic regression algorithm of breast cancer diagnose data by machine learning. In *2018 International Conference on Robots & Intelligent System (ICRIS)*, pages 157–160. IEEE.

[16] Lundin, M., Lundin, J., Burke, H., Toikkanen, S., Pylkkänen, L., and Joensuu, H. (1999). Artificial neural networks applied to survival prediction in breast cancer. *Oncology*, 57(4):281–286.

[17] Marsilin, J. R. and Jiji, G. W. (2012). An efficient cbir approach for diagnosing the stages of breast cancer using knn classifier. *Bonfring International Journal of Advances in Image Processing*, 2(1):01–05.

[18] Powers, D. M. (2011). Evaluation: from precision, recall and f-measure to roc, informedness, markedness and correlation.

[19] Rennie, J. D., Shih, L., Teevan, J., and Karger, D. R. (2003). Tackling the poor assumptions of naive bayes text classifiers. In *Proceedings of the 20th international conference on machine learning (icml-03)*, pages 616–623.

[20] Samuel, A. L. (1959). Some studies in machine learning using the game of checkers. *IBM Journal of research and development*, 3(3):210–229.

[21] Torre, L. A., Bray, F., Siegel, R. L., Ferlay, J., Lortet-Tieulent, J., and Jemal, A. (2015). Global cancer statistics, 2012. *CA: a cancer journal for clinicians*, 65(2):87–108.

[22] Uddin, A. K., Khan, Z. J., and Johirul Islam, A. (2013). Cancer care scenario in bangladesh. *South Asian journal of cancer*, 2(2):102.

[23] van Gerven, M. and Bohte, S. (2018). *Artificial neural networks as models of neural information processing*. Frontiers Media SA.

[24] WHO (2016). Breast cancer: prevention and control.

[25] Wolberg, W., Street, W. N., and Mangasarian, O. L. (1992). Breast cancer wisconsin (diagnostic) data set [uci machine learning repository].

[26] Wu, X., Kumar, V., Quinlan, J. R., Ghosh, J., Yang, Q., Motoda, H., McLachlan, G. J., Ng, A., Liu, B., Philip, S. Y., et al. (2008). Top 10 algorithms in data mining. *Knowledge and information systems*, 14(1):1–37.

[27] Xiao, Y., Wu, J., Lin, Z., and Zhao, X. (2018). Breast cancer diagnosis using an unsupervised feature extraction algorithm based on deep learning. In *2018 37th Chinese Control Conference (CCC)*, pages 9428–9433. IEEE.

[28] Zemouri, R., Omri, N., Devalland, C., Arnould, L., Morello, B., Zerhouni, N., and Fnaiech, F. (2018). Breast cancer diagnosis based on joint variable selection and constructive deep neural network. In *2018 IEEE 4th Middle East Conference on Biomedical Engineering (MECBME)*, pages 159–164. IEEE.

[29] Zhou, Z.-H., Jiang, Y., et al. (2003). Medical diagnosis with c4. 5 rule preceded by artificial neural network ensemble. *IEEE Transactions on information Technology in Biomedicine*, 7(1):37–42.