Machine Learning In Breast Cancer Prognosis and Prediction

by

Shah Abul Hasnat Chowdhury
17301143
Golam Akbar Faruqee
17301085
Sayeed Hassan
17201051
Golam Mostafa Chowdhury Jawad
19101638

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

Department of Computer Science and Engineering
Brac University
May 2022

Declaration

It is hereby declared that

1. The thesis submitted is my/our own original work while completing degree at Brac University.

2. The thesis does not contain material previously published or written by a third party, except where this is appropriately cited through full and accurate referencing.

3. The thesis does not contain material which has been accepted, or submitted, for any other degree or diploma at a university or other institution.

4. We have acknowledged all main sources of help.

Student's Full Name & Signature:

Hasnat Chy

Shah Abul Hasnat Chowdhury
17301143

Golam Akbar Faruqee
17301085

S Hassan

Sayeed Hassan
17201051

Golam Mostafa

Golam Mostafa Chowdhury Jawad
19101638

Approval

The thesis/project titled "Machine Learning in Breast Cancer Prognosis and Prediction" submitted by

1. Shah Abul Hasnat Chowdhury (17301143)
2. Golam Akbar Faruqee (17301085)
3. Sayeed Hassan (17201051)
4. Golam Mostafa Chowdhury Jawad (19101638)

Of Spring, 2022 has been accepted as satisfactory in partial fulfillment of the requirement for the degree of B.Sc. in Computer Science on May 19, 2022.

Examining Committee:

Supervisor:
(Member)

Shakila Zaman
Lecturer
Department of Computer Science and Engineering
BRAC University

Co-Supervisor:
(Member)

Faisal Bin Ashraf
Lecturer
Department of Computer Science and Engineering
Brac University

Head of Department:
(Chair)

Sadia Hamid Kazi, PhD
Chairperson and Associate Professor
Department of Computer Science and Engineering
Brac University

Ethics Statement (Optional)

This is optional, if you don't have an ethics statement then omit this page

Abstract

In the human body, cancer is a condition that causes cells to proliferate quickly and uncontrolled across the whole body. It has the ability to arise in any of the billions of cells that build up the human body. Human cells generally become divided and turn into new cells as the requirement for human body. When cells get harmed or turn aged, they perish, and young vesicle replace them. Cancer can take many forms. Cancer is normally designated after the limb or tissues in which it arises. For instance, kidney cancer starts in the kidney, blood cancer starts in the blood cells and breast cancer starts in the human breasts. Cancer in breast is the maximal prevalent and frequent disease in female population all over the world. The majority of women identified with breast cancers are just above 50 in age, but breast cancer may strike anybody at any age. In the developed world, in one out of every eight women is diagnosed with breast cancer. However, early detection can help to prevent deaths and save many lives. This paper focuses on prediction and prognosis of cancer in breast using ML models where the paper provides accuracy of the ML deep learning models in diagnostically identifying 569 patients where 212 malignant and 357 benign Fine Needle Aspirate (FNAs) and its potential accuracy. Also, Recall and the feature numbers in the database is obtained, which is depicted visually. First of all, we have given an overview of ML and deep learning approaches including DT, KNN and Linear SVC and ANN. We examine their BC implications. The Wisconsin breast cancer database (WBCD) is a standard database for assessing results using multiple techniques. This data set shows features such as tumor radius, concavity, texture and fractal dimensions also defined the tumor as Benign or Malignant. After implementing our selected models we find out the most efficient model with respect to precision, recall, F1 score accuracy and confusion metric. We observed that ANN obtains the height accuracy, which is 97.9%. We provided the necessary statistics and graphs in our result part in this paper. We believe that our results may assist lead to more accurate and guided screening in the future.

Keywords: Machine Learning; Linear SVC; DT; ANN; KNN Breast Cancer;

Dedication

To our parents, who have always been there to support, encourage and inspire us to strive for the best in whatever we do.

Acknowledgement

First of all, all praise to the Great Allah because of whom our thesis have been completed without any egregious difficulties.

Secondly, to Shakila Zaman mam who is our advisor is this study, because of her great support and advice in our work. She helped us whenever we needed help. .

Thirdly, Faisal Bin Ashraf sir who also help us by giving many feedbacks.

And lastly to our parents except their support in all respects it may not come to light. Because of their great corroboration and entreaty, we are now on the brink of our graduation.

Table of Contents

List of Figures

List of Tables

Nomenclature

The next list describes several symbols & abbreviation that will be later used within the body of the document

AUC	Area under ROC curve
DT	Decision Tree
FN	false positive
FP	false positive
FPR	false positive rate
KNN	k-nearest neighbors
ROC	Receiver operating characteristic
SVC	Linear Support Vector Classifier
TN	true negative
TP	true positive
TPR	true positive ratel

Chapter 1

Introduction

1.1 ML in Breast Cancer

Nowadays, machine learning is being used massively in the field of research. Lots of surveys and research are being conducted using machine learning approaches. For instance, in agriculture sector, economic sector, industrial sector, medical sector, educational sector conducting research by machine learning technique. ML is a very important sector of Artificial intelligence which engages in different kinds of probabilistic, statistical and optimization technics that allow machines to "learn" from previous instances and to identify difficult to perceive models from huge, obstreperous or complicated dataset [9]. ML is playing a significant role in medical diagnoses. ML uses statistical techniques for enabling machines to enhance with expertise. Machine sight is an undisputed component of pathological dispensation. This includes presage of advancement and reaction, to gather knowledge from clinical routine information and to dig up fatalistic sickness specimen through the diagnosis methodologies including Oxford Electromagnetic Acoustic imaging (Ox- EMA), Hyper Spectral (HS), MRI, ultra sonogram, CT scan, microscopic images and histology etc. However, we can prognoses and predict cancer using ML techniques. By implementing ML techniques in different sorts of data sets (X-ray reports, images, surveys, symptoms etc.) we can predict cancer in the human body. In accordance with the worldwide cancer statistics: 2020 GLOBOCAN approximates of occurrence and fatality globally for 36 types of Cancers have been found in 185 Countries. Worldwide, an approximated 19.3 million latest cancer occurrences and nearly 10.0 million cancer demises happened in 2020. In accordance with the most recent survey of PubMed, statistics exceeding 1500 reports have been revealed on the basis of ML and cancer. Added to that, most of them are about to perceive, categorize, track out or characterize lump and other fatality using ML. Basically, those works are about cancer identity and pathology. Very few papers have been revealed on cancer prediction and prognosis. However, cancer prediction and prognosis is not the same as cancer detection and diagnosis. Indeed, cancer prediction means the identifying of a person or people of prophetic genetically interchange might favor in prescribing the result of cancer treatment, comply for the allowance of patients into distinguished groups for recherché therapeutic regulations. And prognosis is basically the possibility that the distemper will be medicated effectively, and the patient will get well. The most significant part of an individual is affected by cancer are type and trace of the cancer, the period of the disorder, the organs where the cancer has expanded in the body, and the cancer's grade (how dangerous the outlook of the cancer cells in the microscope—a directing of the cancer how fast to enlarge and expand). Another factors that play significant role in prognosis include the biological and genetic features of the cancer affected cells (these features are occasionally named as bio markers that can be ascertained by inelastic lab and imaging exams), the affected person's age and overall general health, and the area where the affected person's cancer reacts for treatment. However, in our research we will focus on cancer prognosis and prediction and our target is to improve the accuracy of cancer susceptibility, sensitivity, cancer outcome prediction and prediction of survival after having cancer-by using machine learning techniques. As we mentioned before, researching in cancer is a very huge field and approximately 36 types of cancer have been found all over the world. It is quite difficult to work on all types of cancers. For the convenience of our research, we narrowed down the topic. Mainly in our study we will work on breast cancer. Breast cancer is the second-biggest risk factor for mortality for women worldwide, with over 8% of women developing the illness over their lifespan [6]. As per a World Health Organization research, over 1 million women will be newly diagnosed with breast cancer per year, with over half a million women dying from the disease [6]. The frequency of breast cancer is expected to rise in the future as the atmosphere continues to deteriorate. In the USA, there had been roughly 182,460 newly diagnosed patients and 40,480 fatalities in 2008 [10]. Because the origins of breast cancer are uncertain, early identification is essential for lowering the death rate. As soon as malignancies are discovered, the enhanced treatment options are available. In shortly, breast cancer is the growth of cancerous tissue in the breast. Symptoms of this disease are breast bulk, changes in breast form and dimension, alterations in breast skin tone, breast pains, and gene changes, among others. Therefore, early identification necessitates a precise and trustworthy diagnostic that can discriminate between benign and malignant tumors. A successful predictive algorithm ought to have a minimal rate of false positives (FP) and false negatives (FN) [8]. Mammography was once the most reliable method for identifying and treating breast cancer. There are a lot of algorithms and techniques are being used in this research field. For instance-RF, PCA, ANN, KNN, Linear SVC, SVM, SSL, DT, YOLO etc. In our study we will focus on ANN, KNN, Linear SVC and DT models and try to find the best outcomes.

1.2 Research Problem

The major obstacles in the breast cancer discovery as well as treatment procedure are re-designing the examined channel, agreeing breast cancer growth flashes, developing preclinical structure, which precisely handles complicated cancers, aforetime treatment, advanced forms of modeling and inflicting clinical examinations, and amending delicacy which could be crucial for croakers as an extra and in advance judgment. Notwithstanding, no ML mortality hazard bodement algorithm has been validated or compared with much used prognostic hands in oncology. It is not clear how various instruments erudition algorithms assimilate and if they can tastily clinicians to have timely colloquies about treatment and end-of- life preferences.

1.3 Research Objectives

This study targets to develop a cancer prognosis and prediction method that will determine if a person is with breast cancer or not by comparing factors. Such as, medical history of a patient, major symptoms and test report including Lab reports, Genetic tests, biopsies of tumor etc. All these works will be conducted using Machine Learning approaches. In this study we have selected four techniques of machine learning, they are ANN, KNN, Linear SVC and DT. Here is the main objectives of our paper :
1. Our first objective is comparing the accuracy of ANN, KNN, Linear SVC and DT in predicting breast cancer.
2. Then, we will observe which method is giving us a better outcome against our dataset in this case.
3. After that, we will analyze everything and try to find out the efficient method to predict breast cancer.

Chapter 2

Literature Review

ML has been used in Cancer Prognosis for a long time and it became a very popular tool. By using many techniques like identifying relationships and patterns from many datasets, the future of Cancer can be predicted. In this paper [26], we see analyzing circulating miRNAs, which have been shown to be a potential class for cancer diagnosis and detection. There are many related works [28], [24], [30], like screening gene expression signatures are also one of the major works behind prediction of cancer. According to a different study,[29] Clustering and regression are two more common machine learning challenges that people face. In instances involving regression, a learning function is used to convert the data into a variable with a real-value. Based on this approach, the value of a predictive variable may then be calculated for each new sample. The objective of a common unsupervised task known as clustering is to find categories or clusters of data items in order to classify those data objects. Using this strategy, each new sample may be placed into one of the previously established clusters based on the similarities in characteristics that they all exhibit. Moreover, gene expression profiles, clinical factors, and histological characteristics are also included in virtually all research as supplementary inputs to the prognostic process. Decision Trees (DTs), Artificial Neural Networks (ANNs), Bayesian Networks (BNs) and Support Vector Machines (SVMs) are among the methodologies utilized in cancer research to construct predictive models, resulting in efficient and optimum decision-making[27]. Machine learning algorithms have the potential to enhance our understanding of cancer development, but they must first be effectively validated before being put to use in clinical settings. From this paper [17], they used J48 decision trees and also showed the performance analysis where they create pruning trees, J48 employs an algorithm from Decision Tree. It's possible that J48 is responsible for the growth of these trees. Malignant tumor is a term that is used to categorize or classify a patient. Data entropy is a notion used in the data mining technique. Each data feature is utilized. Divide the data into smaller parts to come up with a decision. The error rates and accuracy of breast cancer data with 699 tuples and attributes of 10 different types were investigated which shows Correctly Classified Instances of 661 instances and Percentage is 94.5637%. In this paper [22], they showed the accuracy and the test findings confirmed that the FT classification with the greatest properly numbers instance which is 550 has the highest high accuracy 97.7%. Also, it showed that FT is better compared to other five algorithms. From this article [31], author used 699 samples and used decision tree algorithms and accuracy is 99.55% given by 10- fold cross validation to measure

the average error rate.
Latchoumiet al.[23] found that WPSO with SSVM for classification achieved 98.42%. SVM can predict breast cancer better than Naive Bayes, according to Asri et al [13]. Osman et al.[11] used a two-step clustering technique with an effective probabilistic vector support machine to analyze the WBCD with a 99.10 percent classification accuracy. Naive Bayes (NB), Support Vector Machine (SVM), Decision Tree (C4.5) and K-Nearest Neighbor (KNN) Machine Learning (ML) algorithms were compared in terms of performance [19]. This study employed the Wisconsin Diagnosis of Breast Cancer (WDBC) dataset [5]. The best result was 97.13 percent using the SVM algorithm. Using decision tree classifier (CART), Christobel. Y and Sivaprakasam got the result of accuracy is 69.23% in breast cancer datasets.
In this paper [25], author compared the accuracy using two algorithms, ANN and logistic algorithm working with ensemble machine learning algorithms to get better accuracy for diagnosing breast cancer. The author Qasem used MCWS (Marked Controller Watershed) algorithm and got 95% accuracy [14].The experiment demonstrated the usefulness of the SVM rejection model in lowering the FP rate when compared to the results obtained without using the SVM rejection model. [21] Shimizu reached 90% accuracy using Neural networks and Deep Learning. [12] Author Aruna obtained 68-79% accuracy using SVM and Nave Bayes with the UCI database. Al-Hadidi, the author, was able to achieve an accuracy rate of 93.7% by using the DWT tool for image filtering and the BPNN tool for processing. [20]. [18] An accuracy of 84.2% was attained using Adaptive Resonance Theory with the UCI database by Ahmad Junaid.

Chapter 3

Workplan

As research data continues to get complex, having technology that can recognize and predict the situation of cancer disease. Machine learning methods are capable of detecting patterns and making predictions with the data that we took from various people. In order to do so, we proposed a model which can detect any abnormality and risk within the people. The below figure provides a simple work plan of the model design. It will help to understand what we are actually planning to do and what we have in our heads.

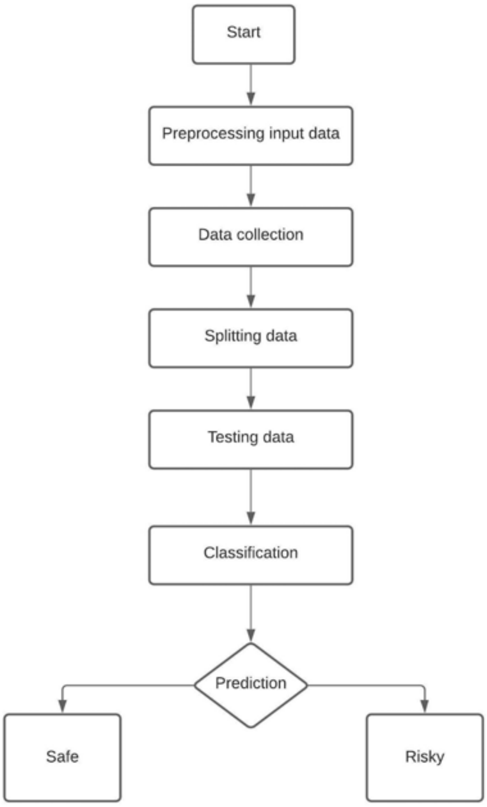
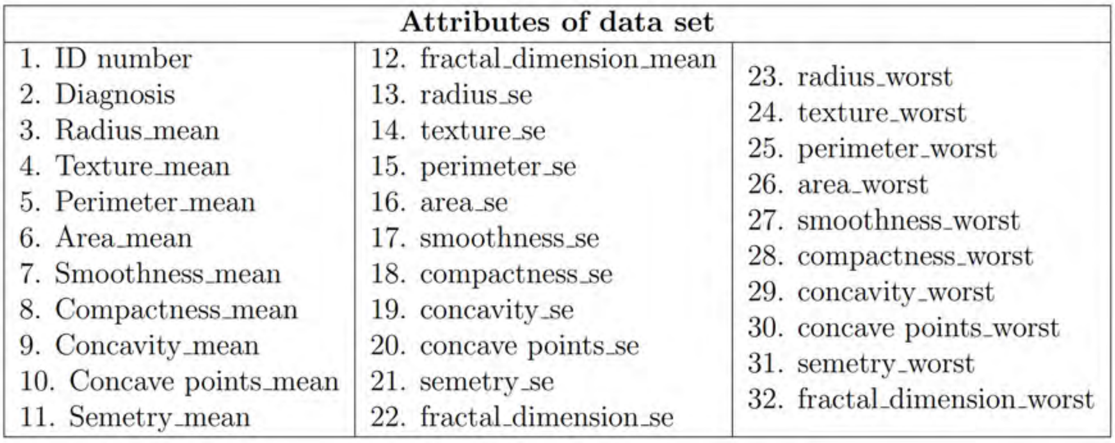


Figure 3.1: Flowchart of the Proposed Cancer Prognosis and Prediction Model

First of all, we will collect data of the patients and then start pre-processing the data. According to our plan, next, we will divide the data into several groups. In the end, we will decide if there is any risk of cancer. If there is risk, we will determine the actual stage of cancer for a specific person.

3.1 Dataset overview

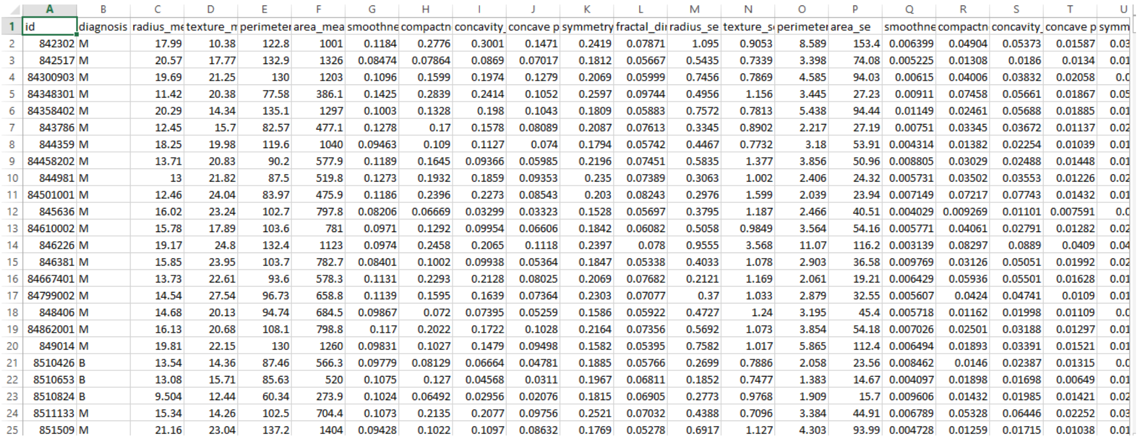
For this research the methods of cancer prediction and prognosis which are SVC, KNN, DT and ANN using machine learning techniques, we are using the Wisconsin breast cancer diagnostic data set for predictive analysis. It is a numerical data set where 32 attributes/column and 570 rows have been found, that means there are 570 individual's data have been stored in this data set. The attributes of our dataset are given in the below table.



Attributes of data set		
1. ID number	12. fractal_dimension_mean	23. radius_worst
2. Diagnosis	13. radius_se	24. texture_worst
3. Radius_mean	14. texture_se	25. perimeter_worst
4. Texture_mean	15. perimeter_se	26. area_worst
5. Perimeter_mean	16. area_se	27. smoothness_worst
6. Area_mean	17. smoothness_se	28. compactness_worst
7. Smoothness_mean	18. compactness_se	29. concavity_worst
8. Compactness_mean	19. concavity_se	30. concave points_worst
9. Concavity_mean	20. concave points_se	31. semetry_worst
10. Concave points_mean	21. semetry_se	32. fractal_dimension_worst
11. Semetry_mean	22. fractal_dimension_se	

Figure 3.2: Attributes of the Dataset

In this data set the Diagnosis attribute is a categorical data where we can find only two features, they are M and B. Here M represents Malignant and B represents Benegin. The rest of the attributes are numerical data. There are total 17 null value in this data set. In radius_mean column there are 9 values are missing, and 8 values are missing in the fractal_dimension_worst column. Our dataset is shown in figure 3.3



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U
1	id	diagnosis	radius_me	texture_m	perimeter	area_mea	smoothne	compactn	concavity_	concave p	symmetry	fractal_di	radius_se	texture_s	perimeter	area_se	smoothne	compactn	concavity_	concave p	symm
2	842302	M	17.99	10.38	122.8	1001	0.1184	0.2776	0.3001	0.1471	0.2419	0.07871	1.095	0.9053	8.589	153.4	0.006399	0.04904	0.05373	0.01587	0.03
3	842517	M	20.57	17.77	132.9	1326	0.08474	0.07864	0.0869	0.07017	0.1812	0.05667	0.5435	0.7339	3.398	74.08	0.005225	0.01308	0.0186	0.0134	0.01
4	84300903	M	19.69	21.25	130	1203	0.1096	0.1599	0.1974	0.1279	0.2069	0.05999	0.7456	0.7869	4.585	94.03	0.00615	0.04006	0.03832	0.02058	0.0
5	84348301	M	11.42	20.38	77.58	386.1	0.1425	0.2839	0.2414	0.1052	0.2597	0.09744	0.4956	1.156	3.445	27.23	0.00911	0.07458	0.05661	0.01867	0.05
6	84358402	M	20.29	14.34	135.1	1297	0.1003	0.1328	0.198	0.1043	0.1809	0.05883	0.7572	0.7813	5.438	94.44	0.01149	0.02461	0.05688	0.01885	0.01
7	843786	M	12.45	15.7	82.57	477.1	0.1278	0.17	0.1578	0.08089	0.2087	0.07613	0.3345	0.8902	2.217	27.19	0.00751	0.03345	0.03672	0.01137	0.02
8	844359	M	18.25	19.98	119.6	1040	0.09463	0.109	0.1127	0.074	0.1794	0.05742	0.4467	0.7732	3.18	53.91	0.004314	0.01382	0.02254	0.01039	0.01
9	84458202	M	13.71	20.83	90.2	577.9	0.1189	0.1645	0.09366	0.05985	0.2196	0.07451	0.5835	1.377	3.856	50.96	0.008805	0.03029	0.02488	0.01448	0.01
10	844981	M	13	21.82	87.5	519.8	0.1273	0.1932	0.1859	0.09353	0.235	0.07389	0.3063	1.002	2.406	24.32	0.005731	0.03502	0.03553	0.01226	0.02
11	84501001	M	12.46	24.04	83.97	475.9	0.1186	0.2396	0.2273	0.08543	0.203	0.08243	0.2976	1.599	2.039	23.94	0.007149	0.07217	0.07743	0.01432	0.01
12	845636	M	16.02	23.24	102.7	797.8	0.08206	0.06669	0.03299	0.03323	0.1528	0.05697	0.3795	1.187	2.466	40.51	0.004029	0.009269	0.01101	0.007591	0.0
13	84610002	M	15.78	17.89	103.6	781	0.0971	0.1292	0.09954	0.06606	0.1842	0.06082	0.5058	0.9849	3.564	54.16	0.005771	0.04061	0.02791	0.01282	0.02
14	846226	M	19.17	24.8	132.4	1123	0.0974	0.2458	0.2065	0.1118	0.2397	0.078	0.9555	3.568	11.07	116.2	0.003139	0.08297	0.0889	0.0409	0.04
15	846381	M	15.85	23.95	103.7	782.7	0.08401	0.1002	0.09938	0.05364	0.1847	0.05338	0.4033	1.078	2.903	36.58	0.009769	0.03126	0.05051	0.01992	0.02
16	84667401	M	13.73	22.61	93.6	578.3	0.1131	0.2293	0.2128	0.08025	0.2069	0.07682	0.2121	1.169	2.061	19.21	0.006429	0.05936	0.05501	0.01628	0.01
17	84799002	M	14.54	27.54	96.73	658.8	0.1139	0.1595	0.1639	0.07364	0.2303	0.07077	0.37	1.033	2.879	32.55	0.005607	0.0424	0.04741	0.0109	0.01
18	848406	M	14.68	20.13	94.74	684.5	0.09867	0.072	0.07395	0.05259	0.1586	0.05922	0.4727	1.24	3.195	45.4	0.005718	0.01162	0.01998	0.01109	0.0
19	84862001	M	16.13	20.68	108.1	798.8	0.117	0.2022	0.1722	0.1028	0.2164	0.07356	0.5692	1.073	3.854	54.18	0.007026	0.02501	0.03188	0.01297	0.01
20	849014	M	19.81	22.15	130	1260	0.09831	0.1027	0.1479	0.09498	0.1582	0.05395	0.7582	1.017	5.865	112.4	0.006494	0.01893	0.03391	0.01521	0.01
21	8510426	B	13.54	14.36	87.46	566.3	0.09779	0.08129	0.06664	0.04781	0.1885	0.05766	0.2699	0.7886	2.058	23.56	0.008462	0.0146	0.02387	0.01315	0.0
22	8510653	B	13.08	15.71	85.63	520	0.1075	0.127	0.04568	0.0311	0.1967	0.06811	0.1852	0.7477	1.383	14.67	0.004097	0.01898	0.01698	0.00649	0.01
23	8510824	B	9.504	12.44	60.34	273.9	0.1024	0.06492	0.02956	0.02076	0.1815	0.06905	0.2773	0.9768	1.909	15.7	0.009606	0.01432	0.01985	0.01421	0.02
24	8511133	M	15.34	14.26	102.5	704.4	0.1073	0.2135	0.2077	0.09756	0.2521	0.07032	0.4388	0.7096	3.384	44.91	0.006789	0.05328	0.06446	0.02252	0.03
25	851509	M	21.16	23.04	137.2	1404	0.09428	0.1022	0.1097	0.08632	0.1769	0.05278	0.6917	1.127	4.303	93.99	0.004728	0.01259	0.01715	0.01038	0.01

Figure 3.3: Dataset

We have found out the correlation of each column in our data set. In figure 3.4, we have shown all the correlation coefficient of each column in the correlation matrix.

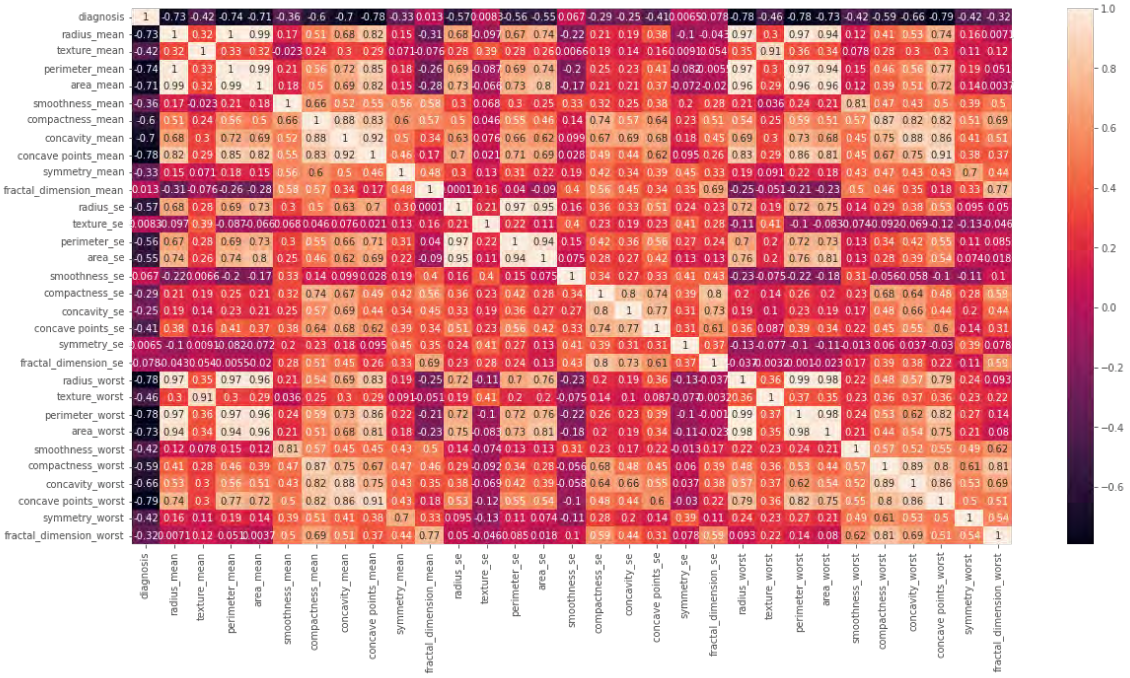


Figure 3.4: Correlation Matrix

3.1.1 DATA PREPROCESSING

We dropped the irrelevant column (feature) such as 'id'. There are a total 17 null values in this data set. In the radius_mean column there are 9 values missing and 8 values are missing in the fractal_dimension_worst column. We dropped these null values from our dataset. To construct our model algorithms, we selected "Diagnosis" column as our target column where we find two object type unique values which are "M" and "B". For our convenience, we converted these object type values into binary type values where "M" is represented by "0" and "B" is represented by "1". We observed that the number of "0" in our target column is 212 and the number of "1" is 357. Figure 3.5 shows the frequency of our target column

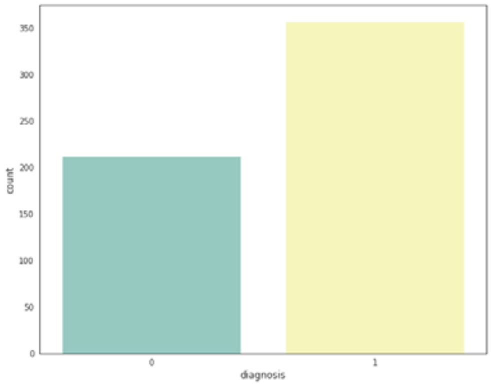


Figure 3.5: Frequency of Cancer Diagnosis

3.1.2 SPLIT DATA

We separated the columns into two categories to educate the model algorithms, they are label and feature. The source columns of the dataset are named features, and the result column we're trying to predict is named label. The label in our dataset is 'diagnostic,' while all the other columns are features. The entire dataset was then separated into training (75%) and testing (25%) sets, with the classes distributed evenly across the dataset.

Chapter 4

Methodology

For this research we are using methods of cancer prediction and prognosis which are Linear SVC, Artificial Neural Network (ANN), Decision Tree (DT) and K-nearest neighbors (kNN) using machine learning technique.

4.1 Linear SVC

The Support Vector Machine (SVM) was first introduced in 1995. Cancer diagnosis and prognosis are increasingly relying on supervised machine learning classification algorithms. In short, SVM's purpose is to develop a model that assumes the key parameters of the sample data given just the sample datasets features. Basically, SVM looks at crucial samples from all classes, which are referred to as support vectors, and then splits them as far as possible using these support vectors to construct a linear function. After that, SVM is used to generate a mapping from an input vector to a heavy space in order to choose the optimum hyper plane for classifying the dataset. [7] They uses this linear classifier to maximize the marginal distance between the selected hyper—plane and the closest data point. Different Kernel functions may be used for the decision function. Custom kernels may be defined as well as the common ones.Talking about kernel, there come SVC. They are essentially various implementations of the same algorithm. The SVM module (e.g., SVC) is a container over the libsvm library and supports several kernels, whereas LinearSVC is built on liblinear and only accepts a linear kernel. Since these implementations are different, in reality users will obtain various outcomes, the most notable ones being that Lin- earSVC only supports a linear kernel, is quicker and can scale a really well. SVC take slightly different sets of parameters and have distinct mathematical formulations. LinearSVC, on the contrary, is another (quicker) deployment of Support Vector Classification for the scenario of a linear kernel. Note that LinearSVC does not take parameter kernel, since this is expected to be linear. It also lacks several of the properties of SVC, such support. The purpose of a Linear SVC (Support Vector Classifier) is to accommodate to such data users supply, providing a "best fit" of support vectors that classifies, or characterizes, all data. From that, after receiving the support vectors, one can then input some characteristics to the classifier and see what the "predicted" category represents. In this paper, we will use Linear SVC.
From the figure 4.1, we can observe the graph view of these models, where 0 represents a loss and 1 represents a win. As a consequence, if the predictive final value is lower than 0.5, the outcome is marked a loss; if it is greater than 0.5, the outcome

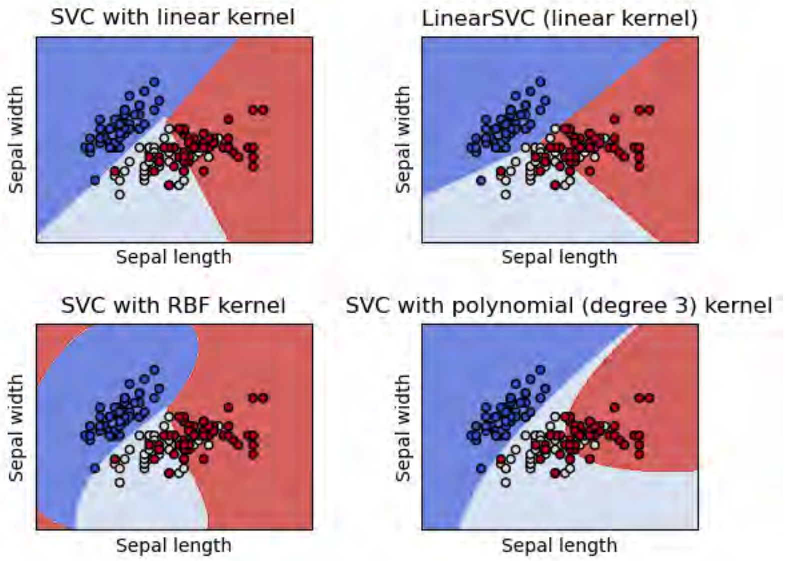


Figure 4.1: Linear SVC

is labeled a win.
The process that is follows by linear SVC is given below:
1.It finds lines or limits that categorise the training dataset.
2. It picks the line with the greatest distance from the closest data points.

The accuracy rate is 95.80% and it is similar with the Decision Tree Algorithm. In addition, it has the highest precision score which is 98.84% compared to others. But, it has the lowest recall score, which is 94.44%. F1 score is 96.60% and it is close to Decision Tree's F1 score.

4.2 Artificial Neural Network

The ANN (Artificial Neural Network) is a computer system designed to automatically perform the following brain functions: B. Generating new information through learning, generation, and discovery [4]. There are typically 3 layers: an input layer, one or more hidden layers, and an outer layer [1]. Each layer contains a certain number of interconnected neurons or nodes. Each neuron connects with other neurons through connection weights and communication links. Signals travel along the neuron due to the weight of the connection. Each neuron receives a set of inputs from other neurons in proportion to the weight of the connection and produces an output signal that other neurons can produce. To create an ANN model, the network goes through two processes, and they are testing and training. For training, the network is trained to predict the output based on the input. We also test storing failures or training data in the network for testing and use to predict the output. When the error under test reaches the desired tolerance, the network training process ends. The most common and widely used algorithm is the Back-Propagation (BP) algorithm [2]. BP is classified into two phases: forward propagation and back-propagation.
The BP learning algorithm uses a linear reduction algorithm. The BP algorithm

is used to improve network performance by changing the weights along the ramp to reduce overall error. When the mean squared error (MSE) stops decreasing and starts to increase, training stops, indicating overtraining.

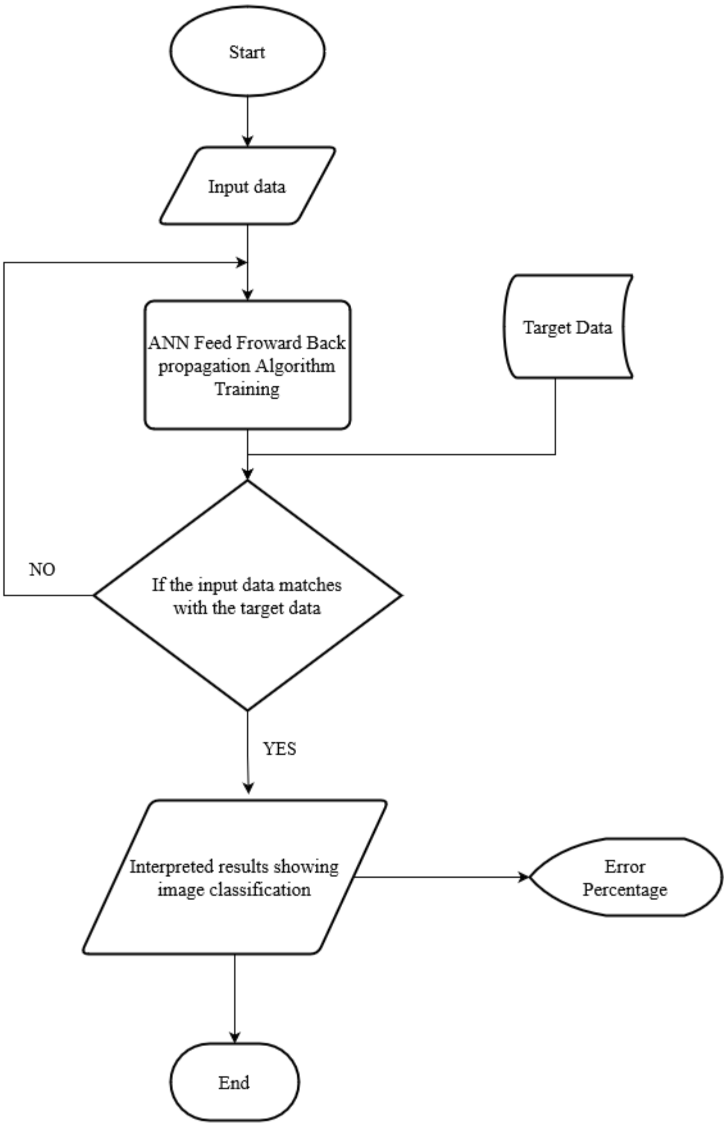

Start
Input data
ANN Feed Froward Back propagation Algorithm Training
Target Data
NO
If the input data matches with the target data
YES
Interpreted results showing image classification
Error Percentage
End

Figure 4.2: Flowchart showing the description of the ANN-FFBP algorithm

Pros:
1. Traditional programming information, for example, is kept throughout the network instead of in a database.
2. Has the ability to work with imperfect knowledge.
3. Damage to one or more cells of the ANN does not interfere with output data generation.
4. ANN makes decisions by studying events and commenting on similar events.

Cons:
1. ANN depends on the implementation of the equipment, as its architecture requires a processor with parallel computing power.

2. When ANN makes an investigation decision, it does not provide a clue as to why or how.
3. There are no specific rules defining the structure of an ANN.
4. The mapping mechanism defined here has a direct impact on network performance, which depends on the capabilities of the user.

The process for ANN algorithm is follows:

1. To begin the procedure, delegate randomized weights to all links.
2. Determine the activation rate of hidden nodes utilizing inputs and links ($input > hiddennodes$).
3. Determine the rate of activation of exit buttons through using activation rates of hidden buttons as well as exit links.
4. Determine the error rate at the exit node and re-calibrate all links between the hidden and exit nodes.
5. Minimize the error at the hidden nodes by using the weights and errors discovered at the exit node.
6. Realign the weights between the hidden nodes and the input nodes
7. Continue the cycle till the convergence conditions are met.
8. Label the activation rate of exit nodes using the last link weight.

The accuracy rate is 94.40% which is lower compared to other implemented algorithms, however similar to KNN. The precision score is 93.62% and it's the lowest precision score among other implemented algorithms. Moreover, it has the highest recall score, which is 97.78%. F1 score is 95.65% and it is slightly higher than KNN.

4.3 Decision Tree

Data mining using the Decision Tree Algorithm is a common practice in the field of Machine Learning. DT uses a tree structure to arrange a number of laws. It is amongst the most useful non-parametric supervised learning methods. To evaluate a classifier's performance, test data is drawn at random from training data after the tree or principles have been established during the process of learning. After accuracy is validated, unmarked data is classified using the tree or principles learned during the learning phase A Decision Tree, like a tree, has a root node, a left subtree, as well as a right subtree. A class label is expressed by a tree's leaf nodes. The conditions on the attributes are represented by the arcs from one node to the next [3]. In the case of Breast Cancer diagnosis, a Decision Tree is an effective tool for classification and prediction. ID3, C4.5, C5, J48, CART, CHAID, SLIQ, SPRINT, ScalParc, and other Decision Tree methods are available to classify the data [15]. CART is an acronym developed by Leo Breiman to define Decision Tree algorithms that may be used to address classification or regression predictive modeling problems. The technique Ross Quinlan used to generate Decision Trees, C4.5 (J48), was previously described. Quinlan's previous ID3 method was expanded upon in C4.5. As a statistical classifier, C4.5 creates Decision Trees that may be used for categorization. In 2008, Springer LNCS published the famous Top 10 Algorithms in Data Mining, which put it at the top of the list.
A 10-fold cross validation is done on the test and training data in this publica-

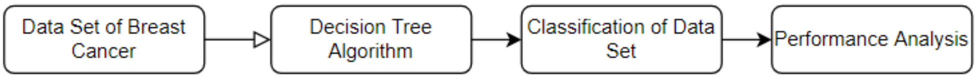


Figure 4.3: Performance Analysis using DT

tion [32]. Using WEKA (Java Toolkit for numerous data mining methods), the J48 method is applied to the dataset, and the data is categorized as "benign" or "malignant" based on the Decision Tree's final result after preprocessing. Figure 4.4 depicts the research flow utilized to construct the model.

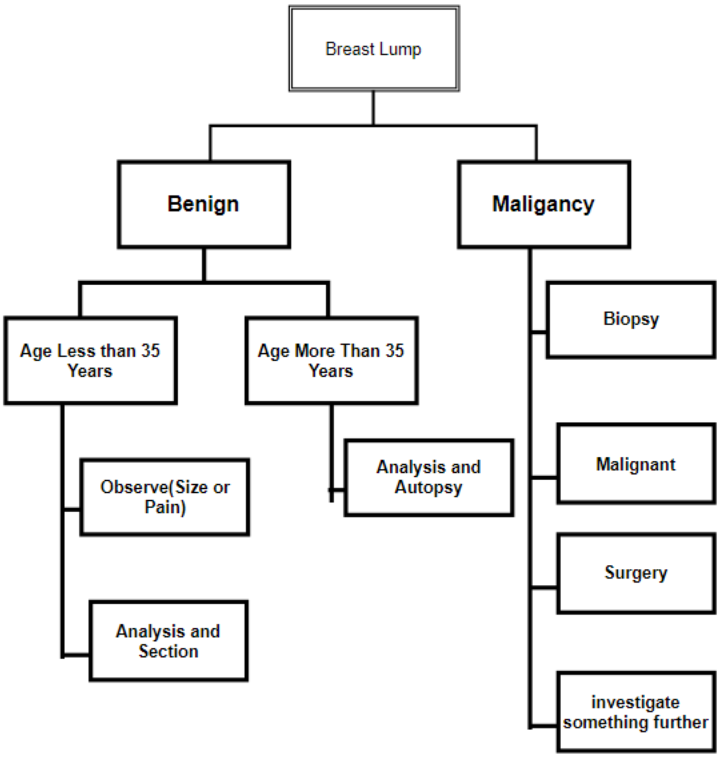


Figure 4.4: Breast Lump Detection Flow Diagram

The treatment choices available to a patient with a breast cancer diagnosis depend on the stage of the illness and the doctor administering the therapy. For the optimal treatment, a variety of aspects are considered, including the type of patient, their age, and their overall health. Figure 4.5 is a flow diagram illustrating a few of the alternatives available to a patient following a breast cancer diagnosis.
The process for the Decision-Tree algorithm is as follows:

1. Take the best characteristic of the given traits at the root of the tree
2. The training dataset should then be divided into subcategories.
3. To produce these splitting subsets, it needs to construct each subset with data of the very same valuation for an input variable .
4. Repeat steps 1–3 upon every subset till the leaf component of every branch is discovered.

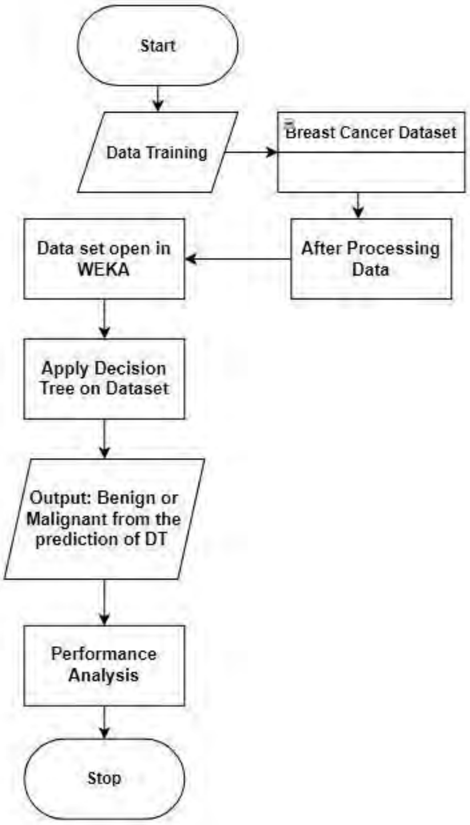


Figure 4.5: Flow Diagram for Patient Diagnosis

The accuracy rate is 95.80% which is the highest accuracy score among other applying algorithms. Precision score is 97.72% and it is quite close to the Linear SVC precision score but higher than KNN and ANN. Recall score is 95.56% similar to KNNs recall score. It has the highest F1 score, which is 96.62%.

4.4 K-Nearest Neighbors

KNN is a supervised machine learning technique for dealing with similarity. KNN is an acronym that stands for K-Nearest Neighbors. It is certainly a classification method that guesses a target variable's class based on a predefined amount of nearest neighbors. It will compute the distance between the instance to be classified and each instance in the training dataset, and thereafter categorize the instance through using the largest proportion of classes of the k-nearest instances. A supervised classification technique, such as k-Nearest Neighbor, needs training data, but an unsupervised algorithm does not need that [16]. The equation of KNN:
The k-nearest neighbor classifier assigns a weight of 1/k to the k nearest neighbors and a weight of 0 to all others. This is applicable to weighted closest neighbor classifiers. That is, where the ith nearest neighbor is assigned a weight wni, with i=1nwni=1. A similar finding holds for the strong consistency of weighted closest neighbor classifiers.
Let Cnwnn symbolize the weighted closest classifier with weights wnii=1n. Subject to regularity conditions on the class distributions the excess risk has the following

asymptotic expansion RR(Cnwnn)- RR(Cbayes)=(B1sn2+B2tn2)1+o(1), For constraints B1 and B2 where sn2=i=1nwni2 and tn=n-2/di=1nwnii1+2/d-(i-1)1+2/d The optimal weighting scheme wni*i=1n, that balances the two terms in the display above, is given as follows: set k*=Bn4d+d wni*= 1k*[1+d2-d2k2/di1+2/d-(i-1)1+2/d] for i= 1,2, k* and wni*=0 for i=k*1+1,....n
With optimal weights, the main term in the asymptotic increase of the excess risk is O(n-4d+4).

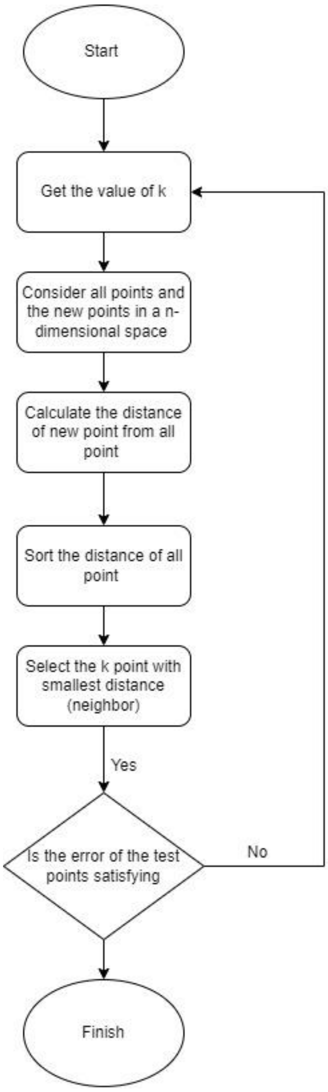

Start
Get the value of k
Consider all points and the new points in a n-dimensional space
Calculate the distance of new point from all point
Sort the distance of all point
Select the k point with smallest distance (neighbor)
Yes
Is the error of the test points satisfying
No
Finish

Figure 4.6: Flowchart of KNN model

The accuracy rate is 94.40% and it is similar to ANNs accuracy rate. Precision score is 95.56% which is higher than ANN but lesser than Decision Tree and linear SVCs precision score. Its recall score is 95.56% and it's higher than linear SVC. Also, its F1 score is also similar to its recall score.

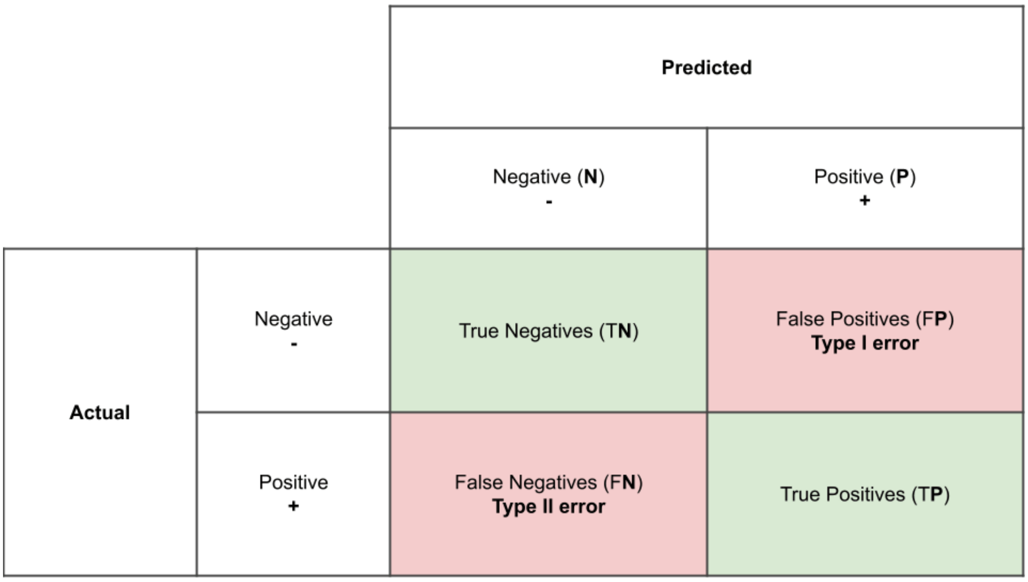
4.5 Reasons behind choosing KNN, DT, Linear SVC and ANN

KNN, Decision Tree (DT), ANN and Linear SVC are 4 popular techniques. Firstly, KNN predicts that objects that are identical are adjacent nearby. To put it another way, elements that are identical are closely together. Actually, it focuses on pattern recognition. In simple words, KNN incorporates the notion of resemblance (also called as distance, proximity, and so on) with some fundamental mathematics, such as determining the distance between graph nodes. Our second algorithm is Decision Tree (DT). When employing a Decision Tree approach, there is no need to standardize or normalize the data that has been collected. We are capable of dealing with both continuous and categorical variables with the help of DT. The pre-processing phases of a Decision Tree making model require less code and analytics, with the exception of usual data pre-processing stages. Then ANN (Artificial Neural Networks) is chosen for its ability to work with incomplete knowledge. Despite the insufficient information, the data can result in terms after ANN training. The relevancy of the missing data determines the level of performance lost in this case. Another advantage of ANN is having a distributed memory. In order for the network to learn, it must first identify the instances and then train it according to the desired output by providing it these examples. The network's outcomes are highly correlated with the examples used, and if the event cannot be shown to the network in all of its dimensions, the network may provide a misleading reading. Lastly, Linear SVC's ultimate aim is to fit the data we provided as well as provide a "best suited" hyperplane which classifies or characterizes given data. Following that, one may input some attributes to the classifier to evaluate whatever the "expected" class is after obtaining the hyperplane. Moreover, it optimizes quadratic hinge loss rather than just hinge loss, but it also marginalizes the extent of the bias. One interesting fact is that the wider the number of data, the faster Linear SVC tends to intertwine.

Chapter 5

Result & Discussion

The data was subjected to feature selection and extraction techniques in order to reduce the scale of variables, resulting in minimal duplicates of the original dataset. The datasets were trained using KNN, Linear SVC, DT, and ANN. The classification of data is produced after all of these have been implemented using Python. We determined the accuracy, F1 score, precision score, and recall score of the four different approaches.



		Predicted	
		Negative (**N**) -	Positive (**P**) +
Actual	Negative -	True Negatives (T**N**)	False Positives (F**P**) **Type I error**
	Positive +	False Negatives (F**N**) **Type II error**	True Positives (T**P**)

Figure 5.1: Representation of the confusion matrix

Precision: Precision is the percentage of relevant examples among the recovered instances in pattern recognition and machine learning classification.Positive predictive value is another name for it.For instance, the precision of our KNN model is 95.56% ,that means when it predicts that the patient has breast cancer it is correct around 95.56% times. Precision in mathematics is defined as:

$$Precision = \frac{TruePositive(TP)}{TruePositive(TP) + FalsePositive(FN)} \tag{5.1}$$

Models	**Precision**
KNN	95.56%
Linear SVC	98.83%
Decision Tree	97.72%
ANN	97.80%

Table 5.1: Precision of Four Models

The precision of our models are shown in below table 5.1 respectively:
Recall: In pattern recognition, classification, and information retrieval, recall is just as important as precision. It is a performance indicator that evaluates the percentage of successfully recognized positive cases.For instance, the recall of our Linear SVC model is 94.44% ,that means it is correct around 94.44% times while predicting the true positive value. It can be expressed mathematically as:

$$Recall = \frac{TP}{FN + TP} \tag{5.2}$$

The recall of our models are shown in table 5.2 respectively:

Models	**Recall**
KNN	95.56%
Linear SVC	94.44%
Decision Tree	95.56%
ANN	98.89%

Table 5.2: Recall Score of Four Models

Accuracy: In machine learning, a variety of measures are used to analyze a model's predicted accuracy. The accuracy metric to use is determined by the machine learning problem. These indicators must be examined in order to determine whether our model is operating properly.For instance, the precision of our ANN model is 97.9% ,it 97.9% correct in breast cancer prediction. It is determined by,

$$Accuracy = \frac{TN + TP}{TP + FN + TN + FP} \tag{5.3}$$

The accuracy of our models are shown in table 5.3 respectively:
F1: In machine learning, the F1-score is one of the most important evaluation criteria. In the establishment of a single statistic, the harmonic mean of a classifier's precision and recall are used by the F1-score.

$$F1score = \frac{2TP}{2TP + FN + FP} \tag{5.4}$$

Models	Accuracy
KNN	94.41%
Linear SVC	95.80%
Decision Tree	95.80%
ANN	97.90%

Table 5.3: Accuracy Score of Four Models

Models	F1 Score
KNN	95.56%
Linear SVC	96.59%
Decision Tree	96.62%
ANN	98.34%

Table 5.4: F1 Score of Four Models

The F1 score of our models are shown in below table 5.4 respectively:
In order to provide a better view of how effectively our proposed classification model works, the confusion matrix was also shown. The picture summarizes and displays the number of correct and incorrect predictions made by our algorithms while detecting authentic and false data.
In the confusion matrix of KNN algorithm we find out respectively that the number of individuals have symptoms and they are cancer positive is 49, the number of individuals do not have symptoms but they are cancer positive is 4, the number of individuals have symptoms but they are not cancer positive is 4 and the number of individuals neither have symptoms nor they are cancer positive is 86. Which are shown in the figure 5.2.

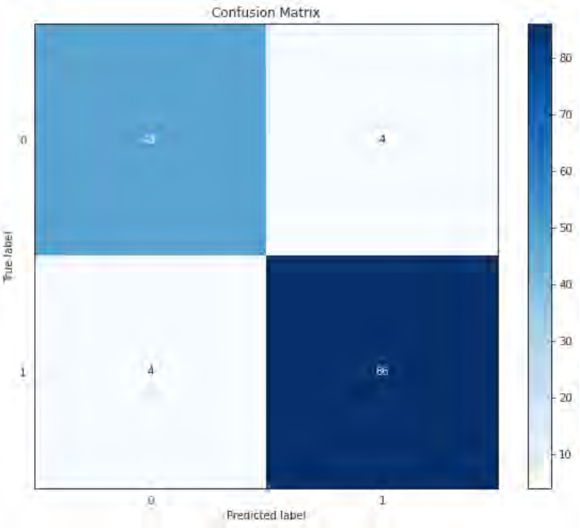


Figure 5.2: Confusion Matrix of KNN

In the pie chart, we see their percentage figure 5.3.

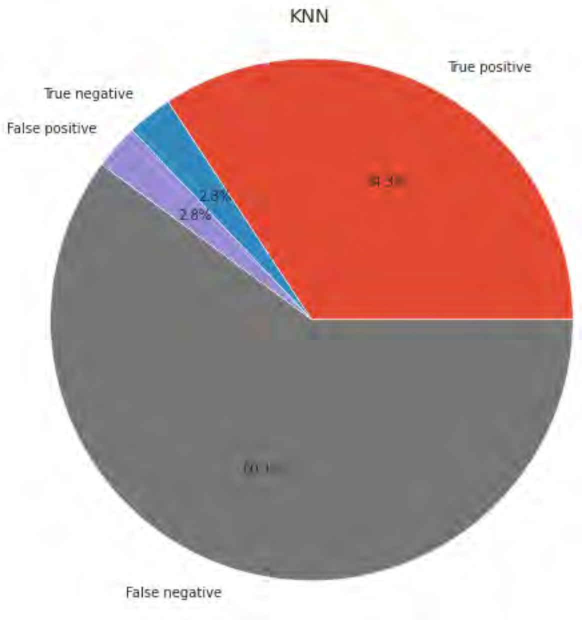


Figure 5.3: Pie chart of KNN

In the confusion matrix of Linear SVC algorithm we find out respectively that the number of individuals have symptoms and they are cancer positive is 52, the number of individuals do not have symptoms but they are cancer positive is 1, the number of individuals have symptoms but they are not cancer positive is 5 and the number of individuals neither have symptoms nor they are cancer positive is 85. As depicted by Figure 5.4

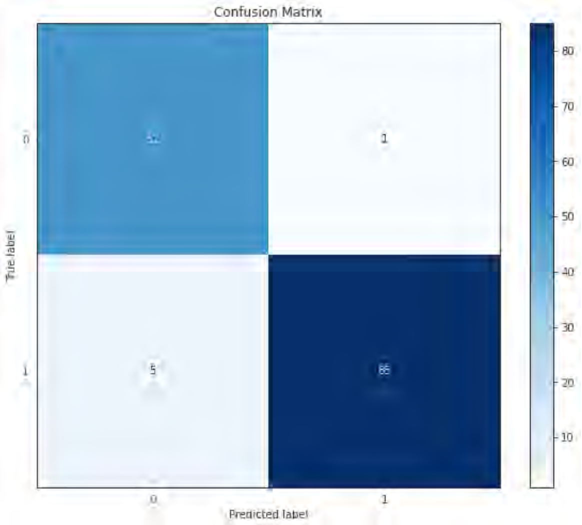


Figure 5.4: Confusion Matrix of SVC

In the pie chart, we see their percentage, figure 5.5.

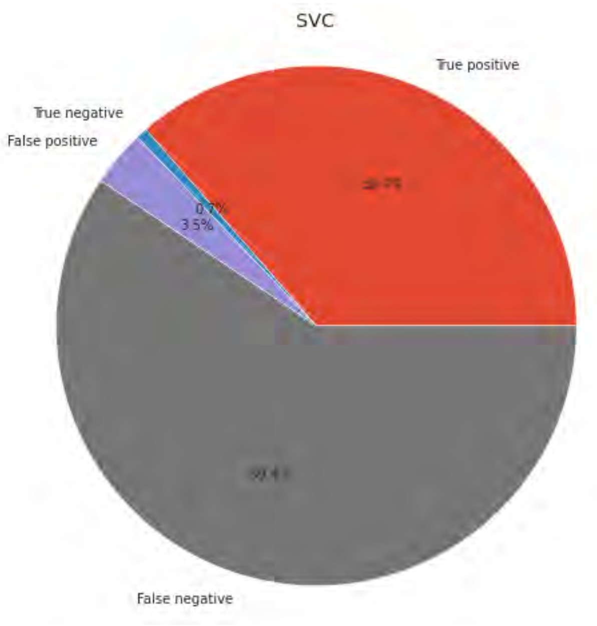


Figure 5.5: Pie chart of SVC

In the confusion matrix of Decision Tree algorithm we find out respectively that the number of individuals have symptoms, and they are cancer positive is 51, the number of individuals do not have symptoms, but they are cancer positive is 2, the number of individuals have symptoms but they are not cancer positive is 4 and the number of individuals neither have symptoms nor they are cancer positive is 86. shown in the figure 5.6.

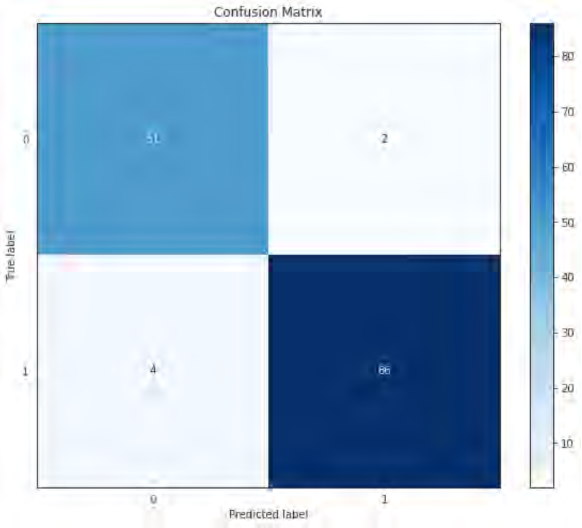


Figure 5.6: Confusion Matrix of DT

In the pie chart, we see their percentage, figure 5.7.

In the confusion matrix of ANN algorithm we find out respectively that the number of individuals have symptoms, and they are cancer positive is 51, the number of individuals do not have symptoms, but they are cancer positive is 2, the number of individuals have symptoms, but they are not cancer positive is 1 and the number of individuals neither have symptoms nor they are cancer positive is 89. Which are

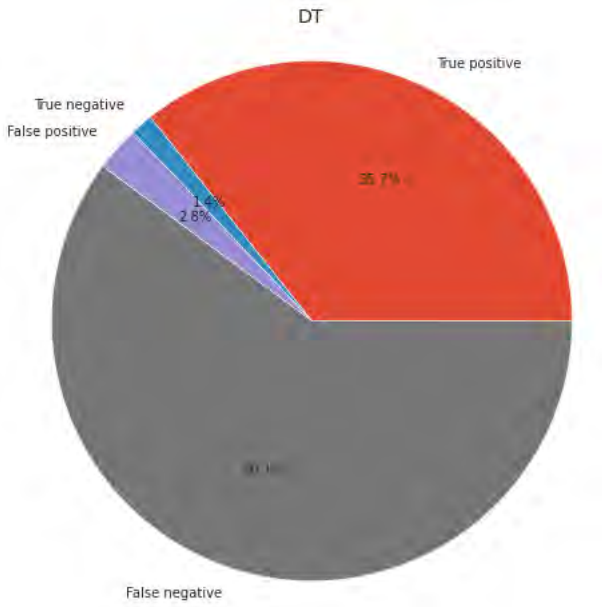


Figure 5.7: Pie chart of DT

shown in the figure 5.8.

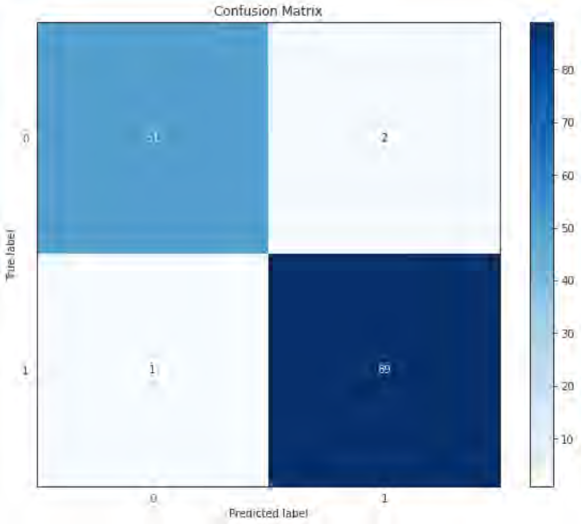


Figure 5.8: Confusion Matrix of ANN

In the pie chart we see their percentage, figure 5.9.
In the figure, 5.10 and 5.11 we are showing the model accuracy and model loss of ANN.

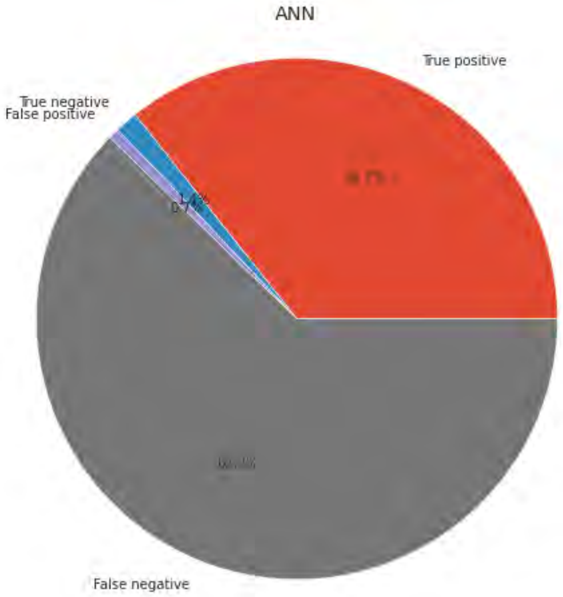


Figure 5.9: Pie chart of ANN

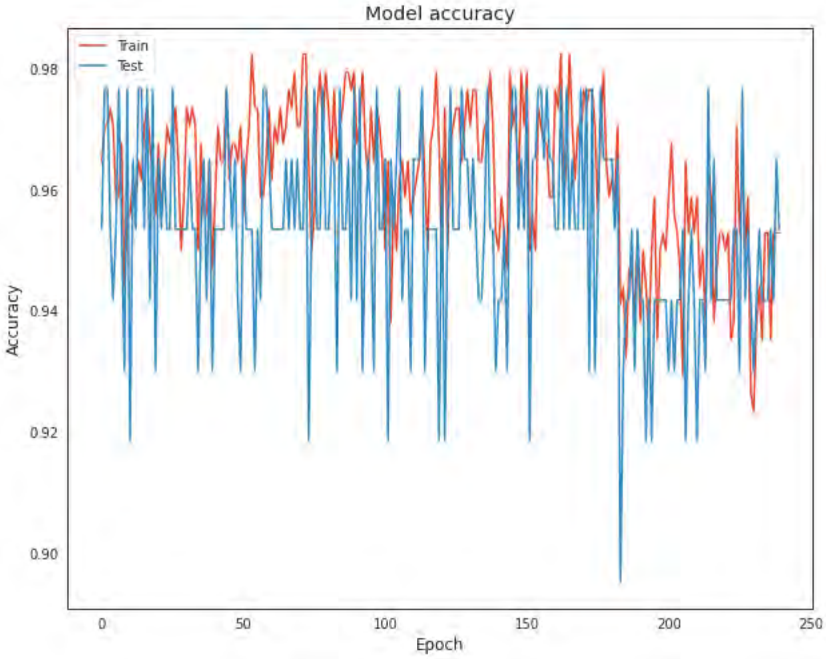


Figure 5.10: Model Accuracy of ANN

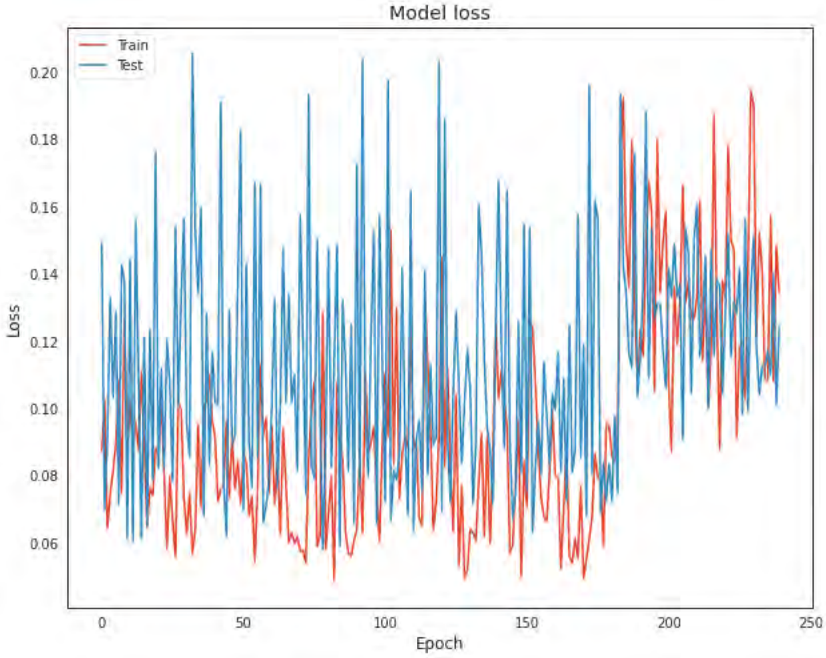


Figure 5.11: Model loss of ANN

For the better understanding of our model we have found out the ROC curves and the value of area under ROC curve which is called AUC. The AUC value of KNN, DT and ANN are respectively 0.970440251572327, 0.95890853249475 and 0.994. This graph shows the performance of our models at all classification terminations. In this graph two parameters are being plotted, they are "True Positive Rate" and "False Positive Rate".True positive rate is being plotted in "Y" axis and the False positive rate in the "X" axis.
True positive rate is denoted by,

$$TPR = \frac{TP}{TP + FP} \tag{5.5}$$

False positive rate is denoted by,

$$FPR = \frac{FP}{FP + TN} \tag{5.6}$$

The ROC curve of our models are shown in figure 5.12 to 5.14 respectively,

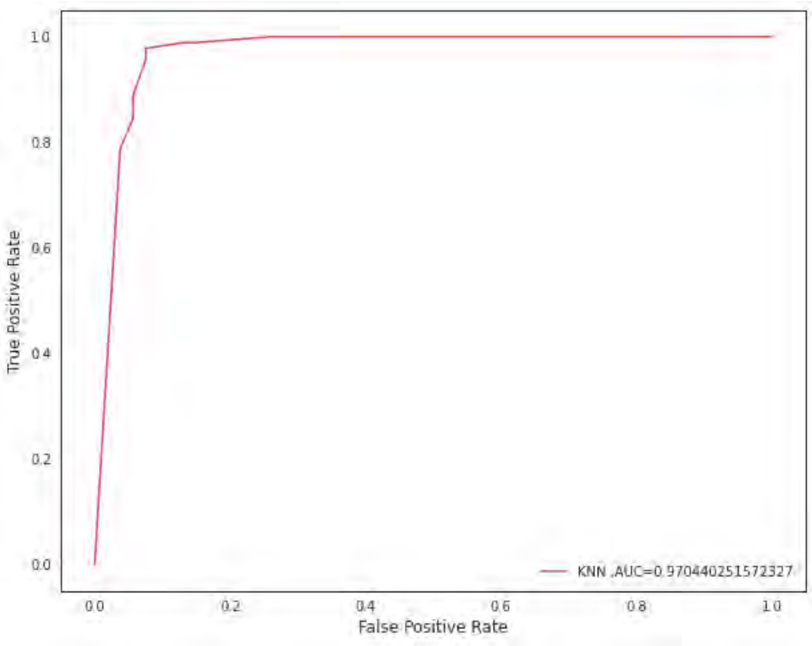


Figure 5.12: ROC curve of KNN

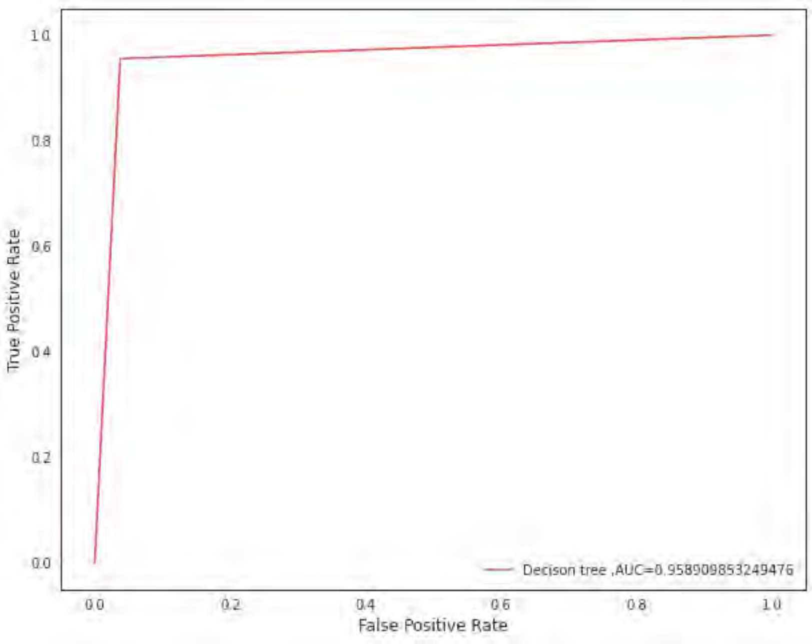


Figure 5.13: ROC curve of DT

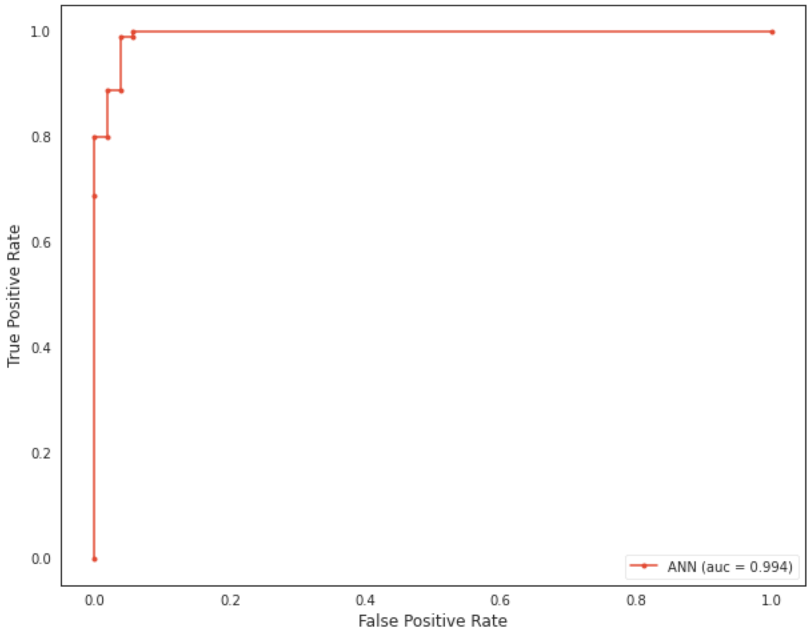


Figure 5.14: ROC curve of ANN

The results of the different classifier algorithms are compared. Table (5.5) shows the performance metrics of our best result.

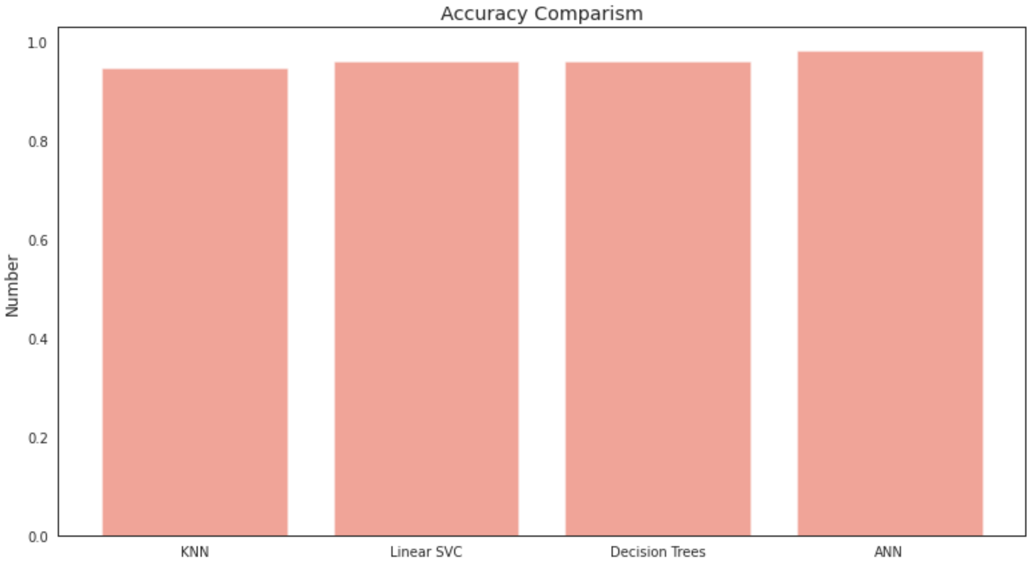


Figure 5.15: Accuracy Comparison

We observe the accuracy compassion of our models in the bar diagram in figure 5.15.

According to the results provided in the table 5.5, the Decision Tree and linear SVC achieved the accuracy, with a score of 95.8%. In summary, Decision Tree outperforms Linear SVC in terms of accuracy and F1 score, while Linear SVC outperforms Decision Tree in terms of precision. We can observe that KNN has the lowest accuracy which is 94.41% respectively precision and F1 of KNN is also the lowest. While ANN has the best accuracy among all the models which is 97.9%.In terms

Models	**F1 (%)**	**Precision (%)**	**Recall (%)**	**Accuracy (%)**
KNN	95.56%	95.56%	95.56%	94.41%
Linear SVC	96.59%	98.83%	94.44%	95.80%
DT	96.62%	97.72%	95.56%	95.80%
ANN	98.34%	97.80%	98.89%	97.90%

Table 5.5: Performance Matrix

of precision Linear SVC outperforms rest of the three models. On the other hand ANN outperforms all the models in terms of F1 and accuracy. So overall ANN is the best fitted model according to our research.

Chapter 6

Conclusion

We attempted to present the fundamentals of machine learning while highlighting their application in breast cancer prediction and prognosis in this review. To be specific, we found some aptitude after various kinds of ML processes are being applied, various kinds of educated data are being mobilized, the cases of breast cancers being studied, types of endpoint prediction being made and overall performance of all of these methods for prediction of breast cancer. Many recent research on breast cancer employed supervised machine learning approaches and classification algorithms to construct prediction models that could accurately forecast the course of the disease. This is what we found after reviewing a number of these studies. Although most studies are typically adequately designed and verified, more attention to experimental design and implementation appears to be required. Improvements in experimental design and biological validation will undoubtedly improve the overall quality, generality, and repeatability of many machine-based classifiers. We tried to compare the outcomes of 4 different types of methods which has been used worldwide. Additionally, another intention was to go through with a wider number of parameters and analyze the outcomes. As breast cancer is a vast area of study, we tried to focus on some very specific points. To sum up, we anticipate that if study quality continues to improve, the use of machine learning classifiers will become much more prevalent in many clinical and hospital settings. Here, it was aimed to present the principles of ML while emphasizing their use in breast cancer prediction and prognosis. We discovered significant patterns connected to the various types of machine learning methods employed, the diversity of training data integrated, the breast cancer cases analyzed, the sorts of Endpoint prediction done, and the overall performance of all these breast cancer prediction approaches. We discovered that the majority of studies aimed at developing predictive models using supervised machine learning and classification algorithms to predict genuine illness outcomes used supervised machine learning methods and classification algorithms after sifting through several breast cancer studies published in recent years. Although most investigations are sufficiently structured and controlled, additional attention to experimental design and implementation appears to be required. Improvements in experimental design and biological validation will undoubtedly increase the overall quality, generality, and repeatability of many machine-based classifiers. We attempted to compare the outcomes of four distinct types of approaches employed all around the world. Another goal is to take into account a wider number of parameters and examine the outcomes. We attempted to focus on very specific areas because breast cancer research is such a broad field. In conclusion, we believe that as study quality improves, the use of machine learning classifiers will become considerably more widespread in various medical settings, including institutions.

6.1 Future Work

After all, we believe that our work is one small step towards making it great one day.Also, we added four different algorithms in our research, which makes it different from other research as other research added one or two machine learning algorithms. Our future goal is to apply the four different ML algorithms Decision Tree, Artificial Neural Network (ANN), KNN and Linear SVC to larger datasets and compare which algorithm provides the best accuracy for breast cancer prediction.We have also plan to work with deep learning and Neural network models where we will use the image data set instead of numerical data set. Today, cancer in breast is becoming a major intimidation to fermale population all over the world. If we can predict cancer in breast through machine learning, it will be very helpful for medicine to start treatment as early as possible.

Bibliography

[1] A. Newell, "A step toward the understanding of information processes: Perceptrons . an introduction to computational geometry. marvin minsky and seymour papert. m.i.t. press, cambridge, mass., 1969. vi + 258 pp., illus. cloth, 12;*paper*,4.95.," *Science*, vol. 165, no. 3895, pp. 780–782, 1969. DOI: 10.1126/science.165.3895.780.

[2] D. E. Rumelhart and J. L. McClelland, "Parallel distributed processing," 1986. DOI: 10.7551/mitpress/5236.001.0001.

[3] R. L. De Mántaras, *Machine Learning*, vol. 6, no. 1, pp. 81–92, 1991. DOI: 10.1023/a:1022694001379.

[4] G. E. Hinton, "How neural networks learn from experience," *Scientific American*, vol. 267, no. 3, pp. 144–151, 1992. DOI: 10.1038/scientificamerican0992-144.

[5] K. P. Bennett and J. A. Blue, "A support vector machine approach to decision trees," *1998 IEEE International Joint Conference on Neural Networks Proceedings. IEEE World Congress on Computational Intelligence (Cat. No.98CH36227)*, vol. 3, 2396–2401 vol.3, 1998.

[6] D. Parkin, "Epidemiology of cancer: Global patterns and trends," *Toxicology Letters*, vol. 102-103, pp. 227–234, 1998, ISSN: 0378-4274. DOI: https://doi.org/10.1016/S0378-4274(98)00311-7. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0378427498003117.

[7] T. Joachims, "Making large-scale support vector machine learning practical," in *Advances in Kernel Methods: Support Vector Learning*. Cambridge, MA, USA: MIT Press, 1999, pp. 169–184, ISBN: 0262194163.

[8] I. Guyon, J. Weston, S. Barnhill, and V. Vapnik, *Machine Learning*, vol. 46, no. 1/3, pp. 389–422, 2002. DOI: 10.1023/a:1012487302797.

[9] J. F. MCCARTHY, K. A. MARX, P. E. HOFFMAN, *et al.*, "Applications of machine learning and high-dimensional visualization in cancer detection, diagnosis, and management," *Annals of the New York Academy of Sciences*, vol. 1020, no. 1, pp. 239–262, 2004. DOI: 10.1196/annals.1310.020.

[10] A. Jemal, R. Siegel, E. Ward, *et al.*, "Cancer statistics, 2008," *CA: A Cancer Journal for Clinicians*, vol. 58, no. 2, pp. 71–96, 2008. DOI: 10.3322/ca.2007.0010.

[11] S. Thirumuruganathan, *A detailed introduction to k-nearest neighbor (knn) algorithm*, May 2010. [Online]. Available: https://saravananthirumuruganathan.wordpress.com/2010/05/17/a-detailed-introduction-to-k-nearest-neighbor-knn-algorithm/.

[12] S. Aruna, S. Rajagopalan, and L. Nandakishore, “An algorithm proposed for semi-supervised learning in cancer detection,” in *International Conference on Sustainable Energy and Intelligent Systems (SEISCON 2011)*, 2011, pp. 860–864. DOI: 10.1049/cp.2011.0487.

[13] K. Bache and M. Lichman, *UCI machine learning repository*, 2013. [Online]. Available: http://archive.ics.uci.edu/ml.

[14] A. Qasem, S. N. H. S. Abdullah, S. Sahran, *et al.*, “Breast cancer mass localization based on machine learning,” in *2014 IEEE 10th International Colloquium on Signal Processing and its Applications*, 2014, pp. 31–36. DOI: 10.1109/CSPA.2014.6805715.

[15] J. R. Quinlan, *C4. 5: programs for machine learning*. Elsevier, 2014.

[16] Z. K. Senturk and R. Kara, “Breast cancer diagnosis via data mining: Performance analysis of seven different algorithms,” *Computer Science Engineering: An International Journal*, vol. 4, no. 1, pp. 35–46, 2014. DOI: 10.5121/cseij.2014.4104.

[17] R. Sumbaly, N. Vishnusri, and S. Jeyalatha, “Diagnosis of breast cancer using decision tree data mining technique,” *International Journal of Computer Applications*, vol. 98, no. 10, pp. 16–24, 2014. DOI: 10.5120/17219-7456.

[18] J. A. Bhat, V. George, and B. Malik, “Cloud computing with machine learning could help us in the early diagnosis of breast cancer,” in *2015 Second International Conference on Advances in Computing and Communication Engineering*, 2015, pp. 644–648. DOI: 10.1109/ICACCE.2015.62.

[19] H. Asri, H. Mousannif, H. A. Moatassime, and T. Noel, “Using machine learning algorithms for breast cancer risk prediction and diagnosis,” *Procedia Computer Science*, vol. 83, pp. 1064–1069, 2016, The 7th International Conference on Ambient Systems, Networks and Technologies (ANT 2016) / The 6th International Conference on Sustainable Energy Information Technology (SEIT-2016) / Affiliated Workshops, ISSN: 1877-0509. DOI: https://doi.org/10.1016/j.procs.2016.04.224. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1877050916302575.

[20] M. R. Al-Hadidi, A. Alarabeyyat, and M. Alhanahnah, “Breast cancer detection using k-nearest neighbor machine learning algorithm,” in *2016 9th International Conference on Developments in eSystems Engineering (DeSE)*, 2016, pp. 35–39. DOI: 10.1109/DeSE.2016.8.

[21] R. Shimizu, S. Yanagawa, Y. Monde, *et al.*, “Deep learning application trial to lung cancer diagnosis for medical sensor systems,” in *2016 International SoC Design Conference (ISOCC)*, 2016, pp. 191–192. DOI: 10.1109/ISOCC.2016.7799852.

[22] N. K. Al-Salihy and T. Ibrikci, “Classifying breast cancer by using decision tree algorithms,” in *Proceedings of the 6th International Conference on Software and Computer Applications*, ser. ICSCA ’17, Bangkok, Thailand: Association for Computing Machinery, 2017, pp. 144–148, ISBN: 9781450348577. DOI: 10.1145/3056662.3056716. [Online]. Available: https://doi.org/10.1145/3056662.3056716.

[23] Y. Tan, *Data mining and big data*, 2017. [Online]. Available: https://link.springer.com/book/10.1007/978-3-319-61845-6.

[24] P. D. J. 16, *Understanding a cancer prognosis*, Dec. 2018. [Online]. Available: https://www.cancertherapyadvisor.com/home/tools/fact-sheets/understanding-a-cancer-prognosis/.

[25] N. Khuriwal and N. Mishra, "Breast cancer diagnosis using adaptive voting ensemble machine learning algorithm," in *2018 IEEMA Engineer Infinite Conference (eTechNxT)*, 2018, pp. 1–5. DOI: 10.1109/ETECHNXT.2018.8385355.

[26] R. Mothkur and K. Poornima, "Machine learning will transfigure medical sector: A survey," in *2018 International Conference on Current Trends towards Converging Technologies (ICCTCT)*, 2018, pp. 1–8. DOI: 10.1109/ICCTCT.2018.8551134.

[27] R. B. Parikh, C. Manz, C. Chivers, *et al.*, "Machine learning approaches to predict 6-month mortality among patients with cancer," *JAMA Network Open*, vol. 2, no. 10, 2019. DOI: 10.1001/jamanetworkopen.2019.15997.

[28] I. Amelio, R. Bertolo, P. Bove, *et al.*, *Cancer predictive studies - biology direct*, Oct. 2020. [Online]. Available: https://biologydirect.biomedcentral.com/articles/10.1186/s13062-020-00274-3.

[29] T. Saba, "Recent advancement in cancer detection using machine learning: Systematic survey of decades, comparisons and challenges," *Journal of Infection and Public Health*, vol. 13, no. 9, pp. 1274–1289, 2020. DOI: 10.1016/j.jiph.2020.06.033.

[30] H. Sung, J. Ferlay, R. L. Siegel, *et al.*, "Global cancer statistics 2020: Globocan estimates of incidence and mortality worldwide for 36 cancers in 185 countries," *CA: A Cancer Journal for Clinicians*, vol. 71, no. 3, pp. 209–249, 2021. DOI: 10.3322/caac.21660.

[31] E. Aboul and A. E. Hassanien, "Classification and feature selection of breast cancer data based on decision tree algorithm," May 2022.

[32] O. Tarawneh, M. Otair, M. Altarawneh, H. Abu Addous, M. Tarawneh, and M. A. Almomani, "Breast cancer classification using decision tree algorithms," *International Journal of Advanced Computer Science and Applications*, vol. 13, Jan. 2022. DOI: 10.14569/IJACSA.2022.0130478.