

**A Study of Breast Cancer Prediction Using Machine Learning Approaches**

**BY**

**MD. ABU RAIHAN**

**ID: 161-15-7189**

**YEASMIN AKTER**

**ID: 161-15-7215**

**DIPTO SARKER**

**ID: 161-15-7110**

This Report Presented in Partial Fulfillment of the Requirements for the  
Degree of Bachelor of Science in Computer Science and Engineering

Supervised By

**MR. ANIRUDDHA RAKSHIT**

Senior Lecturer

Department of CSE

Daffodil International University



**DAFFODIL INTERNATIONAL UNIVERSITY**

**DHAKA, BANGLADESH**

**DECEMBER 2019**

## APPROVAL

This Project/internship titled “**A Study of Breast Cancer Prediction Using Machine Learning Approaches**”, submitted by Md. Abu Raihan , ID No: 161-15-7189, Dipto Sarker ID No: 161-15-7110, Yeasmin Akter ID No: 161-15-7215 to the Department of Computer Science and Engineering, Daffodil International University has been accepted as satisfactory for the partial fulfillment of the requirements for the degree of B.Sc. in Computer Science and Engineering and approved as to its style and contents. The presentation has been held on December 6, 2019.

## BOARD OF EXAMINERS



**Dr. Syed Akhter Hossain**  
**Professor and Head**

Department of Computer Science and Engineering  
Faculty of Science & Information Technology  
Daffodil International University

**Chairman**



**Md. Sadekur Rahman**  
**Assistant Professor**

Department of Computer Science and Engineering  
Faculty of Science & Information Technology  
Daffodil International University

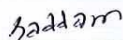
**Internal Examiner**



**Abdus Sattar**  
**Assistant Professor**

Department of Computer Science and Engineering  
Faculty of Science & Information Technology  
Daffodil International University

**Internal Examiner**



**Dr. Md. Saddam Hossain**  
**Assistant Professor**

Department of Computer Science and Engineering  
United International University

**External Examiner**

## DECLARATION

We hereby declare that, this project has been done by us under the supervision of, **Mr. AniruddhaRakshit, Department ofCSE**, Daffodil International University. We also declare that neither this project nor any part of this project has been submitted elsewhere for award of any degree or diploma.

**Supervised By:**

*Aniruddha Rakshit*

---

**Mr. Aniruddha Rakshit**

Senior Lecturer

Department of CSE

Daffodil International University

**Submitted By:**

*Raihan*

---

**Md. Abu Raihan**

ID: 161-15-7189

Department of CSE

Daffodil International University

*Yeasmin*

---

**YeasminAkter**

ID: 161-15-7215

Department of CSE

Daffodil International University

*Dipto*

---

**DiptoSarker**

ID: 161-15-7110

Department of CSE

Daffodil International University

## ACKNOWLEDGEMENT

First, we express our heartiest thanks and gratefulness to almighty Allah for His divine blessing makes us possible to complete the final year project/internship successfully.

We sincerely and heartily grateful to our advisor **Mr. Aniruddha Rakshit, Senior Lecturer**, Department of Computer Science & Engineering, Daffodil International University, Dhaka. Deep Knowledge & keen interest of our supervisor in the field of “*Machine Learning*” to carry out this thesis. Her endless patience, scholarly guidance, continual encouragement, constant and energetic supervision, constructive criticism, valuable advice, reading many inferior drafts and correcting them at all stage have made it possible to complete this project.

We would also like to wish our deepest heartiest gratitude to **Prof. Dr. Syed Akhter Hossain, Head**, Department of CSE for his lot of deepest help to fulfill our final year project and also thanks to other faculty members and the employees of CSE department of Daffodil International University.

Finally, I must acknowledgement with due respect the constant support and patients of my parents.

## **ABSTRACT**

The main aspect of this study is to evaluate the different Machine learning classifiers performance for prediction of breast cancer disease. In this work, we have used six supervised classification techniques for the classification of breast cancer disease. For example: SVM, NB, KNN, RF, DT and LR were used for early prediction of breast cancer. Therefore, we evaluated the breast cancer dataset through sensitivity, specificity, f1 measure and total accuracy. The prediction performance of breast cancer analysis shows that SVM obtained the uppermost performance with utmost classification accuracy of 97.07%. Whereas, NB and RF has achieved the second highest accuracy by prediction. Our findings can be used to help reduce the occurrence of the breast cancer disease through developing a machine learning based predictive system for early prediction.

## TABLE OF CONTENTS

<b>CONTENTS</b>	<b>PAGE</b>
Board of examiners	i
Declaration	ii
Acknowledgements	iii
Abstract	iv
<b>CHAPTER</b>	
<b>CHAPTER 1: INTRODUCTION</b>	<b>1-2</b>
1.1 Introduction	1
1.2 Motivation	1
1.3 Objective	2
1.4 Expected Outcome	2
1.5 Page Layout	2
<b>CHAPTER 2: BACKGROUND</b>	<b>3-5</b>
2.1 Introduction	3
2.2 Related Works	3
2.3 Research Summary	5
2.4 Scope of the Problem	5
2.5 Challenges	5
<b>CHAPTER 3: RESEARCH METHODOLOGY</b>	<b>6-9</b>
3.1 Introduction	6
3.2 Implementation Requirements	6-9
<b>Chapter 4: EXPERIMENTAL RESULTS AND DISCUSSION</b>	<b>10-12</b>

4.1 Introduction	10
4.2 Experimental Results	10
4.3 Descriptive Analysis	10
4.4 Summary	10-12
<b>Chapter 5: Summary, Conclusion, Recommendation and Implication for Future Research</b>	13-14
5.1 Summary of the Study	13
5.2 Conclusion	13
5.3 Recommendations	13
5.4 Implication for Further Study	13-14
<b>APPENDIX</b>	15
<b>REFERENCES</b>	16-17

## **LIST OF FIGURES**

<b>FIGURES</b>	<b>PAGE NO</b>
Figure 3.1: The experimental setup	6
Figure 3.2: Parameters for data analysis	7
Figure3.3: No missing values on breast cancer datasets	7
Figure 3.4: Heat map for correlated columns for breast cancers	8
Figure4.1: The accuracy of six machine learning (Breast Cancer)	10
Figure4.2: Classification Performance Measurements (Breast cancer)	11
Figure4.3: Confusion matrix of classification techniques	11
Figure4.4: ROC curve for Breast Cancer datasets	12



# CHAPTER 1

## INTRODUCTION

### 1.1 Introduction

For perdition of different disease to make the best feasible clinical care decisions it is very essential to take right decision form medical doctors. If the doctor takes their improper decisions, it could be likely to cause interruptions in medical action or drawn to loss of life. We know the medical services is a big commercial viewpoint in every time. The business stream always running in this fields. Patients are always searching there is a good platform for better services. But there is no 100% affordable platform for every patient. Therefore, in this domain need of one excessive platform for problem solves in healthcare and medical fields. Here is my main idea for improving healthcare services is to place more highlighting on early detection of chronic disease and less on treatment and live for better life.

### 1.2 Motivation

Breast cancer are the prominent causes of women death and disability in the global perspective. A report shows that the 508000 women died in 2011 by chronic disease, specially breast cancer [1]. In 2015, around 17.7 million people were global death caused by CHD [2]. The World Health Organization (WHO) estimated that above 23.6 million persons will be dead by 2030, because of such chronic disease[3]. Very few peoples can get their treatment but most of scenario affected by chronic disease treatment is very expensive and complex [4]. Moreover, this reason to takes long time, mistaken or delayed decisions are possible to cause of death. However, the cost of breast cancer diagnosis and replacement is very extreme and it can be calls as extreme level of financial expenses. A study reported that the cancer disease cause the commercial benefits with cost over \$79 billion, and treating people with end stage renal disease cost around \$35 billion [5]. Breast cancer disease is chronic in nature and take long time forcured. This causes most of the patients cannot afford the cost of the cure for cancer disease. Furthermore, chronic disease prediction is most prominent matter for clinical practitioners and medical services center in order to take accurate decision of such disease. Therefore, machine learning based extensive platform can solve these

kidney disease problems through early detection and diagnosis. This work's main aspect is to improve early treatment and diagnosis of kidney disease for people of low-income and developing countries. Hence, our study can be a significant approach for the detecting kidney disease outbreak with machine learning algorithms.

### **1.3 Objective**

In the last 10 years, the growth rate of medical data is going to large amount from enormous arenas [6]. From the art of Machine learning (ML) algorithms have portrayed that purpose to resolve various health and scientific problem [7]. An establishment of several studies show that ML models already have obtained dramatically excessive accuracies in disease based medical problems. However, supervised based models are one of the utmost operative method for the academic and health products on clinical fields. [8]. This work's main aspect is to improve early treatment and diagnosis of chronic disease for people of low-income and developing countries. Hence, our study can be a significant approach for the detecting chronic disease outbreak with machine learning algorithms.

In this work we have a specific research question for breast cancer detection by prediction using computational techniques which is addressed below:

- (1) Does the different type of ML algorithms affect the performance between the various classifiers to prediction of breast cancer?

### **1.4 Expected Outcome**

In this study, the main goal to achieve that the ML predictive model can detect the early breast cancer symptoms through prediction by the experimental model.

### **1.5 Page Layout**

The rest of the work is ordered as following, chapter one describes the objectives of this thesis, motivation behind this thesis, research scope and thesis organization. Chapter 2 depicted the literature review and related works in these clinical areas. And the materials and methodology are described with the evaluation benchmark of different classification techniques in Section 3. Therefore, the performance results and discussion are demonstrated in section four. Finally, the conclusions and future viewpoint of research and recommendations are deliberated in section five.

## CHAPTER 2

### BACKGROUND

#### 2.1 Introduction

Our main aspect is to develop a system using machine learning for early forecast of cancer disease from patient's data. Through related studies were done on applying and using several ML classifiers to determine early detection and prediction of breast cancer using ML techniques.

#### 2.2 Related Work

However, the outcomes of the previous work on machine learning used in breast cancer prediction as follows: Jain et al. [12] presented a survey to attribute assortment and machine learning techniques for identification and forecast of chronic disease. This work focused on a comprehensive review of different feature selection methods and their advantage and limitation. The contribution of this study is to use adaptive with parallel classification techniques for chronic complaint prediction. Bartz-Kurycki et al. [12], introduced a new model to forecast "neonatal surgical site infections (SSI)" using diverse classification processes. Accuracy of area under the curve for each model was similar. The contribution of this study is to examine the hybrid model and other models with fewer and more clinically relevant variables. Carvalho et al. [13] presented a new hybrid method to sustain the early verdict of breast cancer. This study tries to find optimum accuracy to provide sustenance to verdict in circumstances whereas Bayesian Network does not provide a satisfactory outcome. The contribution of the study is to advance an automatic device to contribute a precise identification and prediction of breast cancer. Kumari [14], presented a new prediction system that can predict the occurrence of breast cancer at early stage by analyzing nominal set of attributes that has been selected from medical datasets. The KNN classifier obtain the best performance (99.28%) than others classifiers. The contribution of this study is to use the proposed system to predict the breast cancer at early stage with greatly reduces the cost of treatment and improves the quality of life. Tapak et al. [15], introduced a comparative study between Naïve Bayes, Random Forest, AdaBoost, Support Vector Machine, LSSVM, Adabag, Logistics Regression and LDA to predict breast cancer

survival and metastasis. LR and LDA were achieved the highest accuracy (86%). The SVM and LDA have superior sensitivity in comparison to other classifiers. The contribution of this study is to use SVM to predict existence of breast cancer. Asri et al. [16] presented a comparative study between SVM, DT (C4.5), NB, K-NN to predict early stage of breast cancer. The intelligent techniques are applied on WEKA data mining tool. Experiment results show that the SVM have the best performance accuracy, it is 97.13%. The contribution of this study to use SVM to predict the early stage of breast cancer. Chougrad et al. [16] developed a deep convolutional neural networks based computer aided treatment system. The CNN model achieved the best performance, it is 98.94%. And they tested the CNN model on independent database and they've got the accuracy 98.23% and 0.99 AUC. The contribution of this study to use the high performer classifiers within the proposed structure and that can be used to forecast the patients are "benign or malignant". Wang et al. [17], presented a new model to use breast cancer diagnosis based on patient's historical data from clinical data. The proposed WAUCE model reduces the variance by around 97.98% and increase accuracy by 33.34%. The contribution of this study can be further applied to safer, more reliable illness diagnosis process. Madhuri & Bharat et al. [18], present a comparative study to diagnosis the breast cancer patients through supervise machine learning techniques. They applied multiple machine learning algorithms including LR, RF, DT and Multi-layer perception. Multi-layer perception gives high performance compare to others algorithms. Layla & Hana [19], implement a feed forward back propagation network (FFBPN) to classify the "benign cancer or malign cancer". They showed that for an Artificial Neural network best design for classification is that three hidden layer and twenty-one neurons in every hidden level. Propose design gives highest accuracy 98%. Amrane et al. [20] presented a comparison between two machine learning classifier NB and KNN to provide an effectively diagnosis breast cancer patients. The comparison result was KNN gives high accuracy at 97.51% and NB has 96.19% accuracy. Al-hadidi et al. [21], Proposed a model to detect the breast cancer disease with a higher accurateness. Their model was divided into two part, first one was processing the images for extract the features and second one was used two supervise machine learning techniques to get the accuracy. Xiao et al. [22], introduce a new strategy for gene expression analysis to five different classification algorithms with deep learning methods. The contribution of this study is shown to be accurate and effective prediction results for cancer prediction.

### **2.3 Research Summary**

The above study done on different types of empirical works from various researchers. We have already seen from the previous study that the breast cancer and also with medical research is growing than we think. Some of the well-established work already prove this declare it well. Though, more efficient sources are not present, but prospects is that this area is becoming more imaginative each after passing everybody.

### **2.4 Scope of the problem**

To date, machine learning classification techniques have created a significant impact and obligation in the chronic disease research society for primary discovery of chronic disease. Moreover, ML algorithms are given more accurate results in chronic disease prediction as compared to others data classification techniques [8][9]. Many of studies already shows that the supervised based classification techniques have obtained excellent accuracies in the field of disease prediction [10][4][11] Motivated by this, the authors have used six prominent ML techniques for prediction and proper treatment of chronic patients. The main goal of this study is to inspect the performance measurement of various prominent supervised methods and gained more efficient outcome by reducing extremely cost of diagnosis and dialysis of chronic diseases. For this study, six supervised learning techniques were used including “KNN, Support Vector Machine, Decision Tree, Random Forest, Naïve Bayes and Logistics Regression”. Moreover, the classifiers performance of selected learning techniques is evaluated using the confusion matrix and different statistical methods. Henceforth, the outperform classification technique will be donated for the decision support system and diagnosis of chronic disease.

### **2.5 Challenges**

We have facing certain problems in this work. The collection real time health data from different device or sources are very complex and at the mean time analysis of this data will be challenged. Therefore, we have to first to build a predictive model which will be able to breast cancer from patient’s data. This task is really very challenges for us.

# CHAPTER 3

## MATERIALS AND METHODS

### 3.1 Introduction

In this chapter, we will introduce our experimental setup, data collection process, data preprocessing techniques with various validation process. Therefore, we will present the tools and techniques. Now we will discuss all of things in this chapter 3.

### 3.2 Implementation Requirements

The proposed experimental setup including ML systems has been presented in this section. The intelligent breast cancer detection technique contains of a four steps to take this decision. Phase 1 focuses on extracting and combined data from diverse health systems with devices. In phase 2 usages to store huge amount of medical data. Therefore, the phase 3 usages ML based classifiers to training data of cancer disease dataset. In addition, phase 4 exemplifies the outcome of the breast cancer detection system for the clients. In this study, we have focused only the machine learning phase (phase 3). For further study, we are currently developing in this full system architecture.

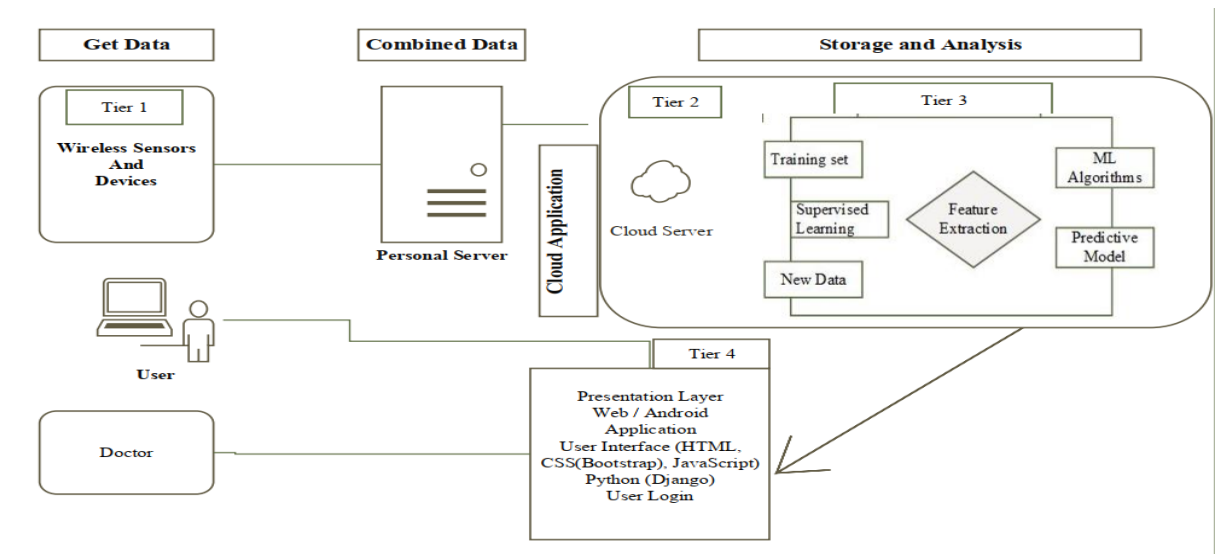


Figure 3.1. The experimental setup

### Data Collection

In this study, we use the Wisconsin Breast Cancer data (Original) by the “University of Wisconsin Hospitals Madison, Wisconsin, USA”[23]. The breast cancer dataset contains 699 breast cancer patients’ records. Moreover, the datasets contain the Benign: 458 (65.5%) samples and Malignant: 241 (34.5%) samples. However, I chose the particular parameters for data analysis which are summarized in figure 3.3

No	Factor	Information Factor	Description
1	Id	Numerical	Id
2	Clump Thickness	Numerical	(1-10)
3	Uniformity of Cell Size	Numerical	(1-10)
4	Uniformity of Cell Shape	Numerical	(1-10)
5	Marginal Adhesion	Numerical	(1-10)
6	Single Epithelial Cell Size	Numerical	(1-10)
7	Bare Nuclei	Numerical	(1-10)
8	Bland Chromatin	Numerical	(1-10)
9	Normal Nucleoli	Numerical	(1-10)
10	Mitoses	Numerical	(1-10)
11	Class	Bengin or Malignant	2 for benign, 4 for malignant

Figure 3.2 Parameters for data analysis

### Data Pre-Processing

From the Wisconsin Breast Cancer data, it contains 699 breast cancer patient's data including 11 parameters. We see that the column 'class' has high correlation with all columns except ID Number which has no significance and needs to be removed. Therefore, we removed 'Id' parameter from the data set. If the 'ID number' column is not removed, that the accuracy is affected when we conducted the analysis. In this dataset, there is no missing data that shown in the figure 3.6. But there are some NaN values, it is denoted by '?'. Therefore, we used the dropna() function to remove the NaN values. After cleaning the datasets, we have 683 entries including 10 parameters. Hence, we didn't find any correlated column in this breast cancer dataset (figure).

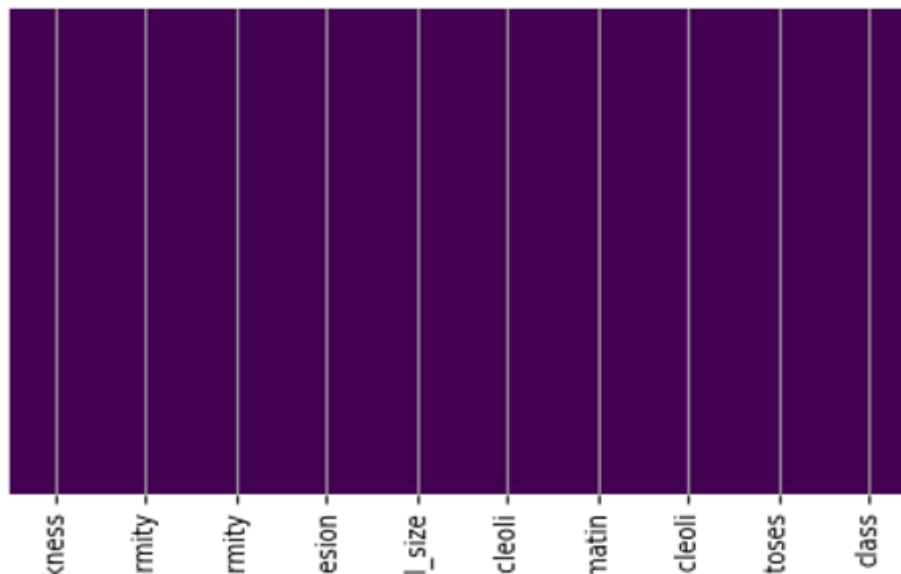


Figure 3.3. No missing values on breast cancer datasets

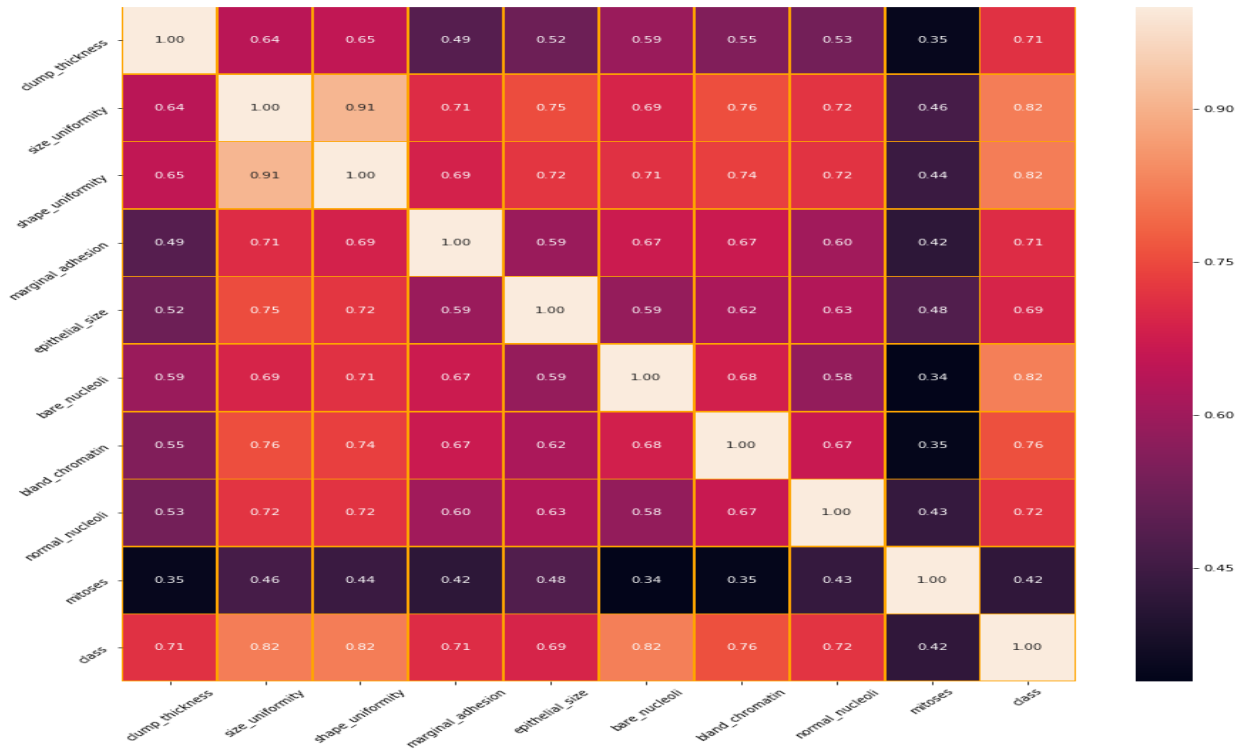


Figure 3.4. Heat map for checking correlated columns for breast cancers

### Evaluation Criteria

In this thesis, we used six machine learning techniques for the early prediction of breast cancer disease. Therefore, the performance measurements of the classifiers are appraised by different statistical procedures. Such as confusion\_matrix (True\_Positive, False\_Positive, True\_Negative, False\_Negative), Recall<sub>i</sub>, Precision<sub>i</sub>, f1- measure etc.[24].

The computation method of the measurement considerations are as follows,

$$\text{Accuracy}_i = (\text{TP}_i + \text{TN}_i) / (\text{TP}_i + \text{FP}_i + \text{TN}_i + \text{FN}_i) \quad (1)$$

$$\text{TPR}_i \text{ or Sensitivity}_i \text{ or Recall}_i = \text{TP}_i / (\text{TP}_i + \text{FN}_i) \quad (2)$$

$$\text{Specificity}_i = \text{TN}_i / (\text{TN}_i + \text{FP}_i) \quad (3)$$

$$\text{Precision}_i = \text{TP}_i / (\text{TP}_i + \text{FP}_i) \quad (4)$$

$$f1_i = 2 * (\text{Recall}_i * \text{Precision}_i) / (\text{Recall}_i + \text{Precision}_i) \quad (5)$$

$$\text{False Positive Rate} = 1 - \text{Specificity}_i \quad (6)$$

The f1\_measure is denoted by the weighted norm of the recall<sub>i</sub> and precision<sub>i</sub>. To classify as a better classifier this the value will be 1 and for the lowest performance, it will be 0.



## **Software and Tools**

In the present study all analysis was implemented in Python version 3.7.0 using Anaconda Distribution including Jupyter Notebook. The version of the notebook server is: 5.6.0-3badce9.

## CHAPTER 4

### RESULT AND DISCUSSION

#### 4.1 Introduction

In this chapter presents analysis of cancer disease detection, we will discuss about the general analysis process of the work. We have tested our model through several measurement experiment. Then we will present the performance of the models and compare it with other classifiers.

#### 4.2 Experimental Results

In this thesis, I have accompanied several analyses to examine the six ML based supervised techniques for diagnosis and forecast of cancer disease. The performance comparison and performance measure of six machine learning classifiers for chronic disease prediction as detailed in the following.

The performance of the selected ML classifiers shows in figure 4.1.1. The SVM classifier attained the uppermost performance with supreme prediction accuracy of 97.07 percent whereas the next maximum classification accuracy is succeeded by NB and RF (i.e. 97%). Moreover, KNN, DT and LR shows the almost same performance by attaining 96% accuracy.

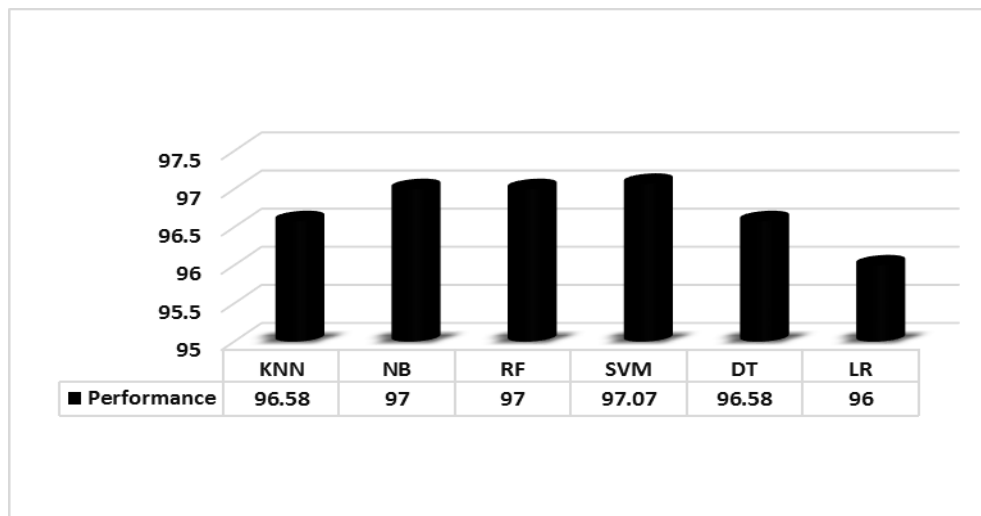


Figure 4.1. The accuracy of six machine learning (Breast Cancer)

#### 4.3 Descriptive Analysis

According to the performance measurements of six classification techniques are illustrated in figure 4.1.2. The results evidently show that the DT and LR reached to the highest precision (97%). NB achieved the highest sensitivity, it's 100%. And NB also achieved the worst specificity (92%). Considering f1 measure, all of classifiers shows

the same performance, it's above 95%, respectively. Figure 4.1.3.demonstratedthe confusion matrix of forecast results for “Naïve Bayes, Random Forest, Support Vector Machine, Decision Tree, KNN and Logistics Regression algorithms”.

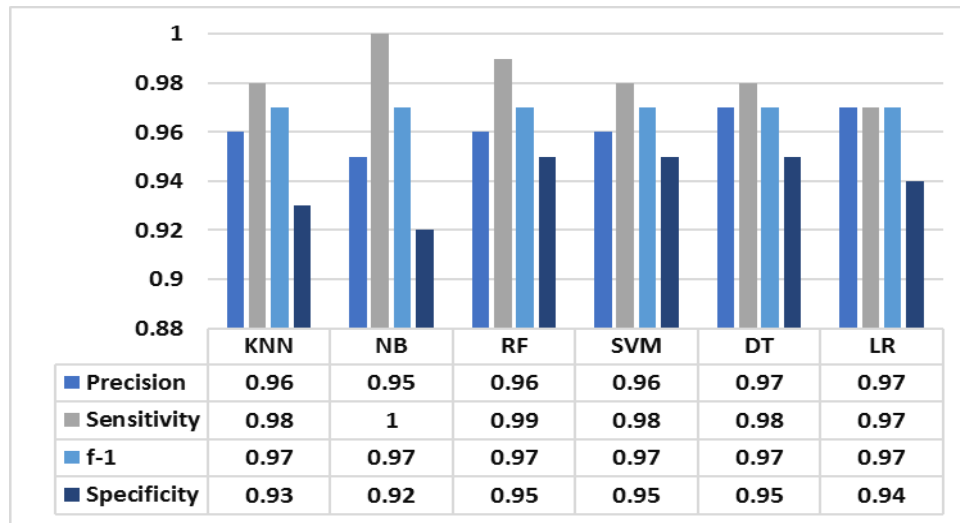


Figure 4.2. Classification Performance Measurements (Breast cancer)

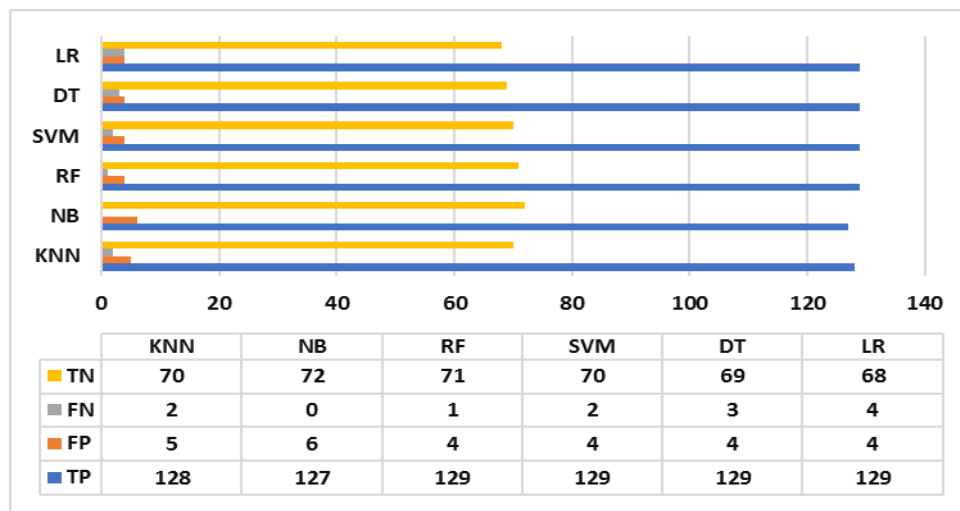


Figure 4.3. Confusion matrix of classification techniques

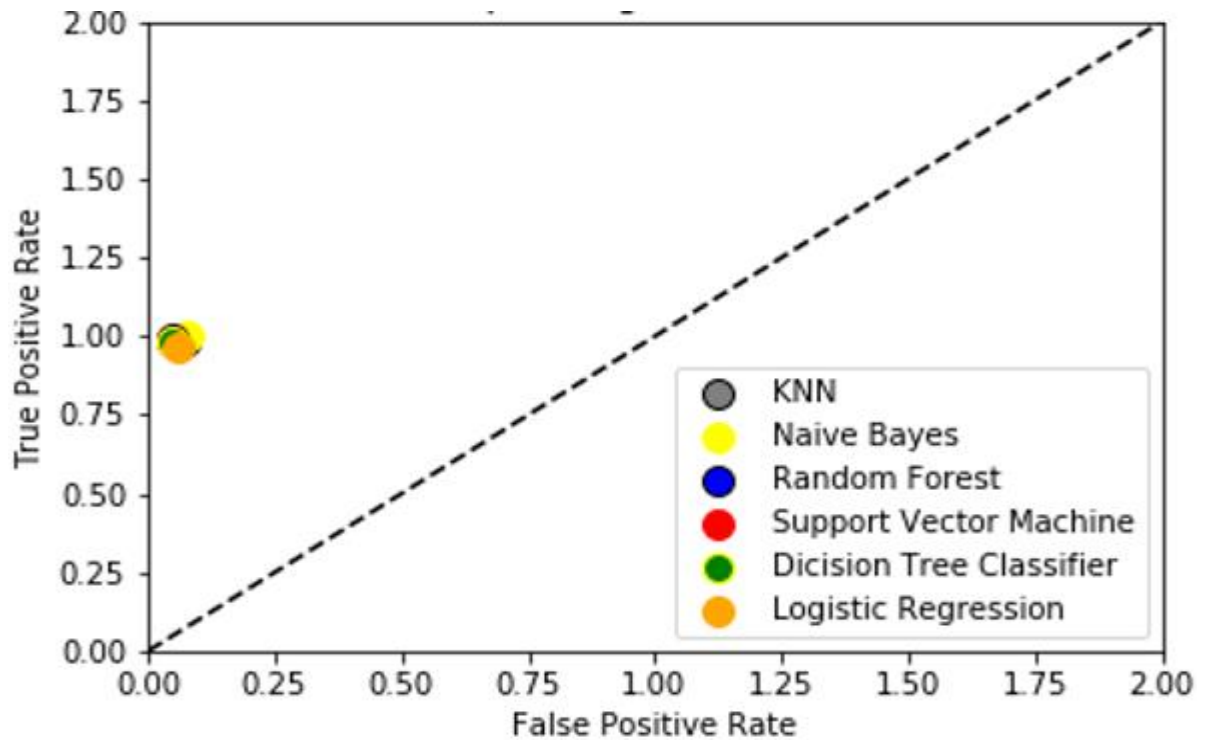


Figure 4.4. Receiver Operating Characteristics curve for Breast Cancer datasets

The prediction result shows the classifiers outcome above 95% for cancer disease detection.

#### 4.4 Summary

Moreover, the six machine learning algorithms are doing very good for cancer disease prediction. In our experiment, it is very essential to recognize about the “receiver operating characteristics (ROC) curve”, which is grounded on “true positive rate (TPR) and false positive rate (FPR)” of these detection results. According to ROC curve (figure 4.1.4.), RF and NB outperformed (Kidney Disease) all other techniques. Furthermore, KNN (Breast Cancer) and SVM (Liver Disease) achieved highest AUC (area under curve) for ROC.

## **CHAPTER 5**

### **CONCLUSION AND RECOMMENDATION**

#### **5.1 Summary of the Study**

There have been many of research study on “machine learning in health care” areas. The outcome of like these technologies are taking a revolutionary transformation in our computing life. Recently, we have got some wonderful applications on this area. But very few of research works on breast cancer prediction are available. In our study, we have followed some approaches to classify the breast cancer disease prediction.

#### **5.2 Conclusion**

In this study, we have depicted several ML based classification techniques. Therefore, we deliver an experimental process on ML based system for the early prediction of breast cancer disease. Therefore, we completed the presentation of the six algorithms which are deployed in the forecast of cancer diseases and assessed by their results using a statistical technique namely is confusion matrix. The experimental performance shows that the Naïve Bayes and Random Forest has achieved the outperform than the other classifiers within cancer datasets. This inspection has usage six ML techniques for the prediction of cancer disease based on some attributes.

#### **5.3 Recommendation**

In addition, this study is part of a project that has the purpose to develop a real time-based computerized tool to give more precise treatment to normal events and make a superior decision to complex situations. The application will be able to early detect in cancer disease in a few minutes and notify the real condition with extreme likelihood of having disease. This application can be remarkably beneficial in low-income countries where is lack of medical institutions and as well as specialized practitioners.

#### **5.4 Implication of Further Study**

In my experiments, related to most work in the study, each classification algorithms were trained and evaluated on a training dataset that comprises both of positive values and negative values. Moreover, the work can be helpful for chronic disease diagnosis and detection by collecting data from different devices and health related sensors,

clinical and medical center and can deliver more accurate results for disease prediction and diagnosis. In my research perspective, there are several directions for the future work in this area of research. We only investigated to some popular supervised machine learning algorithms, it can be choosing more algorithm for build the accurate model of these chronic disease prediction and performance can be more improved. In summary, I have emphasized the research trend and possibility in relation to cancer research and clinical analysis by ML based techniques.

## Appendix: Breast Cancer Dataset

### Sample of Wisconsin Breast Cancer Data [23]

id	clump_thi	size_unifc	shape_un	marginal_	epithelial	bare_nucl	bland_ch	normal_n	mitoses	class
1000025	5	1	1	1	2	1	3	1	1	2
1002945	5	4	4	5	7	10	3	2	1	2
1015425	3	1	1	1	2	2	3	1	1	2
1016277	6	8	8	1	3	4	3	7	1	2
1017023	4	1	1	3	2	1	3	1	1	2
1017122	8	10	10	8	7	10	9	7	1	4
1018099	1	1	1	1	2	10	3	1	1	2
1018561	2	1	2	1	2	1	3	1	1	2
1033078	2	1	1	1	2	1	1	1	5	2
1033078	4	2	1	1	2	1	2	1	1	2
1035283	1	1	1	1	1	1	3	1	1	2
1036172	2	1	1	1	2	1	2	1	1	2
1041801	5	3	3	3	2	3	4	4	1	4
1043999	1	1	1	1	2	3	3	1	1	2
1044572	8	7	5	10	7	9	5	5	4	4
1047630	7	4	6	4	6	1	4	3	1	4
1048672	4	1	1	1	2	1	2	1	1	2
1049815	4	1	1	1	2	1	3	1	1	2
1050670	10	7	7	6	4	10	4	1	2	4
1050718	6	1	1	1	2	1	3	1	1	2
1054590	7	3	2	10	5	10	5	4	4	4
1054593	10	5	5	3	6	7	7	10	1	4
1056784	3	1	1	1	2	1	2	1	1	2
1057013	8	4	5	1	2 ?		7	3	1	4
1059552	1	1	1	1	2	1	3	1	1	2
1065726	5	2	3	4	2	7	3	6	1	4
1066373	3	2	1	1	1	1	2	1	1	2
1066979	5	1	1	1	2	1	2	1	1	2
1067444	2	1	1	1	2	1	2	1	1	2
1070935	1	1	3	1	2	1	1	1	1	2
1070935	3	1	1	1	1	1	2	1	1	2
1071760	2	1	1	1	2	1	3	1	1	2
1072179	10	7	7	3	8	5	7	4	3	4

## REFERENCES

- [1] "WHO | Breast Cancer: Prevention and contro", [www.who.int/cancer/detection/breastcancer/en/index3.html](http://www.who.int/cancer/detection/breastcancer/en/index3.html) [Accessed: 01-Nov-2019]
- [2] "WHO | World Health Organization.",<https://www.who.int/cancer/prevention/diagnosis-screening/breast-cancer/en/>. [Accessed: 01-Nov-2019].
- [3] K. Purushottam. Saxena, and R. Sharma, "Efficient Heart Disease Prediction System," *Procedia Comput. Sci.*, vol. 85, pp. 962–969, Jan. 2016.
- [4] P. Singh, S. Singh, and G. S. Pandi-Jain, "Effective heart disease prediction system using data mining techniques.," *Int. J. Nanomedicine*, vol. 13, pp. 121–124, 2018.
- [5] "Chronic Kidney Disease Basics | Chronic Kidney Disease Initiative | CDC." [Online]. Available: <https://www.cdc.gov/kidneydisease/basics.html>. [Accessed: 12-Dec-2018].
- [6] M. R. Ahmed, M. Arifa Khatun, A. Ali, and K. Sundaraj, "A literature review on NoSQL database for big data processing," *Int. J. Eng. Technol.*, vol. 7, no. 2, pp. 902–906, 2018.
- [7] M. Razu Ahmed, S. M. Hasan Mahmud, M. Altab Hossin, H. Jahan, and S. Rashed Haider Noori, "A Cloud Based Four-Tier Architecture for Early Detection of Heart Disease with Machine Learning Algorithms," 2018, IEEE 4th International Conference on Computer and Communications., 2018
- [8] A. K. Dwivedi, "Analysis of computational intelligence techniques for diabetes mellitus prediction," *Neural Comput. Appl.*, pp. 1–9, Apr. 2017.
- [9] S. M. H. Mahmud and R. Ahmed, "Machine Learning Based Unified Framework for Diabetes Prediction," 2018.
- [10] M. Heydari, M. Teimouri, Z. Heshmati, and S. M. Alavinia, "Comparison of various classification algorithms in the diagnosis of type 2 diabetes in Iran," *Int. J. Diabetes Dev. Ctries.*, vol. 36, no. 2, pp. 167–173, Jun. 2016.
- [11] M. Kukar, I. Kononenko, C. Grošelj, ... K. K.-A. intelligence in, and undefined 1999, "Analysing and improving the diagnosis of ischaemic heart disease with machine learning," Elsevier.
- [12] D. Jain and V. Singh, "Feature selection and classification systems for chronic disease prediction: A review," *Egypt. Informatics J.*, Apr. 2018.
- [13] D. Carvalho, P. R. Pinheiro, and M. C. D. Pinheiro, "A Hybrid Model to Support the Early Diagnosis of Breast Cancer," *Procedia Comput. Sci.*, vol. 91, pp. 927–934, Jan. 2016.
- [14] M. Kumari, "Breast Cancer Prediction system," *Procedia Comput. Sci.*, vol. 132, pp. 371–376, Jan. 2018.
- [15] L. Tapak, N. Shirmohammadi-Khorram, P. Amini, B. Alafchi, O. Hamidi, and J. Poorolajal, "Prediction of survival and metastasis in breast cancer patients using machine learning classifiers," *Clin. Epidemiol. Glob. Heal.*, Oct. 2018.
- [16] H. Asri, H. Mousannif, H. Al Moatassime, and T. Noel, "Using Machine Learning Algorithms for Breast Cancer Risk Prediction and Diagnosis," *Procedia Comput. Sci.*, vol. 83, pp. 1064–1069, Jan. 2016.
- [17] H. Wang, B. Zheng, S. W. Yoon, and H. S. Ko, "A support vector machine-based ensemble algorithm for breast cancer diagnosis," *Eur. J. Oper. Res.*, vol. 267, no. 2, pp. 687–699, Jun.



- 2018.
- [18] M. Gupta and B. Gupta, "A Comparative Study of Breast Cancer Diagnosis Using Supervised Machine Learning Techniques," in 2018 Second International Conference on Computing Methodologies and Communication (ICCMC), 2018, pp. 997–1002.
  - [19] L. Abdel-Ilah and H. Šahinbegović, "Using machine learning tool in classification of breast cancer," Springer, Singapore, 2017, pp. 3–8.
  - [20] M. Amrane, S. Oukid, I. Gagaoua, and T. Ensari, "Breast cancer classification using machine learning," in 2018 Electric Electronics, Computer Science, Biomedical Engineerings ' Meeting (EBBT), 2018, pp. 1–4.
  - [21] 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.
  - [22] X. Liu et al., "A Hybrid Classification System for Heart Disease Diagnosis Based on the RFRS Method," *Comput. Math. Methods Med.*, vol. 2017, p. 8272091, 2017.
  - [23] W. H. Wolberg and O. L. Mangasarian, "Multisurface method of pattern separation for medical diagnosis applied to breast cytology.," *Proc. Natl. Acad. Sci.*, vol. 87, no. 23, pp. 9193–9196, Dec. 1990.
  - [24] "ConfusionMatrix." [http://www2.cs.uregina.ca/~hamilton/courses/831/notes/confusion\\_matrix/confusion\\_matrix.html](http://www2.cs.uregina.ca/~hamilton/courses/831/notes/confusion_matrix/confusion_matrix.html), [Accessed: 20-Dec-2018].

## Breast Cancer Prediction

### ORIGINALITY REPORT

<b>18%</b>	%	%	<b>18%</b>
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

### PRIMARY SOURCES

<b>1</b>	<b>Submitted to Daffodil International University</b> Student Paper	<b>2%</b>
<b>2</b>	<b>Submitted to National College of Ireland</b> Student Paper	<b>2%</b>
<b>3</b>	<b>Submitted to University of Dammam</b> Student Paper	<b>1%</b>
<b>4</b>	<b>Submitted to Victoria University</b> Student Paper	<b>1%</b>
<b>5</b>	<b>Submitted to Guru Nanak Dev Engineering College</b> Student Paper	<b>1%</b>
<b>6</b>	<b>Submitted to CONACYT</b> Student Paper	<b>1%</b>
<b>7</b>	<b>Submitted to Infile</b> Student Paper	<b>1%</b>
<b>8</b>	<b>Submitted to University of Sheffield</b> Student Paper	<b>1%</b>
<b>9</b>	<b>Submitted to Universiti Kebangsaan Malaysia</b>	