# Performance Analysis of Breast Cancer: A Machine Learning Approach

Tania Khatun [1]
Dept. of Computer Science and Engineering
Daffodil International University
Dhaka, Bangladesh
tania.cse@diu.edu.bd

Md. Ashiqul Islam [2]
Dept. of Computer Science and Engineering
Daffodil International University
Dhaka, Bangladesh
ashiqul15-951@diu.edu.bd

Md. Sagar Hossen [3]
Dept. of Computer Science and Engineering
Daffodil International University
Dhaka, Bangladesh
sagar15-1504@diu.edu.bd

Sabiha Jannat Anni [4]
Dept. of Computer Science and Engineering
Daffodil International University
Dhaka, Bangladesh
sabiha15-1376@diu.edu.bd

Md. Mahfuzur Rahman Utsho [5]
Dept. of Software Engineering
Daffodil International University
Dhaka, Bangladesh
mahfuzur35-175@diu.edu.bd

Mst. Fatematuz Zohura [6]
Dept. of Computer Science and Engineering
Daffodil International University
Dhaka, Bangladesh
fatematuz15-3195@diu.edu.bd

Robaiya Akter Rimi [7]
Shahabuddin Medical College
Gulshan-2, Dhaka, Bangladesh
robaiyaakter00@gmail.com

*Abstract—* Nowadays, breast cancer is the most emerging disease among women both in developed as well as developing countries. Due to increased life prospects, increased urbanization, and the relinquishment of western societies, the rareness of breast cancer is supersizing in the developing world. Even it became a second popular cause of cancer that has already been announced. It's very hard to identify the early symptom of this type of cancer for reducing numerous death. Different methods of machine learning and data mining techniques are using for medical diagnosis. In this study, four machine-learning algorithms are applying to analyze breast cancer in the inflammation stage and dig up the most cabbalistic and non-cabbalistic risk factors. To analyze breast cancer data from the Coimbra dataset from the UCI machine learning repository to create accurate prediction models for breast cancer. For getting better performance and to get higher accuracy Naïve Bayes (NB), Random Forest (RF), Multilayer Perceptron (MLP), Simple Logistic Regression (SLR) are using to find out some higher accuracy sequentially 70%, 68%, 85%, and 75%. Among all the above algorithms a better accuracy was achieved using Multi-layer Perceptron. Linear Regression (LiR) models are applying to dig up the most cabbalistic and non-cabbalistic risk factors of breast cancer. These results will help the doctor to detect breast cancer easily in the early stage and take the necessary steps.

*Keywords— Machine Learning, Random Forest, Naïve Bayes, Multilayer Perceptron, Simple Logistic Regression, UCI Machine Learning Repository, Inflammation Stage.*

## I. INTRODUCTION

Early detection of the disease has become an important issue in the medical field due to the increasing population of the world. With the rapid growth of the population, the risk of dying from cancer is progressive each day. "Cancer" is a scary disease for most people. But in women, "breast cancer" can be even more frightening, because it is directly related to the woman's body [1]. However, men can also get breast cancer, but men are much less likely to get the disease than women. Breast cancer is a major concern in Bangladesh. According to the World Health Organization, an exceeding 15,000 populace are diagnosed with breast cancer in Bangladesh every year, of which more than 98 percent are women, and a very small amount of men are also diagnosed with breast cancer. According to expert doctors, the number of people suffering from breast cancer is rapidly growing. According to medical science, when some breast cells grow abnormally, those irregular and extra cells divide into tumors or lumps and expansion to diverse virtue of the organism care of lymphatic

and other blood vessels [2]. The reasons for the increasing number of breast cancer cases in Bangladesh and around the world are given below: i) there have been many changes in our way of life and eating habits. It can happen if someone has a family history of breast cancer. ii) If someone menstruates before the age of twelve or has late menopause, they are also at risk. iii) Radioactivity also increases the risk of breast cancer. It can be caused by late adoption or those who do not have children, or not breastfeeding the child. iv) If you have more fat and animal meat than vegetables or fruits in your diet, eating more processed foods increases your risk of breast cancer. v) Those who are overweight also have a risk of breast cancer. vi) Prolonged use of birth control pills or hormone injections can lead to breast cancer. vii) Breast cancer also increases with age. After the age of 50, this risk increases a lot. Breast cancer can be detected through various tests. These are- 1. Mammography 2. Ultrasonography 3. MRI 4.FNAC 5. Biopsy/meat examination Treatment The treatment of breast cancer is mainly divided into several parts. These are- 1. Surgery 2. Chemotherapy 3. Radiotherapy 4. Hormone therapy 5. Targeted therapy.

In this research, four algorithms called Naive Bayes, random forest, Simple logistic regression, and Multi-layer perception are using to identify breast cancer. According to the algorithms of machine learning and deep learning, the amount of detection of breast cancer in the human body is-70% for Naive Bayes, 68% for Random Forest, 85% for Multi-layer Perceptron, and 75% for Simple Logistic Regression. In the study, a linear regression model is applied to analyze the most significant and non-significant risk factors of breast cancer [3].

TABLE 1. Abbreviation

| Abbreviation | Full form |
|---|---|
| UCI | University of California Irvine |
| MRI | Magnetic Resonance Imaging |
| FNAC | Fine Needle Aspiration Cytology |
| MSP | Minimum Support Price |
| DBT | Direct Benefit Transfer |
| BCRAT | Breast Cancer Risk Assessment Tool |
| MDT | Maggot Debridement Therapy |
| AUC | Area Under the Curve |
| DSS | Decision Support System |
| RO | Reverse Osmosis |
| CAD | Computer Aided design |
| CNN | Convolutional Neural Network |
| USOL | University School of Open Learning |
| MSE | Mental Status Examination |
| MCP | Membrane Cofactor Protein |
| ROC | Receiver Operating Characteristic |
| LDA | Left Displaced Abomasum |
| HOMA | Homeostatic Model Assessment |
| BMI | Body Mass Index |
| WSI | Weekly Stress Inventory |
| DRS | Designated Record Set |
| ELM | Extreme Learning Machine |

## II. RELATED WORK

To detect breast cancer in the early-stage researchers [4] used screening mammography but after creating new MSP images from DBT this will improve the accuracy of screening mammography worldwide from 14% to more. The research gap is to handle a large amount of data smoothly. They [5] used triple assessment to detect breast cancer in the inflammation stage. Researchers [6] used diagnostic accuracy to build up computer vision methods of a full spectrum of breast biopsy samples. They [1] works for the breast cancer prediction of Chinese women. They used XGBoost, RF, and Deep Neural Network (DNN) methods. XGBoost is highly reliable and performs well in a small dataset. If noisy data are determined XGBoost can overfit the data. They [7] works for the research and prevention of Breast Cancer. They used Decision Tree (DT), RF, SVM, Neural Network (NN), and LR for preventing Breast Cancer. DT does not require normalization data. Its training process is relatively expensive. They [8] analyzed MDT decisions for published cancer guidelines. They [9] works with multiparametric MRI to detect breast cancer early. They [10] works with large and complex visual data and the picture of the tumor. They used a machine learning method named DRS and get the best outcome. They [11] works with the symptoms of breast cancer. They used a machine-learning algorithm and got 0.82 precision. To detect the germline single nucleotide polymorphisms (SNPs) they [12] used Random Forest Regression (RFR). RF is flexible for both classification and regression problems but it needs more time to determine the class. They [13] works with 10 fold cross-validation method.

They [14] works with Breast Cancer Prognosis using ML. They used DSS and RO for exact prognosis with an accuracy of 86%. They [15] works with unpredictable subtypes of breast cancer, and they used a multi-level diagnostic approach. They [3] work to dig up the diagnostic accuracy with the most cabbalistic and non-cabbalistic risk factors of chronic kidney disease. They [16] works with CAD tools for the development of an automated breast cancer diagnosis. They used Tabu search for their research. They [17] works with the advanced diagnosis of breast cancer. They used an MRI dataset for the best outcome. They got 99.5% sensitivity. They [18] works for the prediction and diagnosis of breast cancer. They used two datasets and Support Vector Regression (SVR), Lasso Regression (LaR), Kernel Ridge Regression (KRR), K-neighborhood Regression (KNR), and Decision Tree Regression (DTR) methods. They got the best result using cross-validation. To detect and visualize the prognostic breast cancer they [19] used RF and got the highest accuracy of 82.7%. Researchers [20] apply genomic research to specify functions of transparent factors to place the state. They showed ConSReg perfectly recognizes the transparent factors (23.5-25) % better than the enrichment-based approach. Author [21] provides an optimized automatic software testing model through differential evolution and ant colony optimization hybrid model to improved accuracy and

reliability in software testing. Author [22] minimizes the classification error with better accuracy using a modified version of ELM.

They [23] propose an effective approach for identifying breast cancer. They used Gradient Tree Boosting (GTB) method. To classify the data of breast cancer they [24] used SVM, C4.5, NB, KNN. Among them, they found the highest accuracy of 97.13% with SVM. SVM is very effective in high-dimensional spaces. They [25] works with breast ductal carcinoma for predicting Breast Cancer using WSI. They [26] works with data mining techniques to improve predictive models for breast cancer recurrence. They used C4.5, Support SVM, and ANN and find the best accuracy of 0.957 with SVM. They [27] works to predict the breast tumor early. They used LR, KNN, SVM, NB, DT, RF, and Rotation Forest. Among them, the logistic regression model gives the maximum accuracy of 98.1%. KNN stores the training dataset and learns from it only at the time of making real-time predictions. This makes the KNN algorithm much faster than other algorithms. They [28] used two consecutive methods to the prognosis of breast cancer. They found 98% accuracy using SVM and LDA. They [29] works for breast cancer prediction. They used six machine learning algorithms to predict breast cancer. They [30] works to predict breast cancer early. They used J48, NB, and Sequential Minimal Optimization (SMO) methods. And Wisconsin Breast Cancer (WBC) and Breast Cancer datasets. They[31] works for disease prediction using different machine learning algorithms. They used SVM, NB, and RF for the best result. They got the highest accuracy 53% using RF. NB is suitable for solving multi-class prediction problems. NB assumes that all predictors (or features) are independent, rarely happening in real life. They [32] works for breast cancer prediction and treatment using RF, KNN, and NB. They [33] works with the CAD system for breast cancer. They provide an efficient result from 320 cases. To detect and identify the cancer cells they [34] used NB, SVM, and Artificial Neural Network (ANN). They got the result of 98.99%. To detect the disease they [35] used SVC, K means clustering, RF, CNN, and LR techniques. They got the highest accuracy of 98.4% using CNN.

## III. PROPOSED METHOD

Breast cancer is the most sensitive issue in clinical science and this is the more expensive and time-effective diagnosis system all over the world. Identifying breast cancer in the early stage is very challenging. An effective model can help us to diagnose breast cancer in the inflammation stage. This research work has proposed a model to predict breast cancer based on ML. Also, this research work is applying NB, RF, SLR, and MLP to forecast breast cancer systematically. The main purpose of this research work is to collaborate medical science with artificial intelligence to find out the best algorithm performance and factor investigation and predict the results more accurately and time-consuming. This study helps the doctor and physicians to diagnose breast cancer and take necessary steps for their patients.
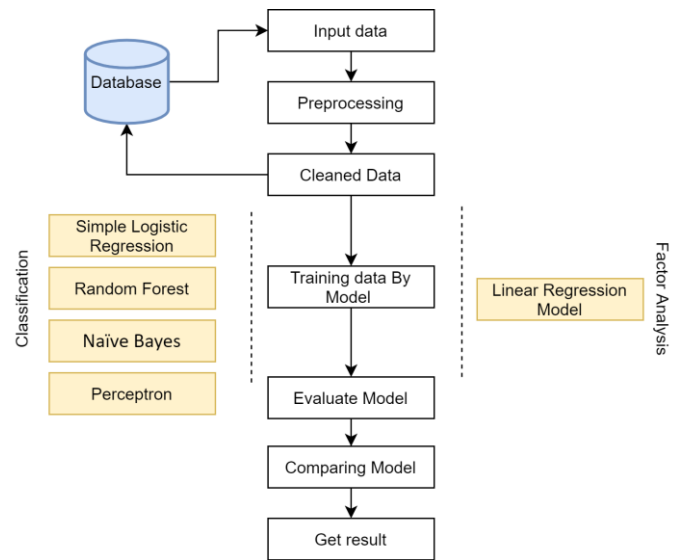


Fig.1. Block Diagram of Proposed Scheme

## IV. EXPERIMENT ANALYSIS

This section, describe the implementing our model and we have applied four different ML model to forecast the performance of breast cancer data.
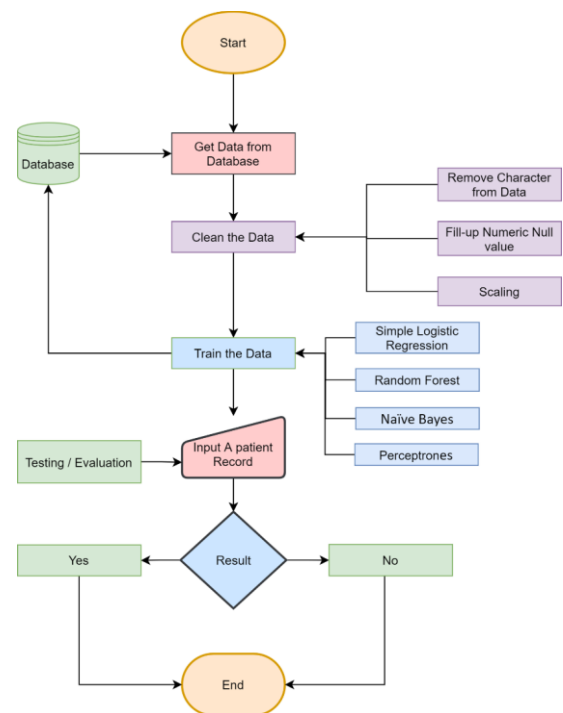


Fig.2. Flowchart of Proposed Scheme

Breast cancer is more prevalent among relatives of patients than overall people. The greater prevalence of breast cancer in monozygotic twins of patients than among dizygotic twins or

siblings means that genetic differences, people's lifestyles are responsible for the majority of cases of kinship swarm [36]. The datasets used in this article, as well as the pre-processing of our data and the various classification models used to forecast female breast tumors, are described in this portion. Some steps are following to fulfill the terms.

## A. Algorithms

**Step 1:** Start The procedure.
**Step 2:** Input Data M.
**Step 3:** Clean the data M: remove character C, Null value N, Scaling S.
**Step 4:** Data train with a model.
**Step 5:** Testing the recorded data.
      if breast cell data is highly match
         it is affected cell
      else if a healthy person.
**Step 6:** End Procedure.

## B. DATASET DESCRIPTION

To construct reliable breast cancer prediction models, data mining strategies are applying to study breast cancer data from the Coimbra dataset from the UCI machine learning repository [34]. There are accurate 116 instances in this dataset where the characteristics of the data set are multivariate. Here, 64 patients are affected with breast cancer and 52 are well, clinical characteristics were detected or assessed. There are ten empirical predictors and a binary dependent variable that suggest the occurrence or absence of breast cancer. The predictors are anthropometric data and parameters, which can be obtained by routine blood sampling [37]. If these predictors are reliable, prediction models based on them may be used as breast cancer biomarkers. There we have used 10 attributes in the data sets for classification and the label of attributes are healthy controls and patients.

TABLE 2. Dataset attribute Description

| Attribute | Description |
|---|---|
| Age (years) | Patients age in years. |
| BMI (kg/m2) | The numerical value of the patient's weight with the patient's height. |
| Glucose (mg/dL) | A glucose level under 140 mg/dL (7.8 mmol/L) is ordinary. A scrutinizing someplace in the scope of 140 and 199 mg/dL (7.8 mmol/L and 11.0 mmol/L) exhibits prediabetes. |
| Insulin (μU/mL) | Insulin may be a chemical made by your pancreas that controls the measure of glucose in your cardiovascular system out of nowhere. Ordinary scope of insulin at Fasting < 25 mIU/L, a half-hour after glucose organization 30-230 mIU/L, and 1 hour after glucose administration 18-276 mIU/L, 2 hours after glucose administration 16-166 mIU/L. |
| HOMA | The normal HOMA-IR value of healthy humans ranges from 1.7-2.0. |
| Leptin (ng/mL) | Ordinary scopes of serum BAP, OC, and leptin were 7.9–20.9 U/L, 3.1–12.7 ng/ml, and 2.5–21.8 ng/mL, separately [38]. |
| Adiponectin (μg/mL) | Adiponectin levels in people range from 0.5 to 30 μg/mL, which is 1,000-overlap higher than the degrees of different chemicals, like insulin and leptin (18) |
| Resistin (ng/mL) | For human resistin (MW 11.42 kDa), the concentrations of 25, 50, and 100 ng/ml equate to 2.19, 4.38, and 8.76 nm [39]. |
| MCP-1(pg/dL) | The information was assessed with KIM-E programming (USOL, Prague, Czech Republic); the recognition furthest reaches of the MCP-1 examine was 5.0 pg/mL |
| Class | 1=Healthy controls, 2=Patients |
| Age (years) | Patients age in years. |

## C. DATA PRE-PROCESSING

Pre-processing data is an essential aspect of preparing data for simulation. Data pre-processing entails several steps, including data transformation, data cleaning, and feature selection. Data cleaning and transformation are strategies for eliminating outliers and standardizing data so that it can be used to construct models faster. This segment focuses mostly on submission. This section emphasizes the function selection aspect of data pre-processing and focuses mainly on model implementations. In this section, some procedures are following. First of the part removing the character from data and fill up the Numeric Null value. And then finally doing scaling the data.

## D. MODEL DESCRIPTION

Training data is the data used to train an algorithm or machine learning model to predict the result. If supervised learning or a hybrid approach is using that involves it, the data will benefit from data tagging or annotation. In this stage, four algorithms are factitive for training the data and Linear Regression to discover the hugest and non-critical danger factors.

### D.1. SIMPLE LOGISTIC REGRESSION

Logistic regression is mathematical analysis, as are all regression analyses. LR is applied to characterize information to show the connection between the dependent variable and one nominal, interval, ordinal, or ratio-level independent variable.

$$P = \frac{e^{a+bX}}{1+e^{a+bX}} \tag{1}$$

Here P is the target variable, it varies from 0 to 1. X is the independent variable. In Fig.2. We have shown a sample picture of the Simple Logistic Regression model.
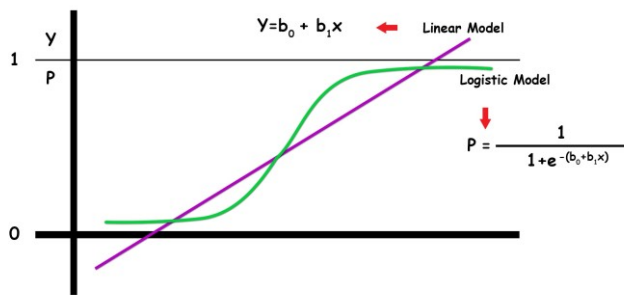
Fig.3. Sample of the process of Simple Logistic Regression Model



## D.2. RANDOM FOREST

Random forest is a nice algorithm to carry out early on in the model development process. It's difficult to make an "evil" random forest because of its simplicity. Random forest is a modular, user-friendly machine learning algorithm that achieves excellent results in most cases without the need for hyper-parameter tuning. It is perhaps one of the most widely used due to its simplicity and flexibility.

$$MSE = \frac{1}{N}\sum_{i=2}^{N}(fi - yi)^2 \quad (2)$$

Here, yi is the actual value, fi is the model returned value, and N is for data points. Random forest performing based on classification data. So we have to often use the Gini index. Gini index can be calculated as:

$$Gini = 1 - \sum_{i=1}^{C}(p_i)^2 \quad (3)$$

Here, Gini is a class and probability to determine each branch of nodes, pi represents the relative frequency of the class. In Fig.3. We have shown a sample model of the Random Forest algorithm.
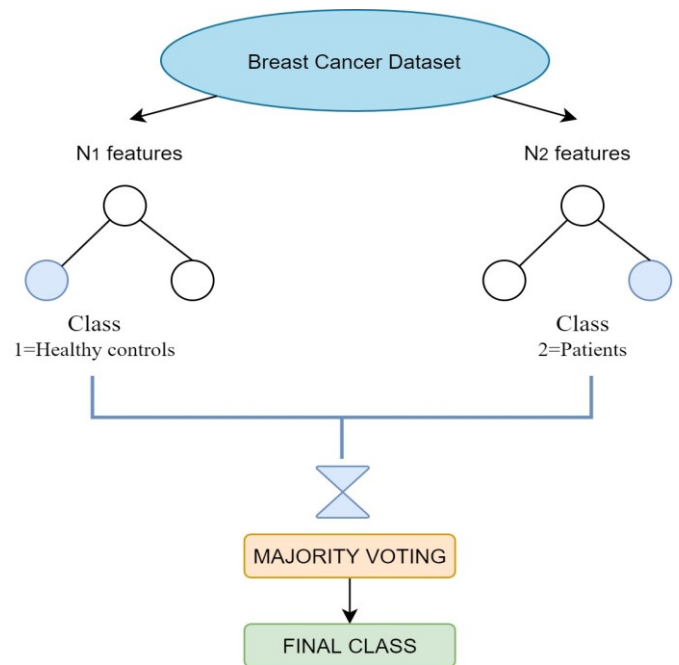
Fig.4. Sample of the process of Random Forest Model

## D.3. NAÏVE BAYES

The Naive Bayes algorithm is a supervised learning model that uses the Bayes theorem to solve classification problems. Based on the likelihood of each trait belonging to the training set, the Naive Bayes algorithm estimates the new data instances class [40]. Nave Bayes forecasts datasets based on the premise that attributes belonging to a class that is distinct from one another.

$$P(M|N) = \frac{P(N|M)P(M)}{P(N)} \quad (4)$$

Here, the Left side Probability of M occurring given evidence and B has already occurred. On the right side, the Probability of N occurring given evidence and M has already occurred in the probability of M occurrence times and divided by probability of N occurrence. In Fig.4. We have shown a sample model of the Naïve Bayes algorithm.
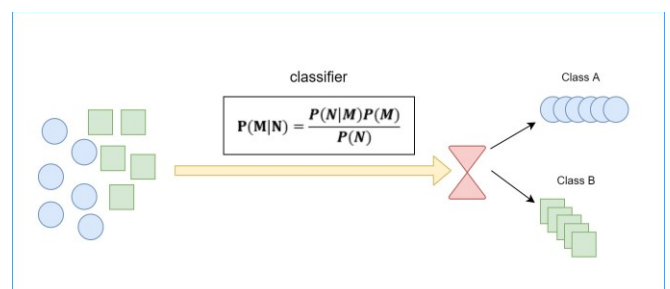


Fig.5.Sample of the process of Naïve Bayes Model

## D.4. MULTILAYER PERCEPTRON

The Perceptron algorithm is a two-class ML model (binary classification). It's a form of the neural network model, and it's the most basic. A single node or neuron processes a row of data and predicts a class name. On the given equation "y" is output w is weight "x" is input and θ is the angle of "w" and "x".

$$y = 1 \, if \, \sum_{i=1}^{n} w_i * x_i - \theta \geq 0 \qquad (5.1)$$

$$y = 0 \, if \, \sum_{i=1}^{n} w_i * x_i - \theta < 0 \qquad (5.2)$$

In Fig.5. We have shown a sample model of the Multilayer Perceptron algorithm.
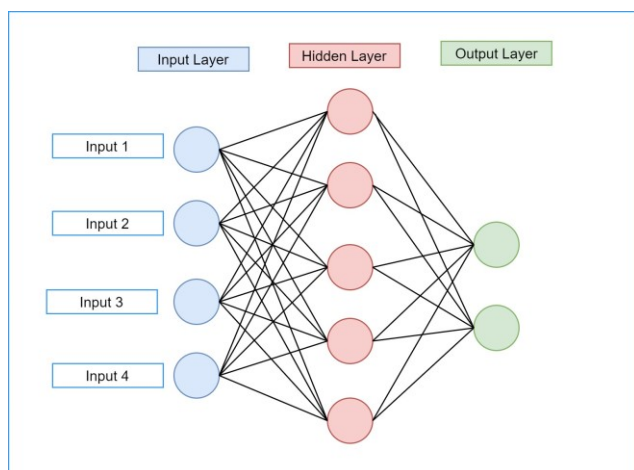


Fig.6. Sample of the process of Multilayer Perceptron Model

## E. TESTING THE DATA EVALUATION

Four algorithms were used to compare data quality: simple logistic regression, random forest, Naive Bayes, and Multi-layer perceptron. After training the data with algorithms then input a patient record as data for testing. For the model validation using all of the algorithms, the dataset was divided into a training set (80%) and a testing set (20%). Each algorithm was evaluated with precision, Recall, F-Measure, and ROC Area.

## F. RESULT ANALYSIS

To make a better prediction using machine learning, several strategies are employing in this paper to choose the best characteristics from a large number of data sets regarding breast cancer using several algorithms. Data mining strategies are factitive to explore breast cancer data from the UCI machine learning repository's Coimbra dataset to construct reliable breast cancer, prediction models [34].
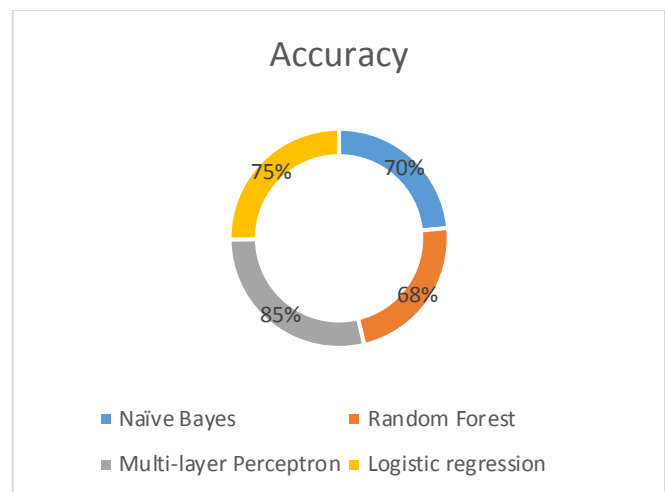


Fig.7. Accuracy pie chart

In this article, four algorithms are employing to achieve improved efficiency and higher precision. These algorithms are Nave Bayes, Random Forest, Perceptron, and Basic Logistic Regression, and by using them, from Fig.7, this can obtain higher accuracy in the following order: 70 %, 68 %, 85 %, and 75%.

TABLE 3. Accuracy, AUC

| Algorithms | Precision | Recall | F-Measure | ROC Area |
|---|---|---|---|---|
| Naïve Bayes | 0.78 | 0.77 | 0.77 | 0.82 |
| Random Forest | 0.68 | 0.68 | 0.68 | 0.78 |
| Perceptron | 0.85 | 0.85 | 0.84 | 0.89 |
| Simple Logistic Regression | 0.75 | 0.75 | 0.74 | 0.81 |

From table- 3, there are different accuracy and AUC where 85% is the best accuracy, precision is 0.85, and 0.89 is the best ROC area which is for the Multi-layer Perceptron algorithm. This is why Multi-layer perceptron performs better than other machine learning algorithms and it's clear to us that the patient and doctor can easily find the existence of the breast cancer of patient's body using all the technology means all model but Multi-layer Perceptron performs better and it is the best fit model for this dataset. Finding the risk factor using the linear regression model. The factor analysis result shows in Table.4.

TABLE 4. Risk Factor Analysis

| Algorithm Name | Factor |
|---|---|
| Linear Regression Model | -0.0275 * BMI + -0.0135 * Glucose + 0.0472 * Insulin + -0.1429 * HOMA+ 0.0083 * Resistin + 0.7811 |

Generally, risk factors are the factors or attributes that cause breast cancer. For analyzing the risk factors of breast cancer a machine learning algorithm is proposed which is linear regression. All the required breast cancer-affected patients' data were culled from the UCI repository. Basically, in the user data, there are some attributes like BMI, Glucose, Insulin, HOMA, Resistin, and Class [41]. Depending on these attributes and their range in the human body, machine learning algorithms like linear regression can analyze breast cancer factors within a very short time with better accuracy. We get an accuracy of 78.11% by using this algorithm. The algorithm was used to predict risk factors rather than predicting breast cancer [6]. As a result, it will be easy for doctors that they can identify the main reasons behind breast cancer. From Table.4, it was clear that Insulin was the main reason for breast cancer and it is the most significant factor. On the other hand, Resistin seems to behave less effect of causing breast cancer. So from the supper table, we can say that Insulin is the most significant, and Resistin is a non-significant factor in breast cancer.

## V. DISCUSSION

Machine learning models are deploying to identify the risk analysis and breast cancer prognosis. Four Machine learning algorithms are applying in the Coimbra dataset. All the algorithms (Naïve Bayes, Random Forest, Multi-layer perceptron, Simple logistic regression) perform very well and their accuracy is very close. From the analysis, we achieve the highest accuracy from Multi-layer Perceptron.

TABLE 5. Comparative Analysis

| Studies | Methods | Accuracy |
|---|---|---|
| Hiba Asri et al. [2] | SVM, C4.5, NB, KNN | 97.13% with SVM |
| Ak et al. [25] | LR, KNN, SVM, NB, DT, RF, Rotation Forest | 98.1% with LR |
| Leili Tapak et al. [27] | NB, RF, SVM, LSSVM, Adabag, LR, LDA | 93% with SVM and LDA |
| Ray A. et al. [30] | SVM, NB, RF, CART | 98.14%% with SVM |
| Ganggayah et al. [32] | DT, RF, NN, exBoost, LR, SVM | 82.7% with RF |

Many researchers perform very well to identify the breast cancer prognosis. Some authors apply a support vector machine and it will provide a good accuracy level (93-98) % which is praised. From Table.5. From the previous work analyzing authors are applying DSS, SVM, C4.5, NB, KNN, LR, DT, RF, Adabag, exBoost, LDA, etc., and they achieve 80% accuracy and they are working with only identify the breast cancer prognosis. From the four Machine learning algorithms, we predict breast cancer with 85% accuracy using Multi-layer perceptron and analysis the risk factor with linear regression, the accuracy of 78.11% to dig up the most significant and non-significant risk analysis. From the result analysis, Insulin is the most cabbalistic and resistin is the non-cabbalistic factor [3]. From the data analysis, we can show that in normal breast cell human bodies glucose ratio is 80-85. Upper than 85 is more dangerous for health and it causes abnormal tissue in the human body. Age is one of the important factors for normal and abnormal tissue. More than 45 aged people are much affected by breast cancer. In general, the ratio of BMI, Insulin, Homa, and Resistin values upper than 20.83, 12.305, 1.011, and 10.96 is dangerous for human breast cells. It causes fatal damages in the breast cell and rapidly grows cancer and tumors cell in the breast. It influences our bodies' metabolic systems. It is a major concern for physicians and doctors to find out the prognosis of breast cancer in the inflammation stage. Our model will help the clinical to detect the disease effectively with low cost and time-consuming process.

## VI. CONCLUSION

A comparative study of the four ML algorithms is presented in this paper for the forecast of breast cancer. Here the accuracy of the classification strategies is evaluated based on the selected ones Classified algorithms. The drawback of the study is the small amount of data. A large dataset can help to improve the performance. In the future image processing technique are applying to identify breast cell mammography. In this study, we have concluded by analyzing the results obtained in the comparative discussion of four ML algorithms. Multi-layer Perceptron has the highest detection rate. Analyzing the ML and DL through the linear regression model, the most common factor we have identified is -0.0472 * Insulin +. This study determines the model performance systematically to diminution the patient's diagnosis rate in the early stage and helps the doctors in clinical practice [27].

## REFERENCES

[1] C. Hou *et al.*, "Predicting breast cancer in Chinese women using machine learning techniques: algorithm development," *JMIR Med. informatics*, vol. 8, no. 6, p. e17364, 2020.

[2] 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, 2016.

[3] M. A. Islam, S. Akter, M. S. Hossen, S. A. Keya, S. A. Tisha, and S. Hossain, "Risk Factor Prediction of Chronic Kidney Disease based on Machine Learning Algorithms," in *2020 3rd International Conference on Intelligent Sustainable Systems (ICISS)*, 2020, pp. 952–957.

[4] W. Lotter *et al.*, "Robust breast cancer detection in mammography and digital breast tomosynthesis using an annotation-efficient deep learning approach," *Nat. Med.*, vol. 27, no. 2, pp. 244–249, 2021.

[5] Z. Salod and Y. Singh, "Comparison of the performance of machine learning algorithms in breast cancer screening and detection: A protocol," *J. Public health Res.*, vol. 8, no. 3, 2019.

[6] E. Mercan, S. Mehta, J. Bartlett, L. G. Shapiro, D. L. Weaver, and J. G. Elmore, "Assessment of machine learning of breast pathology structures for automated differentiation of breast cancer and high-risk proliferative lesions," *JAMA Netw. open*, vol. 2, no. 8, pp. e198777–e198777, 2019.

[7] Y. Li and Z. Chen, "Performance evaluation of machine learning methods for breast cancer prediction," *Appl Comput Math*, vol. 7, no. 4, pp. 212–216, 2018.

[8] A. Tahmassebi *et al.*, "Impact of machine learning with multiparametric magnetic resonance imaging of the breast for early prediction of response to neoadjuvant chemotherapy and survival outcomes in breast cancer patients," *Invest. Radiol.*, vol. 54, no. 2, p. 110, 2019.

[9] Y. Ji *et al.*, "Independent validation of machine learning in diagnosing breast Cancer on magnetic resonance imaging within a single institution," *Cancer Imaging*, vol. 19, no. 1, pp. 1–11, 2019.

[10] S. Sharma, A. Aggarwal, and T. Choudhury, "Breast cancer detection using machine learning algorithms," in *2018 International Conference on Computational Techniques, Electronics and Mechanical Systems (CTEMS)*, 2018, pp. 114–118.

[11] S. Lee *et al.*, "Machine learning on genome-wide association studies to predict the risk of radiation-associated contralateral breast cancer in the WECARE Study," *PLoS One*, vol. 15, no. 2, p. e0226157, 2020.

[12] M. M. Rahman, Y. Ghasemi, E. Suley, Y. Zhou, S. Wang, and J. Rogers, "Machine Learning Based Computer Aided Diagnosis of Breast Cancer Utilizing Anthropometric and Clinical Features," *IRBM*, 2020.

[13] A. Toprak, "Extreme learning machine (elm)-based classification of benign and malignant cells in breast cancer," *Med. Sci. Monit. Int. Med. J. Exp. Clin. Res.*, vol. 24, p. 6537, 2018.

[14] H. Behravan *et al.*, "Machine learning identifies interacting genetic variants contributing to breast cancer risk: A case study in Finnish cases and controls," *Sci. Rep.*, vol. 8, no. 1, pp. 1–13, 2018.

[15] A. Saha *et al.*, "A machine learning approach to radiogenomics of breast cancer: a study of 922 subjects and 529 DCE-MRI features," *Br. J. Cancer*, vol. 119, no. 4, pp. 508–516, 2018.

[16] S. V. J. Jaikrishnan, O. Chantarakasemchit, and P. Meesad, "A Breakup Machine Learning Approach for Breast Cancer Prediction," in *2019 11th International Conference on Information Technology and Electrical Engineering (ICITEE)*, 2019, pp. 1–6.

[17] R. Turkki *et al.*, "Breast cancer outcome prediction with tumour tissue images and machine learning," *Breast Cancer Res. Treat.*, vol. 177, no. 1, pp. 41–52, 2019.

[18] H. Sanyal, P. Saxena, and R. Agrawal, "Innovative Approach for Prediction of Cancer Disease by Improving Conventional Machine Learning Classifier," *Cybern. Cogn. Mach. Learn. Appl. Proc. ICCCMLA 2020*, p. 281, 2021.

[19] N. I. R. Yassin, S. Omran, E. M. F. El Houby, and H. Allam, "Machine learning techniques for breast cancer computer aided diagnosis using different image modalities: A systematic review," *Comput. Methods Programs Biomed.*, vol. 156, pp. 25–45, 2018.

[20] Q. Song, J. Lee, S. Akter, M. Rogers, R. Grene, and S. Li, "Prediction of condition-specific regulatory genes using machine learning," *Nucleic Acids Res.*, vol. 48, no. 11, pp. e62–e62, 2020.

[21] S. Shakya and S. Smys, "Reliable automated software testing through hybrid optimization algorithm," *J. Ubiquitous Comput. Commun. Technol.*, vol. 2, no. 3, pp. 126–135, 2020.

[22] S. R. Mugunthan and T. Vijayakumar, "Design of Improved Version of Sigmoidal Function with Biases for Classification Task in ELM Domain," *J. Soft Comput. Paradig.*, vol. 3, no. 02, pp. 70–82, 2021.

[23] H. Fröhlich, S. Patjoshi, K. Yeghiazaryan, C. Kehrer, W. Kuhn, and O. Golubnitschaja, "Premenopausal breast cancer: potential clinical utility of a multi-omics based machine learning approach for patient stratification," *EPMA J.*, vol. 9, no. 2, pp. 175–186, 2018.

[24] C. Ming, V. Viassolo, N. Probst-Hensch, P. O. Chappuis, I. D. Dinov, and M. C. Katapodi, "Machine learning techniques for personalized breast cancer risk prediction: comparison with the BCRAT and BOADICEA models," *Breast Cancer Res.*, vol. 21, no. 1, pp. 1–11, 2019.

[25] M. F. Ak, "A comparative analysis of breast cancer detection and diagnosis using data visualization and machine learning applications," in *Healthcare*, 2020, vol. 8, no. 2, p. 111.

[26] 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.

[27] 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.*, vol. 7, no. 3, pp. 293–299, 2019.

[28] C.-L. Hsu *et al.*, "Role of fatty liver index and metabolic factors in the prediction of nonalcoholic fatty liver disease in a lean population receiving health checkup," *Clin. Transl. Gastroenterol.*, vol. 10, no. 5, 2019.

[29] S. A. Mohammed, S. Darrab, S. A. Noaman, and G. Saake, "Analysis of breast cancer detection using different machine learning techniques," in *International Conference on Data Mining and Big Data*, 2020, pp. 108–117.

[30] A. Ray, M. Chen, and Y. Gelogo, "Performance Comparison of Different Machine Learning Algorithms for Risk Prediction and Diagnosis of Breast Cancer," in *Smart Technologies in Data Science and Communication*, Springer, 2020, pp. 71–76.

[31] I. Mihaylov, M. Nisheva, and D. Vassilev, "Application of machine learning models for survival prognosis in breast cancer studies," *Information*, vol. 10, no. 3, p. 93, 2019.

[32] M. D. Ganggayah, N. A. Taib, Y. C. Har, P. Lio, and S. K. Dhillon, "Predicting factors for survival of breast cancer patients using machine learning techniques," *BMC Med. Inform. Decis. Mak.*, vol. 19, no. 1, pp. 1–17, 2019.

[33] E. K. Park *et al.*, "Machine learning approaches to radiogenomics of breast cancer using low-dose perfusion computed tomography: predicting prognostic biomarkers and molecular subtypes," *Sci. Rep.*, vol. 9, no. 1, pp. 1–11, 2019.

[34] M. Patrício *et al.*, "Using Resistin, glucose, age and BMI to predict the presence of breast cancer," *BMC Cancer*, vol. 18, no. 1, pp. 1–8, 2018.

[35] M. A. Islam, M. S. Islam, M. S. Hossen, M. U. Emon, M. S. Keya, and A. Habib, "Machine Learning based Image Classification of Papaya Disease Recognition," in *2020 4th International Conference on Electronics, Communication and Aerospace Technology (ICECA)*, 2020, pp. 1353–1360.

[36] P. D. P. Pharoah, A. C. Antoniou, D. F. Easton, and B. A. J. Ponder, "Polygenes, risk prediction, and targeted prevention of breast cancer," *N. Engl. J. Med.*, vol. 358, no. 26, pp. 2796–2803, 2008.

[37] P. D. Hung, T. D. Hanh, and V. T. Diep, "Breast cancer prediction using spark MLlib and ML packages," in *Proceedings of the 2018 5th International Conference on Bioinformatics Research and Applications*, 2018, pp. 52–59.

[38]     J. Iwamoto, T. Takeda, Y. Sato, and H. Matsumoto, "Serum leptin concentration positively correlates with body weight and total fat mass in postmenopausal Japanese women with osteoarthritis of the knee," *Arthritis*, vol. 2011, 2011.

[39]     P. Calabrò *et al.*, "Tissue factor is induced by resistin in human coronary artery endothelial cells by the NF-κB-dependent pathway," *J. Vasc. Res.*, vol. 48, no. 1, pp. 59–66, 2011.

[40]     A. Singh and R. Lakshmiganthan, "Impact of different data types on classifier performance of random forest, naive bayes, and k-nearest neighbors algorithms," 2018.

[41]     M. M. Islam and T. N. Poly, "Machine learning models of breast cancer risk prediction," *BioRxiv*, p. 723304, 2019.