Breast Cancer Prediction using Machine Learning

Arun Sharma[1], Vishal Garg[2], Bhaskar Kapoor[3], Sunil Maggu[4]

[1,2,3,4] *Maharaja Agrasen Institute Of Technology, Delhi, INDIA*

Abstract. Breast cancer poses a serious threat to women, as it is highly morbid and lethal. Doctors find it challenging to develop a treatment plan that could increase patient survival time due to the lack of reliable prognosis models. Therefore, figuring out a method that outcome in the least amount of wrongdoings while improving productivity requires time. In the paper, SVM, Logistic Regression, Random Forest, XGBoost, AdaBoost, k-Nearest Neighbors, Naive Bayes and custom ensemble Classifiers. These algorithms that predict the outcome of breast cancer. The JUPYTER platform is used to conduct all experiments, which are carried out in a simulation environment.

Keywords: Breast Cancer, machine learning, SVM, Logistic Regression, Random Forest, XGBoost, AdaBoost, k-Nearest Neighbors, Naive Bayes.

1 Introduction

The second leading source of mortality for women is breast cancer. In the US in 2016, it is expected that invasive breast cancer in women would result in 246,660 new instances of diagnosis and 40,450 new cases of fatality. Breast cancer accounts for around 12% of all new cancer cases and 25% of all cancers in women. ICTs might be used to cure cancer. Business intelligence (BI) has seen a tremendous transformation, moving from monitoring and decision-making to prediction outcomes thanks to big data, which has emerged as a term associated with data mining, business analytics, and BI. In actuality, big data has improved value extraction from data as well as data size. Data mining techniques, for instance, are rapidly being utilised to research medical science themes because of their high efficacy in outcome prediction, reducing costs for drugs, patient health promotion, boosting healthcare value and quality, and making prompt choices to save lives [1].

There are numerous algorithms for categorizing and predicting the outcomes of breast cancer. In this paper, the performance of various classifiers - SVM, Logistic Regression, Random Forest, XGBoost, AdaBoost, k-Nearest Neighbors, Naive Bayes and custom ensemble Classifiers are compared. These classifiers are among the top 10 data mining algorithms and among the most popular in the research community.

Our goal is to assess these algorithms' accuracy, F1-score, recall and precision and provide them with a custom ensemble classifier to increase their effectiveness and efficiency.

2 Literature Review

In the field of identifying chest harmful development, various evaluations using various considerations and methodologies are used. Different analysts are familiar with different systems and assessments to detect chest pain. We'll look at a few of them here. Barrett and colleagues employed microwave radiometry. To recognize a dangerous chest development in 1977. The warm radiation of the body was measured using a microwave radiometer. They combined microwave and infrared thermographic data, and when they took into account 30 cases, they had a 96% positive disclosure rate. While identifying harmful chest development, the makers examine the errors that mammography missed. Poor radiographic technique and a lack of radiographic disease standards were two major mistakes that were found. In a second review, the endurance examination problem was addressed using ANNs. Results from TheANNs were considered on various datasets that make use of morphometric highlights. The results demonstrate that ANNs were successful in predicting the likelihood of a repeat procedure and isolating patients with poor and excellent forecasts. In 2010, bosom disease was diagnosed using a Computer-Aided Diagnosis (CAD) framework, for instance, ultrasound imaging, which produced more accurate analysis and precision results [2]. A DNA alteration or mutation is one of the causes of breast cancer. When cells start to multiply uncontrollably, cancer develops. Breast cancer cells frequently cluster together to create a lump or an x-ray-visible tumor. The two most prevalent kinds of breast cancer are invasive carcinoma and DCIS. Some are less frequent, such as phyllodes tumors and angiosarcoma. Logistic regression was applied by Wang, D.; Zhang, and Y.H Huang (2018) et al. [8] and resulted in an accuracy of 96.4%. With accuracy of 96.85%, Ak Bugday et al. [9] completed classification on the Breast Cancer Dataset using KNN and SVM. Breast Cancer Prediction and Detection Using Data Mining, by KAYA KELES et al. [10].

3 Preliminaries

A. Support Vector Classifier:

Support Vector Classifier lies in the supervised learning algorithm. It lies in the supervised learning algorithm and are popularly used for classification and regression analysis.
In SVC we provide the labeled data to train the model. Afterwards the data is moved to the testing phase and during the testing phase we perform our trained model with the prediction of new cases.

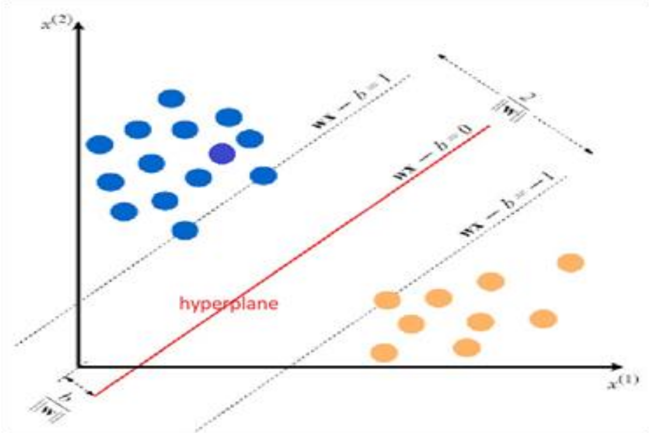


Figure 1. Support Vector Machine

SVC uses a line to divide class A from Class B and this line is commonly called Hyperplane. Hyperplane basically is a kind of boundary between the two classes which decides the new data belongs to which class.

B. Logistic regression:

The Logistic Regression tool forms a model that links a target binary variable—such as a yes/no or pass/fail—to one or more regressors in in order to calculate the estimated chances of each of the target variable's two potential options. Logit, probit, and complements log-log are regularly used logistic regression models. An S-curve can take real values and convert them to values between 0 and 1, but not exactly within those limits.

Y = 1 / (1e^-x)

This sigmoid function is used to convert the independent variable into a expression of probability 0 and 1 with respect to the dependent variable.

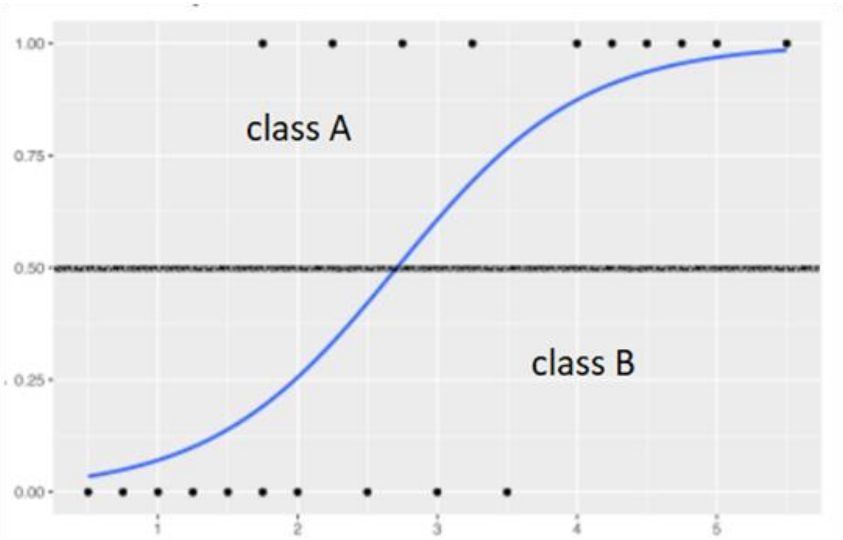


Figure 2. Liner regression

Class A datasets are those above the line going through 0.5, and class B datasets are those below the line. Datasets that pass the 0.5 cutoff are referred to as unclassified.

C. K–Nearest Neighbor Classifier:

The Nearest Neighbor rule regularly outperforms other supervised statistical pattern recognition methods without establishing any presumptions an advance on the distribution of the training data are obtained. Both positive and negative scenarios are included in the training set.

In KNN we use euclidean distance to determine the distance between the nearest training point, and the classification is then done according to the sign of that point .This concept is expanded upon by the k-NN classifier, which assigns the majority sign based on the k nearest points.

When using parametric approaches, whether for classification or density estimation, we presumptively choose a model that is valid over the whole input space. However, in reality, this presumption does not always hold, and if it does not, we might make a significant blunder. One option is to employ a semi-parametric mixing model if we are unable to establish these assumptions and cannot develop a parametric model. [3]

D. Naive Bayes Classifier:

The Naive Bayes classifier tremendously facilitates the process of learning by supposing that qualities are unrelated of class. A Bayesian classifier determines the most likely class for a given sample that is characterized by its feature vector. By assuming that characteristics are not class-dependent, that is, learning such classifiers is possible if a feature vector represents a class made significantly simpler. Despite this irrational presumption, the resulting classifier, known as naive Bayes, is surprisingly effective, frequently competing with considerably more complexed technique.[4]

$$P(A|B) = P(B|A) * P(A) / P(B)$$

-poster probability

E. Decision Tree Classifier:

One of the most well-known machine-learning techniques is the decision tree classifier devised by Quinlan. Decision nodes and child node comprise a decision tree. Each decision node has a set of branches, all of which discusses the output of a test X over a particular characteristic of the input data.

A class that results from a case's decision is represented by each leaf node.It is used for both classification and regression techniques but it is mostly used in classification and tree based data structure is formed.

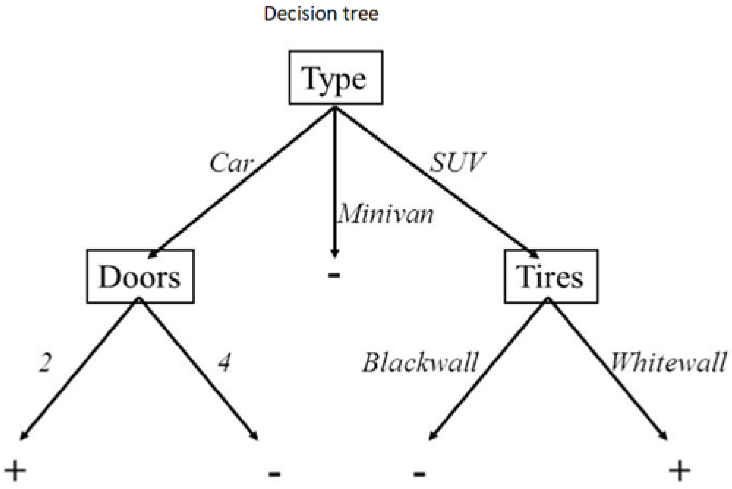


Figure 3. Decision Tree

In the above tree, Type, Doors and Tires are the decision nodes and + and **-** are the leaf nodes.A divide-and-conquer procedure is effectively used for creating a decision tree.[5]

F. Random Forest Classifier:

It is a type of ensemble technique and it uses decision trees in a randomized fashion.

We pick varios samples from the original dataset into the bootstrap dataset. In bootstrap dataset the duplication of sample is allowed but for better results the frequency should be low. Afterwards decision trees are generated randomly from the randomly generated bootstrap data. This randomized manner makes Random forest a much more optimal classifier as compared to decision tree classifier.

The random forest classifier employed in this work grows a tree at each node using randomly chosen characteristics or feature combinations. The training data set was created using a process that involved selecting surrogate N examples of bagging at random for each chosen feature or combination of features.

G. Adaboost Classifier:

The intent of this work is to use RBF SVM as an AdaBoost component classifier. However, how should the s value be set with relation to these RBFSVM component classifiers throughout the AdaBoost repetition? When one s is applied to each RBF SVM component classifier, issues arise.

More specifically, an RBF SVM component classifier that has an excessively large value of s frequently performs poorly. It frequently has classification accuracy below 50% and could not satisfy the AdaBoost provision for a component classifier.[6]

E. XGBoost Classifier:

Extreme Gradient Boosting, or XGBoost, is a type of ensemble technique and it particular use of the Gradient Boosting procedure that utilizes more explicit predictions to choose the greatest tree model. It is most optimal result as compared to other ML Classifiers. It uses a variety of clever techniques that greatly enhance its success, especially with structured data.

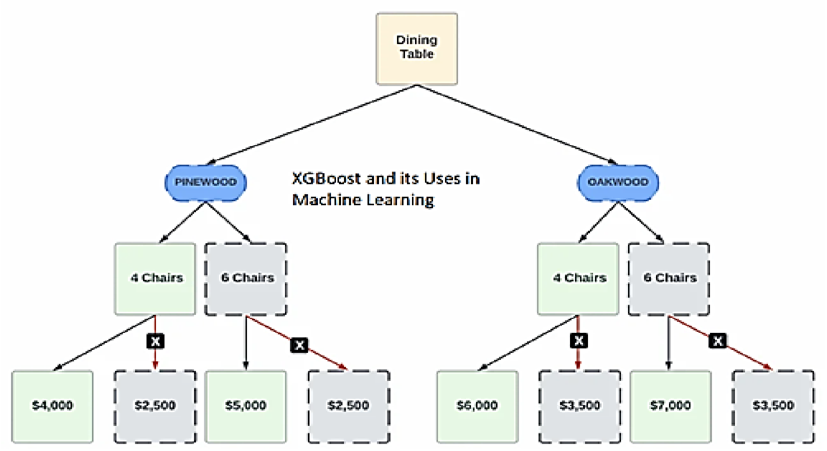
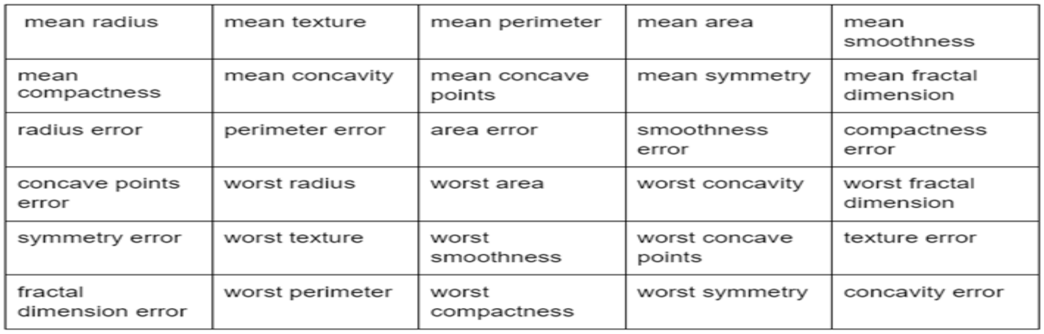


Figure 4. XGBoost Classifier

4 Methodology

4.1 Data Set Description

The breast cancer dataset is part of the UCI machine learning(ML) repository was retrieved. This dataset contains 30 numeric, predictive attributes and the class attributes.This dataset contains 569 instances in which the cases are either benign or malignant. In such cases, 212 (37.25%) of the cases are malignant and 357 (65.90%) are benign. The dataset provided uses a total of thirty features.



mean radius	mean texture	mean perimeter	mean area	mean smoothness
mean compactness	mean concavity	mean concave points	mean symmetry	mean fractal dimension
radius error	perimeter error	area error	smoothness error	compactness error
concave points error	worst radius	worst area	worst concavity	worst fractal dimension
symmetry error	worst texture	worst smoothness	worst concave points	texture error
fractal dimension error	worst perimeter	worst compactness	worst symmetry	concavity error

The provided graph displays the count (Y-axis) of sample users with or without breast cancer. Target 0 denotes benign (patient without breast cancer), while Target 1 denotes benign (patient with breast cancer).

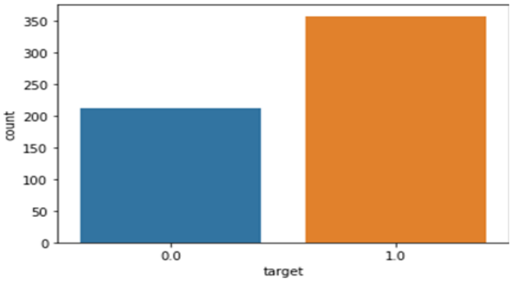


Figure 5. Graph target 0 vs target 1

The below graph shows the pairplot of various columns-

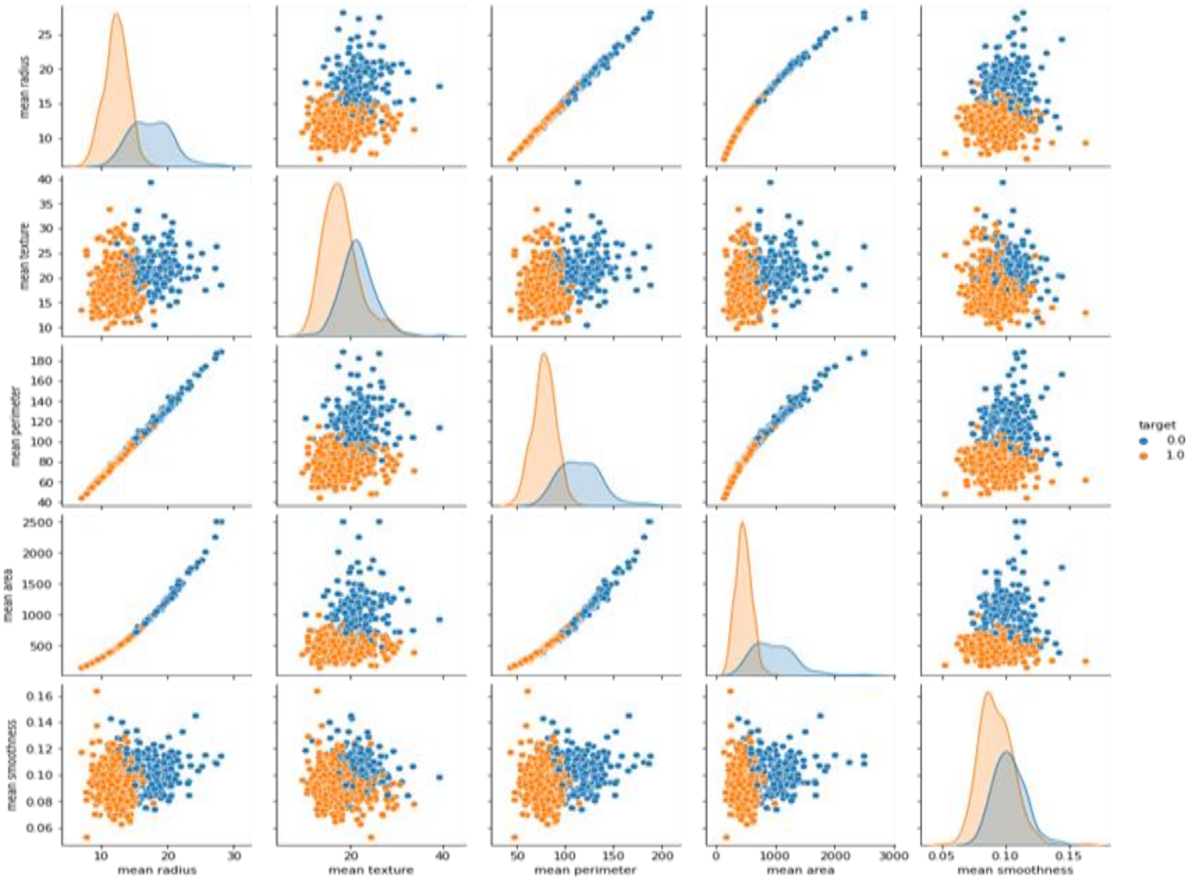


Figure 6. Pair plot of patient with cancer and without cancer

4.2 Training and Testing Phase

The dataset's features are extracted during the training phase, and the suitable model's behavior during the testing phase is then determined. There are two parts to the dataset. These are the phases of testing and training. K fold cross-validation shows that one fold is used for the purpose of testing while k 1 folds are used repeatedly for the purpose of training. Overfitting is prevented by using cross-validation. In our study, data are partitioned using a ten-fold cross-validation technique, eight of which are used for training and the other two for testing in each iteration.

The algorithms used for the analysis are listed below-

- SVM
- Logistic Regression
- Random Forest
- XGBoost
- AdaBoost
- k-Nearest Neighbors
- Naive Bayes

The three best algorithms are selected for ensemble training after a thorough analysis of the algorithms based upon accuracies and precision are XGBoost, AdaBoost and Random Forest.

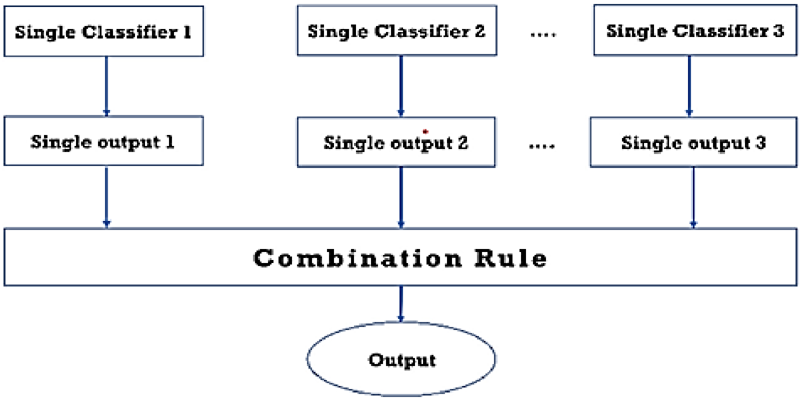


Figure 7. Ensemble classifier using XGBoost, Adaboost and Random Forest

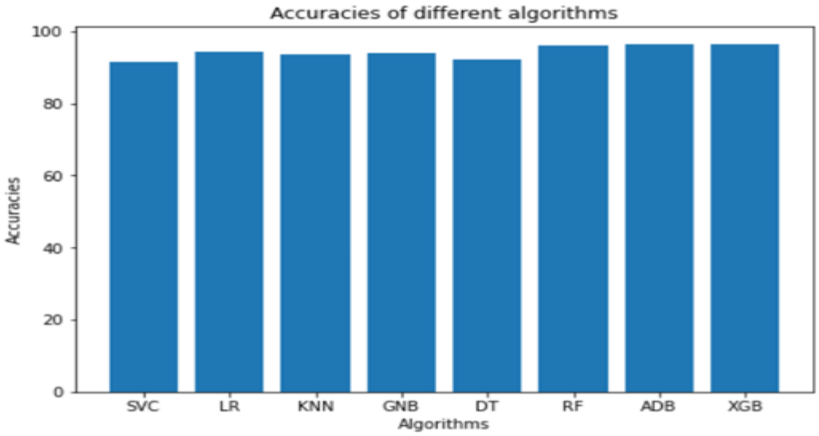
Multiple machine learning models are used in ensemble learning in an effort to improve predictions on a dataset. A dataset is used to train a variety of models, and the individual predictions made by each model form the basis of an ensemble model. The ensemble model then combines the outcomes of different models' predictions to produce the final outcome.

5 Result

Table 1 compare various algorithms based upon accuracy, F1, Recall and Precision.

S.No	Algorithm	Accuracy (%)	F1 Score(%)	Recall(%)	Precision (%)
1	Logistic Regression	94.43	95.42	94.79	96.47
2	Random Forest	96.25	97.27	96.87	97.82
3	XG Boost	96.60	98.18	97.91	98.52
4	Support vector classifier	91.50	93.54	92.70	95.20
5	Decision Tree	92.33	95.48	95.35	95.63

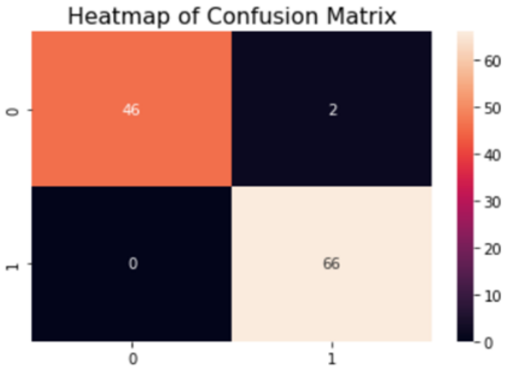
6	ADA Boost	96.42	97.29	97.15	97.44
7	KNN classifier	93.44	93.54	92.70	95.20
8	GaussianNB	93.85	94.56	94.31	94.88
9	Ensemble	96.54	98.18	97.91	98.52



The above graph representation shows the Heatmap of confusion matrix.

Confusion Matrix

Where model errors in categorization issues occurred is shown in the above table. The rows display the actual classes for which the responses were expected. The columns, however, display our forecasts. With the help of this table, it is easy to determine which projections are inaccurate. [7]



6 Conclusion

All supervised classification algorithms were trained to increase accuracy, but you may test a couple that are always in demand. After training various algorithms, we discovered that XGBoost, Random Forest, and Logistic Regression classifiers provided more accuracy than the remaining methods, and after ensembling these methods the accuracy(96.54%), precision(98.18%), F1 score(98.18%) and Recall(97.91%) is increased tremendously.

7 Future Works

The results analysis shows that the combination of n-dimensional data with different selection for features, classification, and dimensionality reduction techniques may offer better tools for inference in this field. It is necessary to conduct additional study in this area to improve the performance of classification algorithms and enable them to make predictions on a wider range of factors. In order to attain high accuracy, we plan to parametrize our categorization systems. We are investigating a variety of datasets and the potential applications of ML Classifiers and techniques to further diagnose breast cancer. We want to maximize accuracy while lowering error rates.

References

1. Asri, Hiba, et al. "Using machine learning algorithms for breast cancer risk prediction and diagnosis." Procedia Computer Science 83 (2016): 1064-1069.
2. Sharma, Shubham, Archit Aggarwal, and Tanupriya Choudhury. "Breast cancer detection using machine learning algorithms." *2018 International conference on computational techniques, electronics and mechanical systems (CTEMS)*. IEEE, 2018.
3. Islam, Mohammed J., et al. "Investigating the performance of naive-bayes classifiers and k-nearest neighbor classifiers." *2007 international conference on convergence information technology (ICCIT 2007)*. IEEE, 2007.
4. Rish, Irina. "An empirical study of the naive Bayes classifier." *IJCAI 2001 workshop on empirical methods in artificial intelligence*. Vol. 3. No. 22. 2001.
5. Stein, Gary, et al. "Decision tree classifier for network intrusion detection with GA-based feature selection." *Proceedings of the 43rd annual Southeast regional conference-Volume 2*. 2005.
6. Li, Xuchun, Lei Wang, and Eric Sung. "AdaBoost with SVM-based component classifiers." *Engineering Applications of Artificial Intelligence* 21.5 (2008): 785-795.
7. Townsend, James T. "Theoretical analysis of an alphabetic confusion matrix." *Perception & Psychophysics* 9.1 (1971): 40-50.
8. Wang, D. Zhang and Y. H. Huang “Breast Cancer Prediction Using Machine Learning” (2018), Vol. 66, NO. 7.
9. B. Akbugday, "Classification of Breast Cancer Data Using Machine Learning Algorithms," 2019 Medical Technologies Congress (TIPTEKNO), Izmir, Turkey, 2019, pp. 1-4.
10. Keles, M. Kaya, "Breast Cancer Prediction and Detection Using Data Mining Classification Algorithms: A Comparative Study." Tehnicki Vjesnik - Technical Gazette, vol. 26, no. 1, 2019, p. 149+.