International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 10s
DOI: https://doi.org/10.17762/ijritcc.v11i10s.7688
***Article Received:** 05 June 2023 **Revised:** 22 July 2023 **Accepted:** 12 August 2023*

Comparision of Different Classifiers for Prediction of Breast Cancer

Tintu P B[1], Dr. S Manju Priya[2]
[1]Research Scholar
Department of Computer Science,
Karpagam Academy of Higher Education Coimbatore,
Tamil Nadu, India
tintupadikkal@gmail.com
[2]Professor
Department of Computer Science,
Karpagam Academy of Higher Education, Coimbatore, Tamil Nadu, India
smanjupr@gmail.com

Abstract – The cell formed in the breast are known as breast cancer. It occurs mainly in women and it may occur rarely in men also. It is considered as the most common ailment that can lead to large number of death in females every year. In spite of the factuality that cancer is treatable and can be relieve if treated at its early stages; many patients are screened for cancer only at a very late stage. Data mining technique such as classifications provides an efficient technique to classify data, where these methods are commonly used for diagnostic decision making. The Machine learning techniques propound various methods such as statistical and probabilistic methods which allow system to learn from past experiences to distinguish and identify patterns from a standard dataset. The research work presents a review of machine learning techniques which can be used in breast cancer disease detection by applying algorithms on breast cancer Wisconsin data set. Algorithms such as Navies Bayes, Random Forest, Support Vector Machine, Adaboost and Decision Trees were used. The result outcome shows that Random Forest performs better than other techniques.

Keywords- Data Mining algorithm, Breast Cancer, Wisconsin Breast Cancer (WDBC), Classification, Navies Bayes, Random Forest, Support Vector Machine, Adaboost and Decision Trees.

I. INTRODUCTION TO BREAST CANCER

Among various types of cancers breast cancer is considered as the second largest aspect of demise in women. Tumor which grows in breast is classified as malignant or benign. At a starting stage, if the cancer is diagnosed as benign, it can be treated and cured. It is believed to be one of the most frequent types of cancer in women, causing death of about 1.5 million people globally [1]. Presently , in the United States of America, out of eight women one is at the risk for breast cancer, with the highest ratio in the age group of 35 and 50[2]. Examination of the latest information's reveals that after the discovery of disease can increase the survival rate to 88%. if diagnosed within six years the survival rate is decreased to 80.3% [3]. Australia, the United States and Europe (Western countries) have confirmed a high proportion of breast cancer patients. In certain countries, the proportion has increased in the 20th century, leading to an increase in mammography and biopsy having a significant impact on general changes in reproductive patterns .. The formatter will need to create these components, incorporating the applicable criteria that follow. Information and Communication Technology plays an important role in cancer treatment. For example, data mining technologies are for reducing drug rates, generating real-time decisions in people's lives ,predicting outcomes, supporting patient health, , and performing well in developing customized quality of care. The data is generated in large volumes in the field of biomedicine is extremely fast [4]. The data which is generated is rich in source. Information and knowledge discovery can be performed with the help of Machine learning techniques. The data which is available can be noisy data which requires preprocessing. The data set is taken from UCI repository, Wisconsin breast cancer diagnostic dataset which has different types of breast cancer dataset [5]. The dataset consist of 569 instances out of which malignant cases are 212 and benign cases are 357. In this data set there is no missing attribute value. The attributes in the wdbc dataset are as follows: number for ID, M or B for diagnosis. For each cell nucleus ten real value features were calculated like radius, concativity, perimeter, texture, compactness, area, smoothness, fractal dimensions, concave points, and symmetry. From each images 30 features were resulted based on computation of features like the standard error, mean, worst or largest. Here field 3 represents mean radius; field13 calculated as Radius SE

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 10s
DOI: https://doi.org/10.17762/ijritcc.v11i10s.7688
***Article Received:** 05 June 2023 **Revised:** 22 July 2023 **Accepted:** 12 August 2023*

and field 23 is represented as worst radius. All the values for the feature were recorded as significant four digits.

II. LITERATURE REVIEW

Researchers have realized the importance about fighting against breast cancer and also to explain machines learning techniques for analyzing performance for diagnosis of breast cancer .Mrs. M. Sathya and Dr .S.Manju Priya[6] proposed , a method for selecting feature genes using modified whale optimized algorithm from the microarray dataset. Using the concept helps to select relevant feature genes which can be used for dimensionality reduction of the dataset and also identify the target genes which are responsible for causing cancer. The performance of the proposed method was analyzed and studied by using microarray dataset for different cancer. Authors Hiba , Hajar, Hassan and Thomas [7] analyzed six different algorithms of machines learning which were used in classification of breast cancer into two categories namely malignant ore begin. Weka was used as a tool for experiments in Wisconsin breast cancer original data set .On analysis the classification accuracy was improved by using feature selection techniques .Sunita Soni , Shweta Kharya [8] they aimed to improve Naive Bayes classifier performance for which they proposed a method which incorporates weighted concept to traditional Naive Bayes. They used dataset for breast cancer from machine learning repository UCI. Researchers Bharat and Madhuri [9] examined different algorithms like Random forest, Linear regression, Multi layer Perceptron and Decision tree algorithm and these algorithms were used to improve the accuracy of prediction on breast cancer data. Through analysis Multilayer perceptron has greater performs when compared to other algorithms. Rung Ching Chen, Bin Dai, Wei-Wei Zhang and Shun-Zhi Zhu [10] used breast cancer diagnostic dataset where the combination algorithms of random forest were tested to enhance the accuracy of breast cancer classification. Zehra Gromilic, Sabina Halilovic ,Ahmed Osmanovic, Layla Abdel Ilah and Adnan Fojnica[11] in their experiment they seek to decrease the number of camouflaged cases, data was collected from UCI machine learning repository, and different tests were performed on expert diagnostic systems on data. The experiment result shows that with 20 neurons single hidden layer neural network with Feed Forward Back propagation and transfer of neuron using TANSIG gives 98.9% higher accuracy in training set and for testing set it gave 99% accuracy. Authors Resul Kara, Zehra Karapinar Senturk [12] in their research seven different machines learning algorithms for predicting breast cancer in new cases were studied, Rapid Miner 5.0 was used as tool dataset which was taken from UCI repository. Hongwei Chen ,Wenbin, Yue , Annette Payne ,Zidong Wang and Xiaohui Liu [13] in their review paper explained several algorithms of machines learning first and the used wisconsin breast cancer data set to provide their applications. They concluded that to improve classification and accuracy different machines learning techniques could be used. Amrane Saliha, Oukid Ikram Gagaoua Tolga , Ensar and Meriem Amrane [14]in their work they performed a contrast between naive bayes classifier and KNN. The comparison performed shows that K nearest neighbor gives 97.5% accurate result. Halilovic, Ahmed Osmanovic et all. [15]. They aimed to decrease the number of cases which are misdiagnosed in their research, the used dataset from UCI machine learning respiratory for testing different diagnostic systems. The Feed Forward Back propagation with single hidden layer network with twenty neurons and transfer using TANSIG gives the higher accuracy with 98.9% accuracy in training set and 99% accuracy in test set. Resul Kara and Zehra Karapinar Senturk[16] in their research work they studied seven different machines learning algorithms which was used to predict breast cancer for news cases. RapidMiner 5.0 was used as tool and data set was taken from UCI Machine Learning Repository. H. Chen ,Wenbin Yue et all. [17] in their paper they presented a review. First they gave detailed explanations of several machines learning algorithms and then they passed their applications in Wisconsin Breast Cancer Data Set. There conclusion was that machines learning techniques present a remarkable improvement in classification and accuracy prediction. T. Ensari ,Meriem Amrane et all.[18] in their research work they did a comparison between Naive Bayes classifier and K Nearest neighbor. The result shows that and K Nearest neighbor give more accurate result with the accuracy of 97.51%.

III. DIFFERENT CLASSIFIERS USED

Data Mining in medical field is a critical area of study which is used to show patterns from facts which became used to expand portending replications. Information is stored as digital order in medical organizations. The information contains details about all the patients whose details are provided by fitness care providers. By the usage of such conventional methods, it is difficult to extract significant facts. Data mining are utilized in situations where a big series of healthcare information is used for analysis. Data mining equipment assists us to find out facts from unknown styles. Techniques like classification, clustering and prediction are a few strategies that are used on scientific facts. This overview paper affords Data Mining strategies within side the area of scientific region [19]. In the field of medical diagnosis, Breast cancer is the most notable disease which is increasing every year. A comparative analysis of five widely used machine learning techniques is performed on Wisconsin Breast Cancer Dataset to predict the breast cancer: Decision Tree, Support vector machine, Navies Bayes, Ada boost and Random Forest.

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 10s
DOI: https://doi.org/10.17762/ijritcc.v11i10s.7688
***Article Received:** 05 June 2023 **Revised:** 22 July 2023 **Accepted:** 12 August 2023*

A. Naive Bayes

Naive Bayes algorithm is primarily based on Bayes theorem which gives an assumption of independence amongst different predictors. A Naive Bayes algorithm assumes how a selected characteristic of a category present is unreliable to the characteristic presence. Naive Bayesian are statistical classifiers , which can predict class membership probabilities such that a given sample will belong to a particular case. The Naive Classifiers assumes that the effect of an given class attribute value is independent of the values of other attributes. Advantage of the Bayes classifier is that it requires only a small amount of data for training to estimate the parameters such as means and variances of the variables which are necessary for classification. Naive Bayes performs better in many real world situations like Medical Diagnosis, Spam Classification and Weather forecasting. It is used when dimensionality of input is high. [20]

B. Support Vector Machine

The support vector machine also known as SVM.which is a supervised classification models, that is mainly applied in the areas of cancer diagnosis [21]. The algorithm assembles samples from all groups which are critical and these samples are known as support vectors These support vectors distributes the classes by creating a linear function. Algorithm can be applied to map high dimensionality space between the input vectors. SVM aims to find the various suitable hyper planes which separate the data set into different classes [22]. The aim of Linear classifier is to maximize the distance between the nearest data point and decision hyper plane which is called the limited distance . SVM is preferred [23] due to its performance in classification performance. In n- dimensional space every data item are plotted as coordinates where n is the total number features being used for classification and the value of each feature is represented by the coordinates of the data point. A decision hyper plane is used as data points of diverse classes using maximum margin. In the hyper plane data points which lie closer are called as Support Vectors.Maintaining the Integrity of the Specifications

C. Random Forest

The set of rules for random selection forests was created in 1995 by TinKam Ho. The usage of the random subspace technique in Ho's formulation is a technique proposed through means of Eugene Kleinberg a random wooded area classifier Meta estimator which suits some of selection tree classifiers on numerous sub-samples. Predicting accuracy and managing of over fitting makes use of averaging method to enhance the prediction [24]. The process of combining multiple classifiers for solving complex problems is based on concept of ensembling technique. It is used to improve the model performance. Random forest combines multiple trees for predicting class from dataset. Some decision trees may predict the correct output and while others may not. But together, all the trees predict the correct output. Two assumptions for a better Random forest classifier are the predict accurate result than guessed result and they must have low correlations from every tree

D. Decision tree J48

Decision tree[25] is a supervised machine learning algorithm which can be used for both regression and classification.The main concept behind Decision tree is divide and conquer methodology which has two partitions namely numerical and Nominal partition. Depending on attribute value the can be used for creating pruned or unpruned tree from training dataset which are labeled. In decision tree algorithm, the dataset can be split into very small subsets. And each one of the attribute of data set could use splitting for decision making. When all subset belong to same or single class the splitting stops. The J48 algorithm creates the node for decision in tree by conjecturing the contemplate value of the class [26]. C4.5 decision tree algorithm [27] is mainly used for experimental analysis. In this method a tree can have categorical as well as discrete attributes. The order in which attributes will be placed . Accuracy of each rule generated is estimated to determine the order in which attribute will be placed in the decision tree. The most common approach is to minimize the average error of every node.

E. Ada Boost

The adaptive boosting algorithm also known as AdaBoost was proposed by Yoav Freund and Robert Shapire for generating strong classifier from a set of weak classifiers[28] . The algorithm maintains collection of weights over training data and it is adjusted after each weak cycle.The weights which are misclassified by current weak learner will be increased and the weights of the samples which are correctly classified will be decreased. AdaBoost is one of the most fast convergence ,promising and easy to be implemented machine learning algorithm. No prior knowledge is required about the weak learner and it can be easily combined with other methods, such as support vector machine. some of the applications are Face Detection and Facial Expression Recognition.It is a meta estimator classifier which places a classifier on the actual data set and then it fits additional copies of the classifier on the same dataset . This subsequent classifiers focus on difficult cases and locale the weights of faultily classified instances can be adjusted [29].

IV. PROPOSED WORK

The proposed work uses different machine learning algorithm for comparison of Random forest , Ad boost ,Naïve Bayes,

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 10s
DOI: https://doi.org/10.17762/ijritcc.v11i10s.7688
***Article Received:** 05 June 2023 **Revised:** 22 July 2023 **Accepted:** 12 August 2023*

Decision Tree and SVM classifiers. These classifiers were applied on WBDC dataset which consists of 569 instances which was obtained from Wisconsin dataset from UCI repository. The classification was performed in Google Colab and the following result was obtained. Result is formulated in table 1.

TABLE 1 : COMPARISON OF EXECUTION TIME AND CLASSIFICATION ACCURACY

SNO	CLASSIFIER	EXECUTION TIME in seconds	CLASSIFICATION ACCURACY%
1	Naive Bayes	0.01	84%
2	SVM	0.01	88%
3	Random Forest	0.50	92%
4	Decision tree J48	0.01	84%
5	AdaBoost	0.12	84%

A. Data set

Wisconsin Breast Cancer Diagnostic Data Set is used for the work [30]'. This dataset is taken from Kaggle repository which consists of 569 records and 32 attributes. The features for this data set are taken from tissues in breast of patients using Fine Needle Aspiration. In FNA model a thin needle is injected into the abnormal appearing body tissues and sample is collected to make a diagnosis or prediction of disease such as cancerous or non cancerous. This dataset contain 2 classes one is malignant which is a cancerous tumor and other is non-cancerous tumor called benign. There are no missing values in this dataset. It has a class distribution of 212 malignant cells and 357 benign cells. The dataset contains ten major real valued features which are computed for each cell nucleus. Such as Radius, Perimeter, Texture, Compactness, Area , Smoothness , Concavity , Symmetry, Concave points, Fractal dimension . For each image features are calculated. The mean, standard error, and worst are calculated for all these 10 features, which results in 30 features. The Maligant and Benign class distribution is shown in the figure below with features such as texture mean and radius mean.

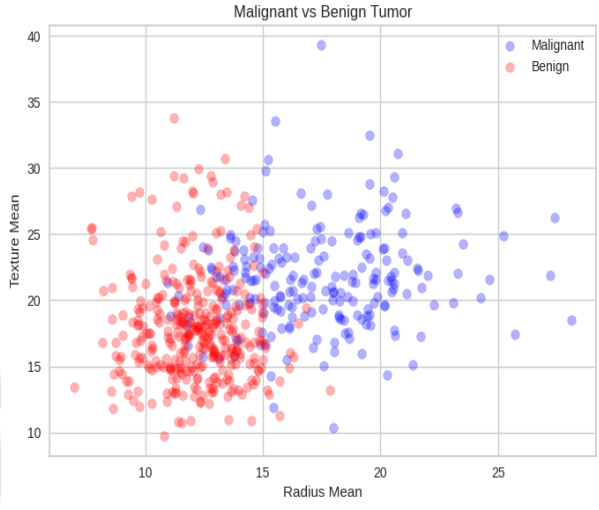

Figure 1. Malignant and Benign class distribution

V. RESULTS AND FIGURES

In this figure1 five different classifiers are used. For each classifier accuracy value is shown in the chart. Random Forest has the highest accuracy of 92%

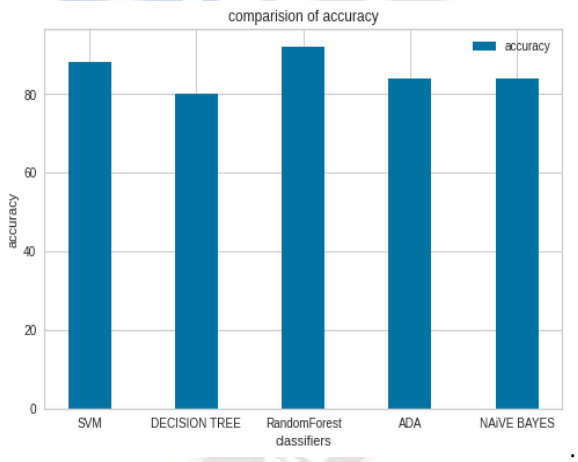

Figure 2. Different Classification algorithms

In this figure2 for each classifier execution time is represented.Minimum execution time is taken by Navie Bayes,SVM, Decision Tree and highest by Random Forest.

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 10s
DOI: https://doi.org/10.17762/ijritcc.v11i10s.7688
***Article Received:** 05 June 2023 **Revised:** 22 July 2023 **Accepted:** 12 August 2023*

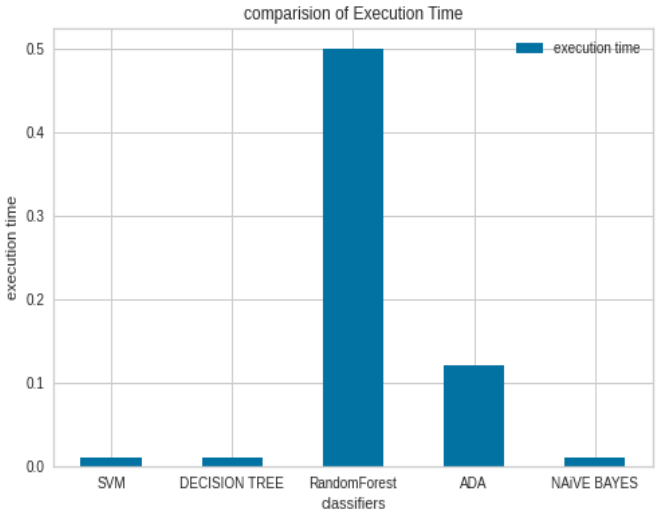

Figure 3. Comparison of Execution Time

TABLE 2: PERFORMANCE MEASURE OF DIFFERENT ALGORITHM

Classification Algorithms	Performance Measures	
SVM	Accuracy	0.88
	Recall	0.88
	Precision	0.87
	F1 measure	0.88
Decision Tree Classifier	Accuracy	0.80
	Recall	0.78
	Precision	0.80
	F1 measure	0.79
Naïve Bayes Classifier	Accuracy	0.84
	Recall	0.83
	Precision	0.83
	F1 measure	0.83
Random Forest Classifier	Accuracy	0.92
	Recall	0.92
	Precision	0.92
	F1 measure	0.92
Ada Boost	Accuracy	0.84
	Recall	0.83
	Precision	0.83
	F1 measure	0.83

VI. CONCLUSION

Breast cancer is the growth of tumors, if it detected in early stage, it can save lives of women and men. By using Machine learning algorithms ,we can classify and predict breast cancer as Benign or Maligant.From the analyze , using WDBC datasets, it is noted that out of five classifiers randomly chosen after classification Random Forest has 92% accuracy. Further works are carried out to provide accurate results in diagnosing breast cancer.

REFERENCES

[1] Lyon IAfRoC, World Cancer Report, International Agency for Research on Cancer Press 2003:188-193.
[2] Breast Cancer Facts & Figures, American Cancer Society-2007
[3] Daniel F. Roses,' Clinical Assessment of Breast Cancer and Benign Breast Disease', In: Breast Cancer: Vol. 2, Ch.14(2005)
[4] Marx, Vivien. "Biology: The big challenges of big data." Nature 498.7453 (2013): 255-260.
[5] UCI Machine Learning Repository and breast cancer site:archive.ics.uci.edu/ml site: ics.uci.edu.
[6] Sathya,Dr.S.ManjuPriya, 'ModifiedWhale Optimization Algorithm For Feature Selection In Micro Array Cancer
[7] Dataset', international journal of scientific & technology research volume 9, issue 03, march 2020
[8] Hiba Asria ,Hajar Mousannifb ,Hassan Al Moatassime c ,Thomas Noël.' Using Machine Learning Algorithms for Breast Cancer Risk Prediction and Diagnosis', Procedia Computer Science 83 (2016) 1064 – 1069
[9] Sunita Soni , Shweta Kharya ,'Weighted Naive Bayes Classifier: A Predictive Model for Breast Cancer Detection', International Journal of Computer Applications (0975 – 8887) Volume 133 – No.9, January 2016
[10] Bharat and Madhuri Gupta ,'A Comparative Study of Breast Cancer Diagnosis Using Supervised Machine Learning Techniques', Proceedings of the Second International Conference on Computing Methodologies and Communication (ICCMC 2018) IEEE Conference Record # 42656; IEEE Xplore ISBN:978-1-5386-3452-3
[11] Zhu, H. Zou, S. Rosset, T. Hastie, 'Multi-class AdaBoost', 2009.
[12] H. Saoud, A. Ghadi, M. Ghailani, and B. Anouar Abdelhakim, 'Application of Data Mining Classification Algorithms for Breast Cancer Diagnosis'. 2018
[13] S. Kharya and S. Soni, 'Weighted Naive Bayes Classifier: A Predictive Model for Breast Cancer Detection', Int. J. Comput. Appl., vol. 133, no. 9, pp. 32–37, Jan. 2016.
[14] Wenbin Yue , Zidong Wang ,Hongwei Chen, Annette Payn and Xiaohui ,' Machine Learning with Applications in Breast Cancer Diagnosis and Prognosis', Designs, 2(2), 13.
[15] G. Rajasekaran, & P. Shanmugapriya. (2023). Hybrid Deep Learning and Optimization Algorithm for Breast Cancer Prediction Using Data Mining. International Journal of Intelligent Systems and Applications in Engineering, 11(1s), 14–22. Retrieved from https://ijisae.org/index.php/IJISAE/article/view/2472
[16] Osmanovic, S. Halilovic, L. A. Ilah, A. Fojnica, and Z. Gromilic, 'Machine Learning Techniques for Classification of Breast Cancer', in World Congress on Medical Physics and Biomedical Engineering 2018, vol. 68/1, pp. 197–200.
[17] H. Saoud, A. Ghadi, M. Ghailani, and B. Anouar Abdelhakim, 'Using Feature Selection Techniques to Improve the Accuracy of Breast Cancer Classification: Special Issue on Data and Security Engineering', 2019, pp. 307–315.
[18] Mr. Ather Parvez Abdul Khalil. (2012). Healthcare System through Wireless Body Area Networks (WBAN) using Telosb Motes. International Journal of New Practices in Management

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 10s
DOI: https://doi.org/10.17762/ijritcc.v11i10s.7688
***Article Received:** 05 June 2023 **Revised:** 22 July 2023 **Accepted:** 12 August 2023*

and Engineering, 1(02), 01 - 07. Retrieved from http://ijnpme.org/index.php/IJNPME/article/view/4

[19] Z. K. Senturk and R. Kara, 'Breast Cancer Diagnosis via Data Mining: Performance Analysis of Seven Different Algorithms', Comput. Sci. Eng. Int. J., vol. 4, no. 1, pp. 35–46, Feb. 2014.

[20] W. Yue, Z. Wang, H. Chen, A. Payne, and X. Liu, 'Machine Learning with Applications in Breast Cancer Diagnosis and Prognosis', Designs, vol. 2, no. 2, p. 13, May 2018.

[21] 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), Istanbul, 2018, pp. 1–4.

[22] Brown, R., Brown, J., Rodriguez, C., Garcia, J., & Herrera, J. Predictive Analytics for Effective Resource Allocation in Engineering Education. Kuwait Journal of Machine Learning, 1(1). Retrieved from http://kuwaitjournals.com/index.php/kjml/article/view/91

[23] Witten, Ian H., Eibe Frank, Mark A. Hall, and Christopher J. Pal.,' Data Mining: Practical machine learning tools and techniques', 2016.

[24] J. Friedman, L. Breiman, R. Olshen and C. Stone,' Classification and Regression TreesWadsworth, Belmont, CA (2014).

[25] B. Dai, R.-C. Chen, S.-Z. Zhu and W.-W. Zhang, 'Using Random Forest Algorithm for Breast Cancer Diagnosis', in 2018 International Symposium on Computer, Consumer and Control (IS3C), Taichung, Taiwan, 2018, pp. 449–452.

[26] Y. Freund, R. Schapiro, 'A Decision-Theoretic Generalization of on-Line Learning and an Application to Boosting', 1995.

[27] Morzelona, R. (2021). Human Visual System Quality Assessment in The Images Using the IQA Model Integrated with Automated Machine Learning Model . Machine Learning Applications in Engineering Education and Management, 1(1), 13–18. Retrieved from http://yashikajournals.com/index.php/mlaeem/article/view/5

[28] Abreu, Pedro Henrique's, et al. "Predicting Breast Cancer Recurrence Using Machine Learning Techniques: A Systematic Review." ACM Computing Surveys (CSUR) 49.3 (2016): 52.

[29] B. Dai, R.-C. Chen, S.-Z. Zhu and W.-W. Zhang, 'Using Random Forest Algorithm for Breast Cancer Diagnosis', in 2018 International Symposium on Computer, Consumer and Control (IS3C), Taichung, Taiwan, 2018, pp. 449–452.

[30] Yan-yan, Ying LU2,'Decision tree methods: applications for classification and prediction', Shanghai Archives of Psychiatry, 2015, Vol. 27, No. 2

[31] http://scikit-learn.org/stable/modules/tree.html (accessed November, 2016)

[32] Patel Brijain, R., and Kaushik K. Rana. "A survey on decision tree algorithm for classification." International Journal of Engineering Development and Research. Vol. 2. No. 1, (2014), IJEDR.

[33] Ana Silva, Deep Learning Approaches for Computer Vision in Autonomous Vehicles , Machine Learning Applications Conference Proceedings, Vol 1 2021.

[34] Zhu, H. Zou, S. Rosset, T. Hastie, 'Multi-class AdaBoost', 2009.

[35] Ruihu Wang, 'AdaBoost for Feature Selection, Classification and Its Relation with SVM , A Review', 2012 International Conference on Solid State Devices and Materials Science.

[36] https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin +(D iagnostic) (Accessed 15 Oct 2017)