Scope
Volume 14 Number 04 December 2024

Improving Breast Cancer Detection through Advanced Machine Learning Techniques

Daizy Deb[1], Ritam Rajak[2], Soumyadeep Sil[3], Avijit kumar Chaudhuri[4], Nilav Darsan Mukhopadhyay[5]
[1,5]Assistant Professor, Brainware University, Barasat
[2,3] Master of Technology Brainware University, Barasat
[4]Associate Professor, Brainware University, Barasat, India

Corresponding Author: **Ritam Rajak**

Abstract
When it comes to the question of female death throughout the world, breast cancer plays a significant role due to which the female mortality rate is high. So, to address this crucial question effective methods are necessary to diagnose breast cancer in an early stage to do proper treatment. Here comes the application of emerging technologies in the form of machine learning. Machine learning shows promising significance in breast cancer prediction. To address these capabilities of machine learning a thorough focused research is needed. So, in this paper, we used four machine learning algorithms named Random Forest, Random Tree, J48, and Multilayer Perceptron. We applied those algorithms to the well-known Wisconsin Diagnostic dataset on Breast Cancer. After applying feature selection techniques named Information Gain, Gain Ratio, ReliefF, and OneR we used machine learning algorithms mentioned above with 10-fold cross-validation to the given dataset. Thus, we got only 8 features which are significant out of 32 features present in the original dataset to predict breast cancer's presence in the human body. We also try to achieve high consistency, sensitivity, and specificity levels by exploring popular ensemble approaches of algorithms in machine learning. By writing this paper we want to establish a comprehensive framework to guide breast cancer prediction using decision-making trees for the benefit of humans.

Keywords: Breast Cancer, Random Forest, Random Tree, J48, Multilayer Perceptron, Machine learning.

1. Introduction and Literature Review

As per the report of the International Agency for Research on Cancer (IARC) in December 2020, lung cancer has been overtaken by Breast Cancer in terms of how frequently it is diagnosed in the female human body [1]. It is a matter of concern when

we notice that the number of cancer cells in the human body after proper diagnosis is 19.3 million in the year 2020 which is almost double of 10 million observed in the year 2000 [2]. As per the data, it is noted that in every five individuals, one has to develop cancer at any moment of their life. The forecast for cancer diagnosis rates in the human body in the future is much scarier because the cancer diagnoses will continue to rise to 50% by 2040 [3]. The surge in terms of cancer-related death numbers is also a matter of thought as it goes up to 10 million in 2020 from 6.2 million in 2000 [4].

To modernize the healthcare system, we have to incorporate information and communication technologies (ICT). Big Data took a revolutionary role in the organization of vast amounts of unstructured, diverse, non-standard, and incomplete medical data [5]. Prediction of trends and decision-making processes can be supported by this data through training and testing of capable machine learning models. This type of machine learning model can reduce the overall treatment cost and make the healthcare diagnostic system more effective.

When it comes to the question of what data mining can do in the field of healthcare, the answer is that it can help in the classification of diseases and also predict the disease [6]. In the case of the healthcare sector, classification can help by sorting data based on predefined categories, and prediction can assist in the determination of future trends of disease based on historical data [7]. Different machine learning algorithms show their capabilities in the prediction of Breast Cancer in its early stage [8]. Among those machine learning algorithms, four algorithms named Random Tree, Random Forest, J48, and Multilayer Perceptron (MLP) performance are compared in this paper [9][10].

In this study, we deployed a 10-fold cross-validation method to split the whole dataset into 10 subsets among which nine are used for training the model and one is for testing the model in rotation [11]. Through this approach, the enhancement of the reliability of the machine learning model's performance is evaluated. After analyzing the data in terms of confusion matrix matrices, the authors of this paper reduced the number of significant features from 32 to 8 only [12]. The models' assessment is done by using various data splits (i.e., 66-34, 50-50, and 80-20) [13]. After this assessment using the reduced 8 significant features, all algorithms achieve over 90% accuracy except Random Forest [14]. Nearly 100% accuracy is achieved by J48, MLP, and Random Tree algorithms in both 80-20 train-test split and 10-fold cross-validation [15][16].

The authors of this paper ranked the features of the original dataset by using four renowned feature selection algorithms named Information Gain, Gain Ratio, ReliefF, and OneR and measured the performances of the models [17]. To determine the most significant features or data points, these techniques play a very important role [18]. These classifiers' performance is compared, followed by the analysis in this comprehensive study by the authors of this paper [19]. The main objective of this

particular research is to find out the most effective machine learning-based classification algorithm to predict the presence of breast cancer in a human body based on the classifiers' accuracy, precision, F1 score, and sensitivity [20].

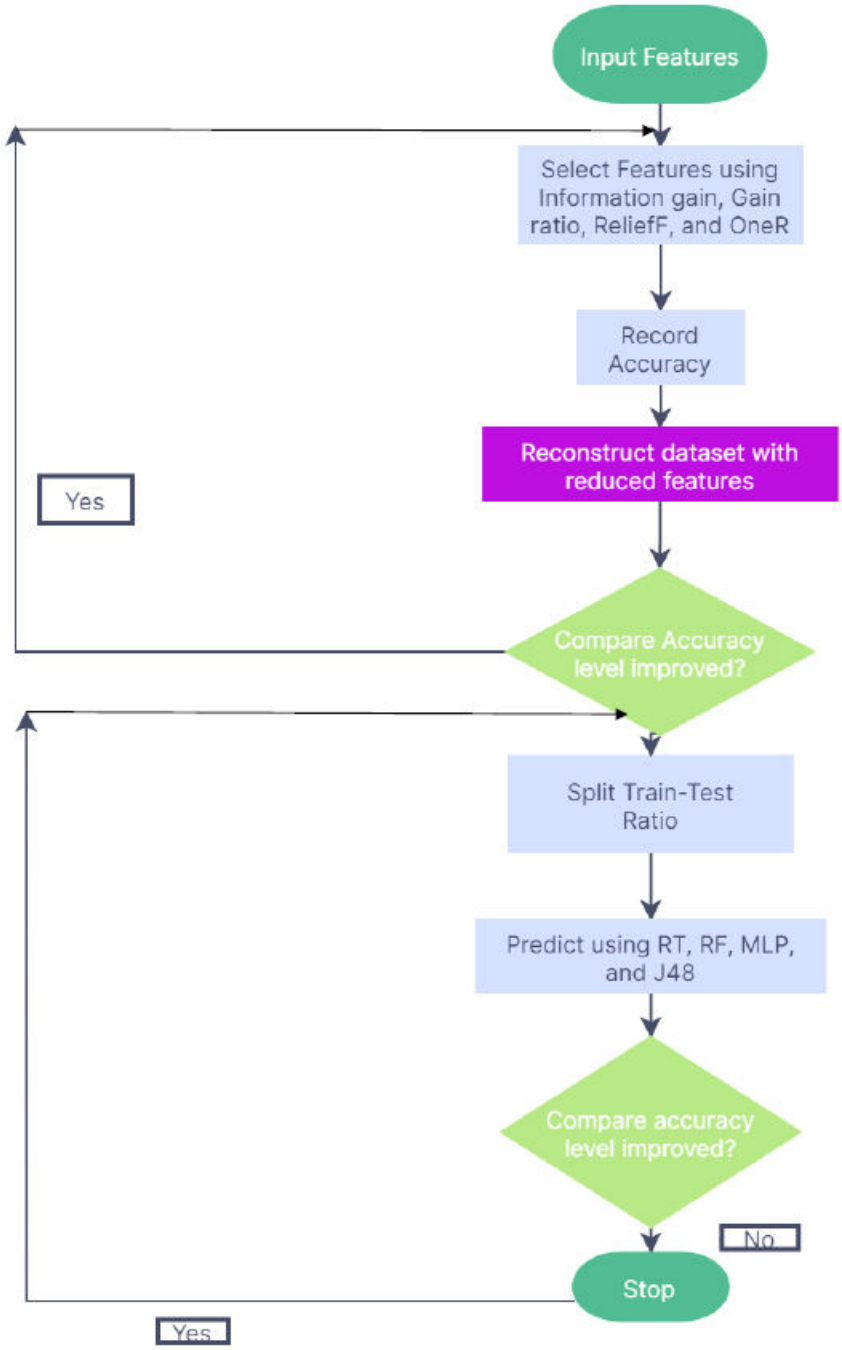


Fig 1: Hybrid Feature Selection and Stacked Generalization Model (HFSSGM) Algorithm

A fatal threat to mankind particularly on the women community is breast cancer. So, it is time to develop a proper machine-learning algorithm that can predict the presence of

breast cancer cells in the human body. Although many researchers have done valuable research in this field, they got low accuracy as compared to this paper. In this paper, the authors used almost all possible algorithms of machine learning to classify whether a lump present in the human body is leading to breast cancer or not and they got almost 100% accuracy. The papers that served as relevant literature for this paper are mentioned below with the accuracy the corresponding researchers got in those papers.

References	Algorithms	Sampling Strategies	Classification Accuracies (%)
Quinlan 1996 [20]	C4.5 DT	10–fold cross validation	94.74
Setiono 1996 [21]	Pruned ANN	50–50 training-testing	96.56
Bennett & Blue 1998 [22]	SVM	5–fold cross validation	97.20
Setiono 2000 [23]	Neuro-rule ANN	10-fold cross validation	97.97
Sarkar & Leong 2000 [24]	k-NN	50–50 training-testing	98.25
	Fuzzy k-NN	50–50 training-testing	98.83
Abbass 2002 [25]	EANN	80–20 training-testing	98.10
Bagui et al., 2003 [26]	k-RNN	10-fold cross validation	98.10
Kiyan & Yildirim 2004 [27]	RBN	50–50 training-testing	96.16
	GRNN	50–50 training-testing	98.80
	PNN	50–50 training-testing	97.00
	MLP	50–50 training-testing	95.74
Polat et al., 2005 [28]	C4.5 + FS-AIRS	10–fold cross validation	98.51
Pach & Abonyi 2006 [29]	F-DT	10–fold cross validation	95.27
Polat & Gne 2007 [30]	LS-SVM	10–fold cross validation	98.53

Akay 2009 [31]	F-score-SVM	10–fold cross validation	99.51
Karabatak & Ince 2009 [32]	AR-ANN	3–fold cross validation	97.40
Marcano-Cedeño et al., 2011 [33]	AMMLP	60–40 training-testing	99.26
Chen et al., 2011 [34]	RS-SVM	80–20 training-testing	96.87
Fan et al., 2011 [35]	CBFDT	75–25 training-testing	98.90
Chen et al., 2012 [36]	PSO-SVM	10-fold cross validation	99.31
Koyuncu & Ceylan 2013 [37]	RF-ANN	50–50 training-testing	98.05
	PSO-ANN	50–50 training-testing	97.36
Medjahed & Saadi 2013 [38]	k-NN (Euclidean)	Holdout method	98.70
Azar & El-Said 2014 [39]	PSVM	4–fold cross validation	96.00
	NSVM	4–fold cross validation	96.57
	LPSVM	4–fold cross validation	97.14
	LSVM	4–fold cross validation	95.43
	SSVM	4–fold cross validation	96.57
Sumbaly et al., 2014 [40]	J48	10–fold cross validation	94.36
Seera & Lim 2014 [41]	FMM-CART-RF	50–50 training-testing	97.29
Bhardwaj & Tiwari 2015 [42]	GOANN	10-fold cross validation	99.26
Nahato et al., 2015 [43]	RS-BPANN	80–20 training-testing	98.60
Kumar et al., 2017 [44]	SVM-Naive Bayes-J48	10–fold cross validation	97.13

Latchoumi & Parthiban 2017 [45]	WPSO-SSVM	5–fold cross validation	98.42

Table 1: Comparison of previous works

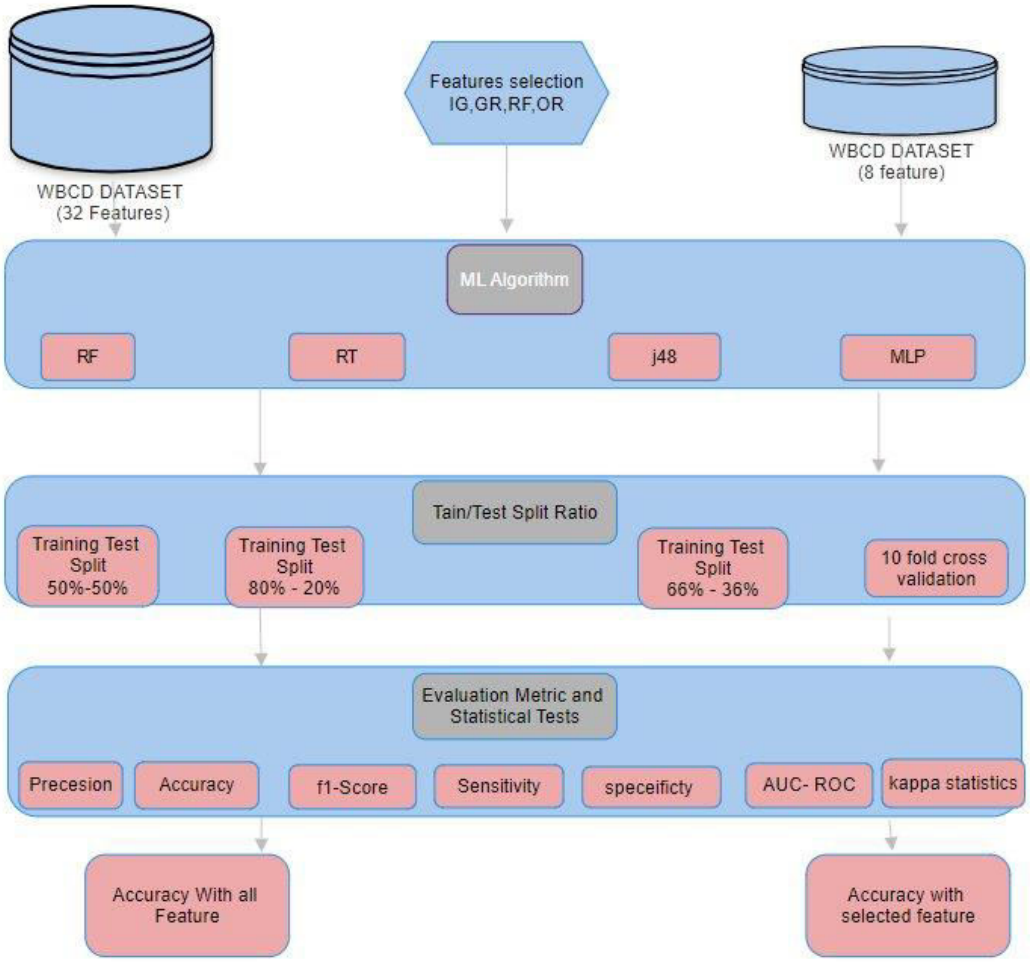


Fig 2: The Architecture for the purposed system

2. Methodologies, Results, and Discussion

2.1. Feature Selection Methods

As the name suggests feature selection methods can be categorized into wrapper and filter methods to select important features among a large set of features present in the dataset which is to be used for training and testing of machine learning algorithms. The wrapper method makes subsets of feature sets and tests which subset is most accurate for a selected ml model. This selection is done by the wrapper method based

on accuracy or particularly based on the performance of the model. This method is akin to selecting the most crucial subset of features that is most accurate for the best models' performance. However, this model's iterative nature makes it computationally costly.

Unlikely, the filter method is focused on the assessment of every feature present in the dataset independently. In this method, different statistical methods like Information Gain, Relief-F, OneR, and Gain Ratio are used to calculate the relevance of each feature on the basis of some predefined criteria. This method is less expensive in terms of time and computational resources.

2.1.1 Information Gain

Information gain selects features based on that feature's usefulness in the prediction of a targeted variable. It calculates the entropy (i.e. the uncertainty) of the target variable before and after the addition of a feature in the dataset. It should be noted that features having higher information gain value are more informative for the target variable [9].

3.1.2 ReliefF

To handle datasets with target variables having both binary or multiclass data points ReliefF is designed by researchers. As per the basic algorithm of relief-F, the Euclidian distances are calculated for each feature vector instance with the targeted feature vector. The process continues iteratively for every instance in the feature vector and ranks them accordingly. This whole process makes it robust for noisy or incomplete datasets [10].

3.1.3 OneR

Holte's One Rule which is popular as the OneR feature selection algorithm is a very simple but effective method for selecting features in a given dataset. In this algorithm, a rule is generated for every feature and then the evaluation of the contribution of that feature to the accuracy of the model. If the feature affects the accuracy of the model by decreasing its value, then the feature is ranked low otherwise high rank is assigned [11].

3.1.4 Gain Ratio

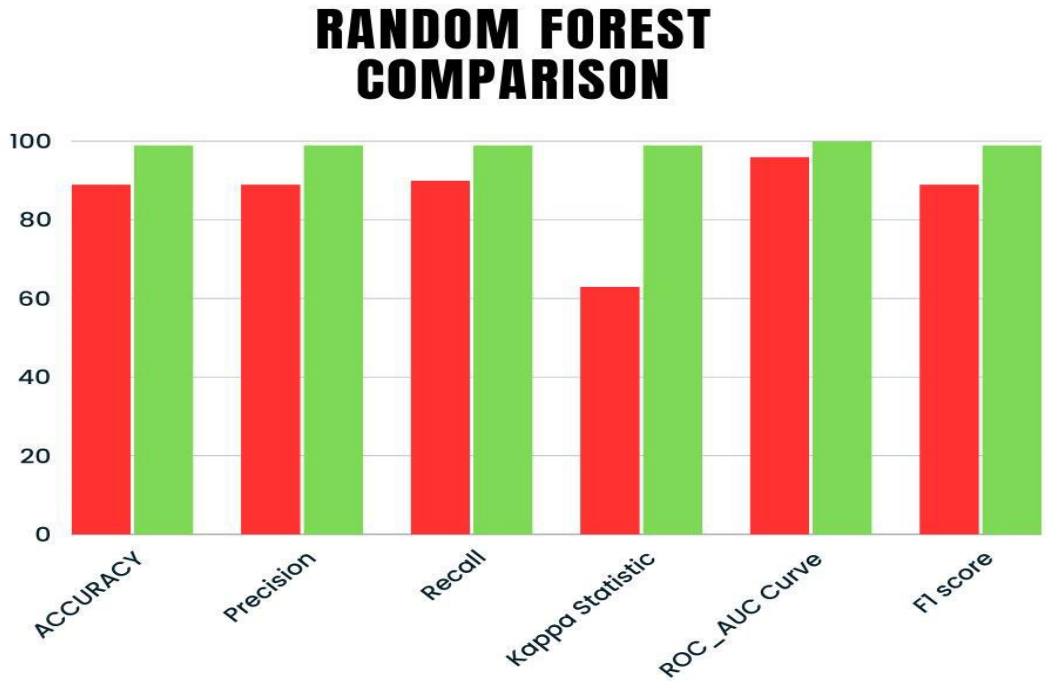
The bias of information gain can be addressed by the gain ratio through consideration of intrinsic information about split points. The number of branches and the distribution of instances among these branches is adjusted by this algorithm at the time of selection of features. The gain ratio provides a more balanced assessment of feature importance across different datasets and splits because it can normalize information gain based on split entropy [12].

3.2 Classification Algorithms

The realm of machine learning contains various algorithms for serving distinct roles. It will be noted that every algorithm has its strengths as well as some limitations. In this research paper, the authors did a comprehensive analysis of prominent algorithms such as Random Trees (RT), Random Forest (RF), Multi-Layer Perceptron (MLP), and J48. In this study, the nuance characteristics of these four algorithms are meticulously scrutinized and by this process, we aim to present this study as a thorough understanding of their performance and contribution in the enhancement of their adaptability and effectiveness for diverse applications [13].

3.2.1 Random Forest

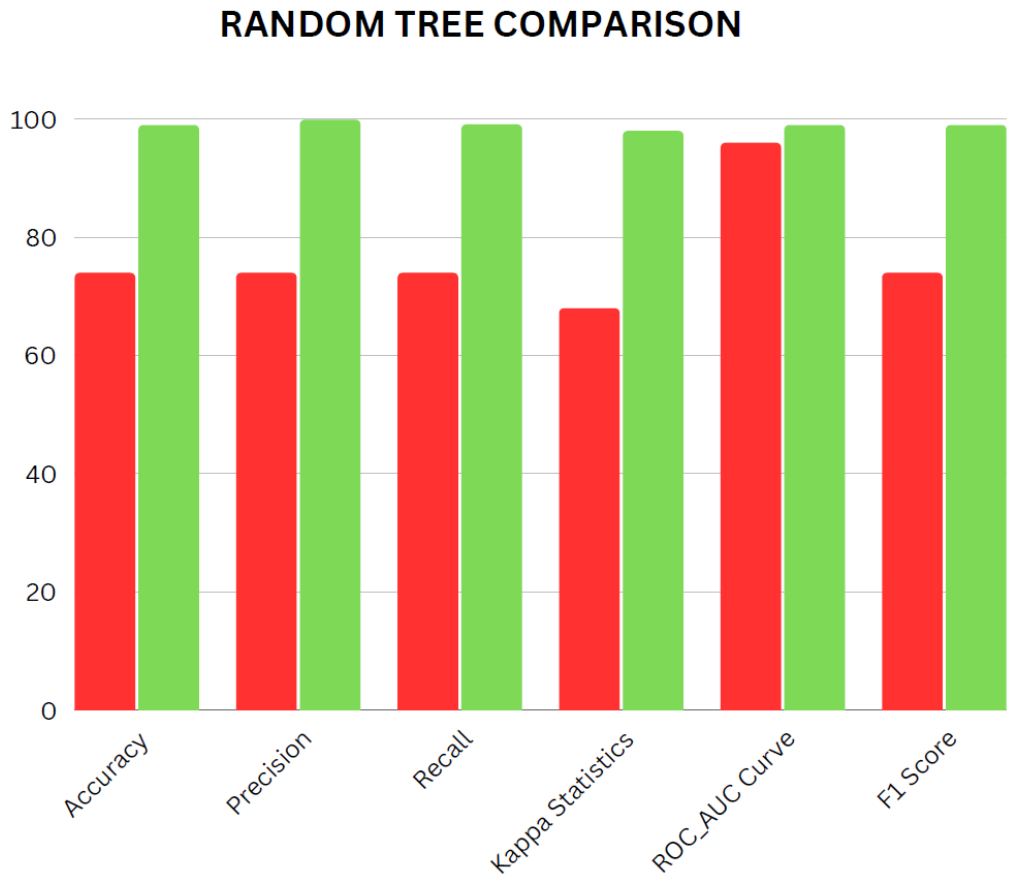
Random forest is a petty good algorithm that is mainly designed to perform decision-making tasks for a computer. As the name suggests it is a collection of decision trees. But the catch is that it is not dependent on only one tree's decision-making capabilities. It takes a branch of decision trees and then takes some trees among them randomly for voting to support or deny a decision. By this process, it can mitigate the probability of making an error by a single decision tree. The reason for the popularity of this algorithm is its robustness. It can handle both simple and complex datasets. It also performs well with heterogeneous data sets. Moreover, it does not become confused when anomalies and outliers are present in the dataset. The most critical feature of a dataset is prioritized by a random forest algorithm to make a decision. Autonomous improvement of the decision-making trees present in a random forest collaboratively is extended to its learning process also. As a result of it the need for manual human intervention in the algorithm is not needed. These phenomena enhance the efficiency and adaptability of this algorithm in various analytical processes [14].



3.2.2 Random Tree:

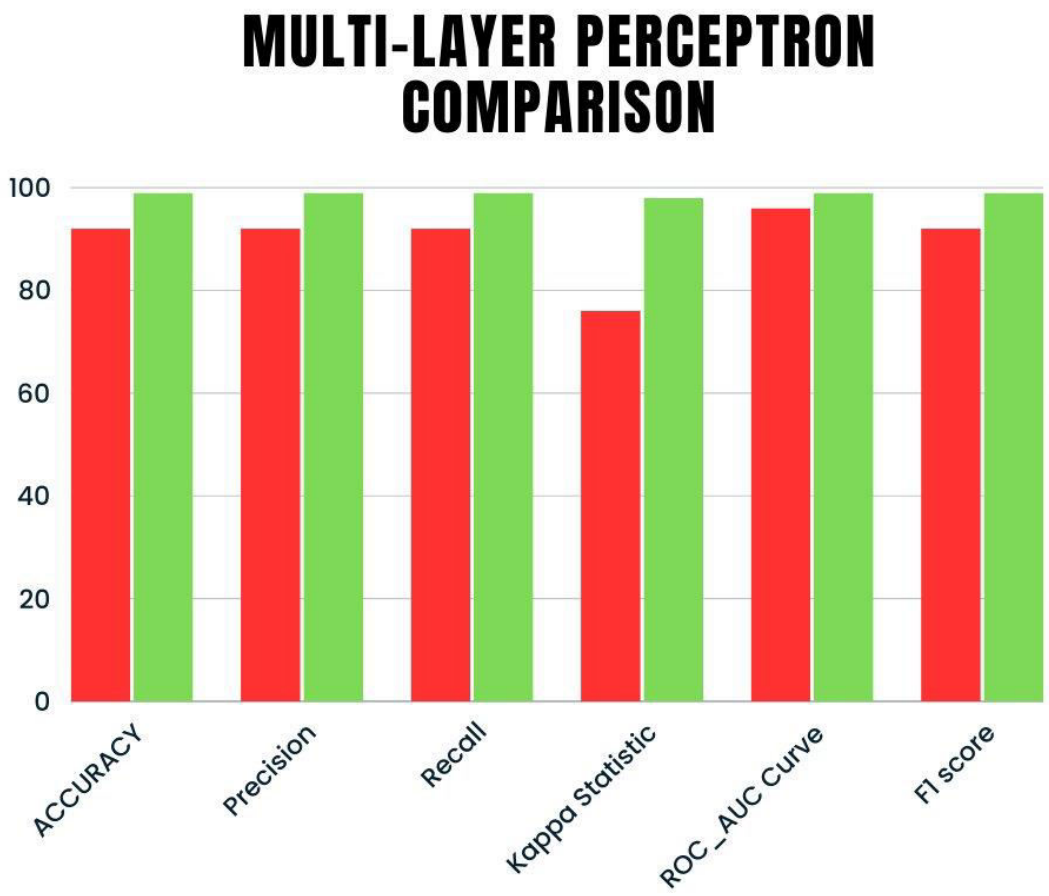
The algorithm mentioned above is also a part of the learning methods by which machines are learned. Multiple models' involvement in ensemble learning is required for the improvement of performance and robustness. In this particular algorithm, multiple decision trees are being used for different tasks like classification and regression. To perform those tasks the algorithm itself divides the whole data set based on features and datapoints. Then for each subset of the original dataset, a decision tree is constructed by the algorithm. After the formulation of all decision tress, the algorithm completes its prediction task by averaging the results obtained from each decision tree. This method proves itself effective when the dataset is large enough and when missing values may or may not be present in that dataset. As the algorithm makes an average of all predictions of all the decision trees, it reduces the chance of overfitting.

The phenomena of overfitting occurs when a model is built too closely to the training data and as a result, it performs poorly on new unseen testing data. At the same time, this algorithm predicts more stable and accurate results based on low variance which is due to its averaging nature. In short, this ensemble algorithm creates a more reliable and efficient model by summing up the strengths of all decision trees and compensating their weaknesses for large-scale machine learning tasks [14].



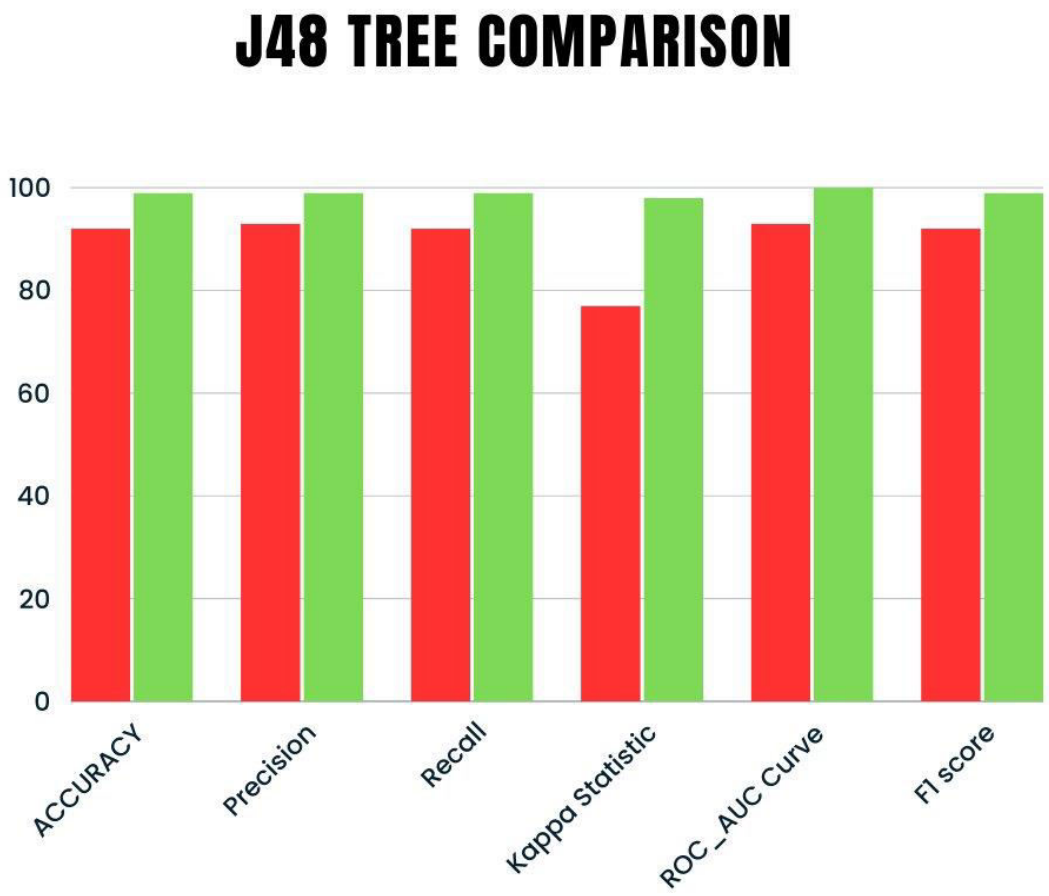
3.3.3 Multilayer Perceptron:

Multilayer Perceptron which is also known as MLP is one type of perceptron among all neural network models present in the Artificial Neural Network family. It is well known for its versatile and powerful nature among machine learning algorithms. To understand the complex relationship between input and output data MLP is widely used by the research community. MLP also performs well enough when smoothening the relationship between input and output data is required. Although MLP performs well with a small amount of data when the question of computational resource and learning rate concerning time arises with a large amount of data in the case of MLP the research scholars opted for another algorithm. As the name suggests MLP is a neural network consisting of one input, multiple hidden, and one output layer. The input layer takes input data and passes it to the hidden layer next to it then multiple hidden layers try to establish the input-output relationship among different data points by adjusting the weight vector and passing a feed to the last layer present in the network which is the output layer. By this process, a model based on MLP is trained to make predictions on unseen testing data [15].



3.3.4 J48 / C4.5 Tree Classifier

A renowned and influential classification algorithm in the field of machine learning is J48 which is also known as C4.5. The basic principle of this algorithm is that it makes a Tree data structure based on information entropy which means whether it meets any new feature in the training data set it creates a tree with two leaf nodes and every branch it creates is a decision and nodes represent a class. It determines the importance of any feature in the dataset by the dataset into two classes and those classes will further be divided into another two classes each. This process goes on iteratively. J48 can deal with missing values by distributing them between all feature classes. This process makes it advantageous for handling datasets with missing values. Pruning is a valuable feature of the J48 tree as it can reduce the size of the decision tree whenever pruning is needed to reduce the phenomena of overfitting and then generalize the model for new unseen data. Continuous data can also be handled by this algorithm by partitioning the continuous data points of a dataset based on a threshold value. Its versatile nature makes it significant among all classification algorithms in the field of machine learning [16].



4. Performance Metrics

A Confusion Matrix is used as one of the necessary tools for evaluating the performance of a learning model. In order to measure performance, the following four key terms are included within the calculation of the Confusion Matrix. To be noteworthy:

1. True Positive (TP): This indicates the number of patients correctly identified as being those who have breast cancer.
2. False Positive (FP): This is that group of patients who do not have breast cancer but are stated to be with the disease.
3. True Negative (TN): This refers to the number of patients accurately diagnosed as patients with no breast cancer.
4. False Negative (FN): This represents those patients diagnosed with breast cancer who are said to have none.

Accuracy: Accuracy is the measure of the extent to which the predictions made by the study are correct relative to all the predictions of the interborder. Equation 1 below depicts the equation accuracy.

$$ACCURACY = \frac{TP + TN}{TP + TN + FP + FN} \quad (1)$$

Precision: Precision is defined as the portion of suspected mental health crisis patients who were correctly identified over the total number of patients identified with any situation of a mental health crisis. Equation 2 explains the precision formula.

$$PRECESSION = \left(\frac{TP}{FP + TP} \right) \tag{2}$$

Recall/sensitivity: As for recall, sensitivity is also known as recall, in this case, it evaluates the proportion of people experiencing a crisis of mental health and correctly recognized by the algorithm, relative to the actual number of people suffering from a mental health crisis. The factor of recency/depth offers the maximum recall or sensitivity ratio amongst concepts. It is physically possible, and equation 3 below illustrates the recall/sensitivities.

$$RECALL = \left(\frac{TP}{FN + TP} \right) \tag{3}$$

Specificity: Accuracy considers the effectiveness of a test in excluding those who do not require the emergency mental health services. It is the non-mental health crisis patients divided by the number of patients that the algorithm predicts will be non-mental health crisis patients. Equation 4 represent the specificity formula

$$SPECIFICITY = \left(\frac{TN}{TN + FP} \right) \tag{4}$$

F1 score: F1 score is a great way to come up with the right balance of between the measures of Precision and Recall. It appears to be a midpoint figure derived from the harmonic average of the level of accuracy and recall (or scale). The formula used in the computation for F1 is indicated in Equation 5.

$$F1SCORE = \frac{2(PRECISION \times RECALL)}{PRECISION + RECALL} \tag{5}$$

AUC-ROC curve: The AUC-ROC Curve is a statistical model that can be applied to different two classes discriminant functions, as for example if a given patient has or not a disease like cancer. He tries to imagine how the model treats the two classes. The ROC component allows the extent to which the total model correctly labels an ailment (True Positives) while at the same time, it also tends to over-diagnose a disease in healthy people (False Positives). The discrimination ability is predicted by the AUC (Area Under the Curves) value, which is the degree of this performance that is like a value closer to 1. Finally, Equations 6, 7 illustrated AUC-ROC curves.

$$TPR = \frac{TP}{TP + FN} \tag{6}$$

$$FPR = \frac{FP}{FP + TN} \tag{7}$$

Kappa statistics: Kappa Statistics is significant assessment that is used in the assessment of the degree of the agreement between two rates who are giving ratings or making judgments as contrasted to the average degree of agreement. It is also useful to find out whether they have an above average, normal, below average, or significantly below average level of agreement. Kappa statistics can be calculated using the formula labelled Equation 8 below [17], [18].

$$K_{STAT} = \frac{A_{OBS} - A_{EXP}}{N - A_{EXP}} \tag{8}$$

5. Dataset Acquisition

The research uses the Breast Cancer Wisconsin Diagnostic database from the University of Wisconsin Hospitals Madison Breast Cancer Database [13]. This dataset comprises of information derived from Breast Cancer Histological Images from Fine Needle Aspiration (FNA) images. These features reflect morphological properties of cell nuclei as seen in these images. The records used to build the dataset amount to 569 with 357 of them being classified as benign, 212 as malignant. It categorizes cases into two classes: 62. They were particulate, benign in 74% and malignant in 37.26%. The dataset comprises 8 integer-valued attributes: To decide which are important, we must analyze The Wagner with GELOM ID, Compactness_se, Concavity_se, the Fractal_dimension_mean, Concavity_mean, Area_mean, the Concave_points_mean, the Concave_points_worst, and, of process, the dependent variable Diagnosis [19].

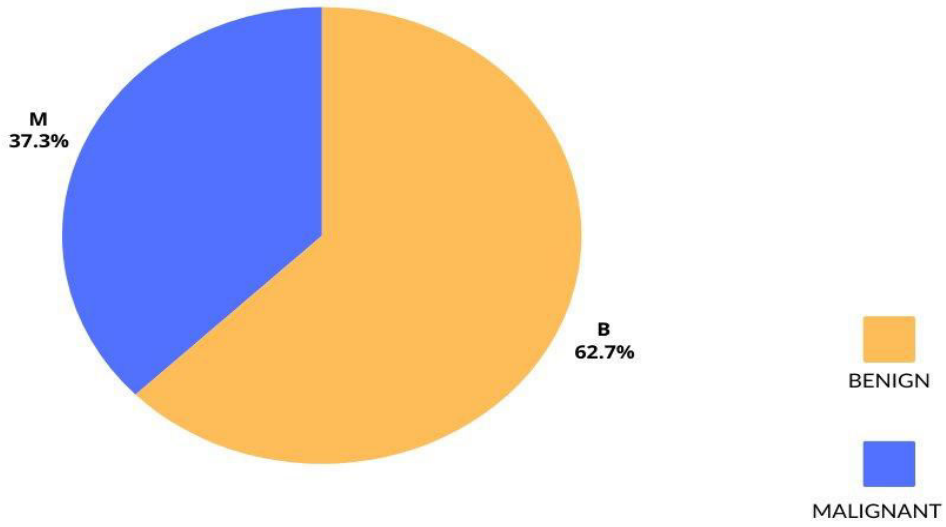


Fig 7: Wisconsin Breast Cancer Diagnostic Dataset

6. Result and discussion

The author discusses the basic strategy to obtain high levels of consistency, sensitivity, and specificity on the given problem. Furthermore, the author also examines the means by which these approaches can be combined. Previous researchers in the field eliminate some features for the aim of increasing accuracy without actually being mindful of the fact that such instances may carry important information. Hence, this paper develops a framework to support Breast Cancer predictions using DMTs to benefit humanity.

The study employs Four feature selection techniques—OneR, ReliefF, Gain Ratio and InfoGain—to rank attribute importance, alongside four classification algorithms: That is MLP, J48, Random Forest (RF) and Regression Trees (RT). The dataset has 32 variables where one variable is function of the other 31 variables.

First, these four feature selection algorithms are applied in the study to determine that it is important to measure levels of breast cancer using eight minimum attributes. In the second stage, the authors fit the four classification algorithms to both datasets: with 32 attributes and with 8 optimized attributes. Table 1 shows the accuracies in percentage according to the 50-50, 66-34, 80-20 and 10-fold CV splits. Namely, J48, MLP, RF and RT with the near to 99% accuracy for the 80-20 data split and RT, J48, MLP, and RF for 10-fold cross validation.

Train – Test Split	Number of Features	J48	MLP	RF	RT
50 - 50	32	0.89	0.81	0.85	0.70
	8	0.95	0.98	0.98	0.94
66 - 34	32	0.89	0.79	0.83	0.70
	8	0.98	0.97	0.97	0.96
80 - 20	32	0.92	0.84	0.85	0.65
	8	0.98	0.95	0.98	0.97
10-fold Cross Validation	32	0.91	0.87	0.89	0.74
	8	0.99	0.99	0.99	0.99

Table 2: Comparison of accuracy

Table 2 reflects performance analysis of precision. Extra 80/20 data split gave 99% precision as well where SGD, J48, MLP, RF and RT performed the best In 10 fold cross validation, all classes gave 99% precision where SGD, J48, MLP, RF performed best. Besides that, we achieved over 90% precision rates in other classifiers that have been displayed in the Table 2.

Train – Test Split	Number of Features	J48	MLP	RF	RT
50 - 50	32	0.90	0.81	0.86	0.71
	8	0.95	0.98	0.97	0.94
66 - 34	32	0.89	0.79	0.85	0.71
	8	0.99	0.97	0.97	0.96
80 - 20	32	0.93	0.84	0.86	0.65
	8	0.98	0.95	0.98	0.97
10-fold Cross Validation	32	0.96	0.87	0.89	0.74
	8	0.99	0.99	0.99	0.99

Table 3: Comparison of precision

Table 3 and table 4 outlines the sensitivity/recall performance analysis The F1-score corresponding to each combination is also outlined in the subsequent table below. The models consolidated 99% sensitivity in all of the data splits: MLP, J48, and RF and the Random Tree was lower though it also did not fall lower than 99%. In table 4, MLP, J48 and RF had F1-score of 99% in both 80-20 data split and 10-fold cross validation.

Train – Test Split	Number of Features	J48	MLP	RF	RT
50 - 50	32	0.89	0.81	0.85	0.70
	8	0.95	0.98	0.98	0.94
66 - 34	32	0.89	0.79	0.83	0.71
	8	0.99	0.97	0.97	0.96
80 - 20	32	0.93	0.84	0.86	0.65
	8	0.98	0.95	0.98	0.97
10-fold Cross Validation	32	0.91	0.87	0.89	0.74
	8	0.99	0.99	0.99	0.99

able 4: Comparison of specificity/recall

Train – Test Split	Number of Features	J48	MLP	RF	RT
50 - 50	32	0.89	0.81	0.84	0.71
	8	0.95	0.98	0.	0.94
66 - 34	32	0.89	0.79	0.82	0.71
	8	0.99	0.97	0.97	0.96
80 - 20	32	0.93	0.84	0.85	0.65
	8	0.98	0.95	0.98	0.97
10-fold Cross Validation	32	0.91	0.87	0.89	0.74
	8	0.99	0.99	0.99	0.99

Table 5: Comparison of f1-score

In this research study, among the four classifiers, the best results were provided by all the classifiers, providing the accuracy, sensitivity, precision rate and f1-score of 99.99%

7. Conclusion

In this comparison, four evaluation methods (Random Forest, Random Tree, Multilayer Perceptron) were used in combination with other attribute selection measures, namely OneR, Gain Ratio, Information Gain, and ReliefF. Such algorithms were utilized optimally to construct a beneficial graphical aid for estimating the mass in question. It makes it possible for researchers to calculate several emergency factors for distinct forms of sicknesses to people. This way, the above-mentioned sorts of assessments can direct the societal evaluations toward higher and better probability of preventing future health conditions in an effective manner. In addition, it will have its benefits in

identifying diseases at a relatively young stage and in giving the required medical care. Each of the methods aims to be the key to early diagnosis of certain diseases and early detection of possible aggravating factors.

In the application on the Wisconsin Breast Cancer Diagnostic dataset (WBCD), five main algorithms were used: Random Forrest, Random Tree, J48, Multilayer Perceptron. Algorithms of this nature can therefore be helpful in determining health deformities in people, based on this research study. Still, it is pertinent to mention that the validity of the derived results does vary depending on the nature and quality of data collected the size of the sample and the selection of the variables.

8. References

1. Nayak, S., & Gope, D. (2017, June). Comparison of supervised learning algorithms for RF-based breast cancer detection. In *2017 Computing and Electromagnetics International Workshop (CEM)* (pp. 13-14). IEEE.
2. Gayathri, B. M., & Sumathi, C. P. (2016, December). Comparative study of relevance vector machine with various machine learning techniques used for detecting breast cancer. In *2016 IEEE International Conference on Computational Intelligence and Computing Research (ICCIC)* (pp. 1-5). IEEE.
3. Asri, H., Mousannif, H., Al Moatassime, H., & Noel, T. (2016). Using machine learning algorithms for breast cancer risk prediction and diagnosis. *Procedia Computer Science*, *83*, 1064-1069.
4. Khourdifi, Y., & Bahaj, M. (2018, December). Applying best machine learning algorithms for breast cancer prediction and classification. In *2018 International conference on electronics, control, optimization and computer science (ICECOCS)* (pp. 1-5). IEEE.
5. Osman, A. H. (2017). An enhanced breast cancer diagnosis scheme based on two-step-SVM technique. *Int. J. Adv. Comput. Sci. Appl*, *8*(4), 158-165.
6. Quinlan, J. R. (2014). *C4. 5: programs for machine learning*. Elsevier.
7. Saoud, H., Ghadi, A., Ghailani, M., & Abdelhakim, B. A. (2019). Using feature selection techniques to improve the accuracy of breast cancer classification. In *Innovations in Smart Cities Applications Edition 2: The Proceedings of the Third International Conference on Smart City Applications* (pp. 307-315). Springer International Publishing.
8. Das, S., Kar, S. P., Sil, S., Molla, A. R., Rajak, R., & Chaudhuri, A. K. (2024). A Multifaceted Approach to Understanding Mental Health Crises in the COVID-19 Era: Using AI Algorithms and Feature Selection Strategies. In *AI-Driven Innovations in Digital Healthcare: Emerging Trends, Challenges, and Applications* (pp. 97-119). IGI Global.
9. Kar, S. P., Molla, A. R., Das, S., Rajak, R., Sil, S., & Chaudhuri, A. K. (2024). Identification of Insecurity in COVID-19 Using Machine Learning Techniques.

In *Medical Robotics and AI-Assisted Diagnostics for a High-Tech Healthcare Industry* (pp. 239-256). IGI Global.

10. Osareh, A., & Shadgar, B. (2010, April). Machine learning techniques to diagnose breast cancer. In *2010 5th international symposium on health informatics and bioinformatics* (pp. 114-120). IEEE.
11. Yue, W., Wang, Z., Chen, H., Payne, A., & Liu, X. (2018). Machine learning with applications in breast cancer diagnosis and prognosis. *Designs*, *2*(2), 13.
12. Amrane, M., Oukid, S., Gagaoua, I., & Ensari, T. (2018, April). Breast cancer classification using machine learning. In *2018 electric electronics, computer science, biomedical engineerings' meeting (EBBT)* (pp. 1-4). IEEE.
13. Ahmad, L. G., Eshlaghy, A. T., Poorebrahimi, A., Ebrahimi, M., & Razavi, A. R. (2013). Using three machine learning techniques for predicting breast cancer recurrence. *J Health Med Inform*, *4*(124), 3.
14. Asri, Hiba, Hajar Mousannif, Hassan Al Moatassime, and Thomas Noel. "Using machine learning algorithms for breast cancer risk prediction and diagnosis." *Procedia Computer Science* 83 (2016): 1064-1069.
15. Gayathri, B. M., Sumathi, C. P., & Santhanam, T. (2013). Breast cancer diagnosis using machine learning algorithms–a survey.
16. Chaudhuri, A. K., Banerjee, D. K., & Das, A. (2021). A dataset centric feature selection and stacked model to detect breast cancer. *International Journal of Intelligent Systems and Applications*, *13*(4), 24.
17. Chaudhuri, A. K., Sinha, D., Banerjee, D. K., & Das, A. (2021). A novel enhanced decision tree model for detecting chronic kidney disease. *Network Modeling Analysis in Health Informatics and Bioinformatics*, *10*, 1-22.
18. Chaudhuri, A. K., Ray, A., Banerjee, D. K., & Das, A. (2021). A multi-stage approach combining feature selection with machine learning techniques for higher prediction reliability and accuracy in cervical cancer diagnosis. *International Journal of Intelligent Systems and Applications*, *10*(5), 46.
19. Chaudhuri, A. K., Das, S., & Ray, A. (2024). An Improved Random Forest Model for Detecting Heart Disease. In *Data-Centric AI Solutions and Emerging Technologies in the Healthcare Ecosystem* (pp. 143-164). CRC Press.
20. Quinlan, J. R. (1996). Improved use of continuous attributes in C4. 5. *Journal of artificial intelligence research*, *4*, 77-90.
21. Setiono, R. (1996). Extracting rules from pruned neural networks for breast cancer diagnosis. *Artificial intelligence in medicine*, *8*(1), 37-51.
22. Bennett, K. P., & Blue, J. A. (1998, May). A support vector machine approach to decision trees. In *1998 IEEE international joint conference on neural networks proceedings. IEEE world congress on computational intelligence (Cat. No. 98CH36227)* (Vol. 3, pp. 2396-2401). IEEE.

23. Setiono, R. (2000). Generating concise and accurate classification rules for breast cancer diagnosis. *Artificial Intelligence in medicine*, *18*(3), 205-219.
24. Sarkar, M., & Leong, T. Y. (2000). Application of K-nearest neighbors algorithm on breast cancer diagnosis problem. In *Proceedings of the AMIA Symposium* (p. 759). American Medical Informatics Association.
25. Abbass, H. A. (2002). An evolutionary artificial neural networks approach for breast cancer diagnosis. *Artificial intelligence in Medicine*, *25*(3), 265-281.
26. Bagui, S. C., Bagui, S., Pal, K., & Pal, N. R. (2003). Breast cancer detection using rank nearest neighbor classification rules. *Pattern recognition*, *36*(1), 25-34.
27. Kıyan, T., & Yıldırım, T. (2004). Breast cancer diagnosis using statistical neural networks. *IU-Journal of Electrical & Electronics Engineering*, *4*(2), 1149-1153.
28. Polat, K., Sahan, S., Kodaz, H., & Günes, S. (2005). A new classification method for breast cancer diagnosis: feature selection artificial immune recognition system (FS-AIRS). In *Advances in Natural Computation: First International Conference, ICNC 2005, Changsha, China, August 27-29, 2005, Proceedings, Part II 1* (pp. 830-838). Springer Berlin Heidelberg.
29. Akay, M. F. (2009). Support vector machines combined with feature selection for breast cancer diagnosis. *Expert systems with applications*, *36*(2), 3240-3247.
30. Polat, K., & Güneş, S. (2007). Breast cancer diagnosis using least square support vector machine. *Digital signal processing*, *17*(4), 694-701.
31. Akay, M. F. (2009). Support vector machines combined with feature selection for breast cancer diagnosis. *Expert systems with applications*, *36*(2), 3240-3247.
32. Karabatak, M., & Ince, M. C. (2009). An expert system for detection of breast cancer based on association rules and neural network. *Expert systems with Applications*, *36*(2), 3465-3469.
33. Marcano-Cedeño, A., Quintanilla-Domínguez, J., & Andina, D. (2011). WBCD breast cancer database classification applying artificial metaplasticity neural network. *Expert Systems with Applications*, *38*(8), 9573-9579.
34. Chen, H. L., Yang, B., Liu, J., & Liu, D. Y. (2011). A support vector machine classifier with rough set-based feature selection for breast cancer diagnosis. *Expert systems with applications*, *38*(7), 9014-9022.
35. Fan, C. Y., Chang, P. C., Lin, J. J., & Hsieh, J. C. (2011). A hybrid model combining case-based reasoning and fuzzy decision tree for medical data classification. *Applied Soft Computing*, *11*(1), 632-644.
36. Chen, H. L., Yang, B., Wang, G., Wang, S. J., Liu, J., & Liu, D. Y. (2012). Support vector machine based diagnostic system for breast cancer using swarm intelligence. *Journal of medical systems*, *36*, 2505-2519.
37. Koyuncu, H., & Ceylan, R. (2013, July). Artificial neural network based on rotation forest for biomedical pattern classification. In *2013 36th International*

conference on telecommunications and signal processing (TSP) (pp. 581-585). IEEE.

38. Medjahed, S. A., Saadi, T. A., & Benyettou, A. (2013). Breast cancer diagnosis by using k-nearest neighbor with different distances and classification rules. *International Journal of Computer Applications*, *62*(1).
39. Azar, A. T., & El-Said, S. A. (2014). Performance analysis of support vector machines classifiers in breast cancer mammography recognition. *Neural Computing and Applications*, *24*, 1163-1177.
40. Sumbaly, R., Vishnusri, N., & Jeyalatha, S. (2014). Diagnosis of breast cancer using decision tree data mining technique. *International Journal of Computer Applications*, *98*(10).
41. Seera, M., & Lim, C. P. (2014). A hybrid intelligent system for medical data classification. *Expert systems with applications*, *41*(5), 2239-2249.
42. Bhardwaj, A., & Tiwari, A. (2015). Breast cancer diagnosis using genetically optimized neural network model. *Expert Systems with Applications*, *42*(10), 4611-4620.
43. Nahato, K. B., Harichandran, K. N., & Arputharaj, K. (2015). Knowledge mining from clinical datasets using rough sets and backpropagation neural network. *Computational and mathematical methods in medicine*, *2015*(1), 460189.
44. Kumar, U. K., Nikhil, M. S., & Sumangali, K. (2017, August). Prediction of breast cancer using voting classifier technique. In *2017 IEEE international conference on smart technologies and management for computing, communication, controls, energy and materials (ICSTM)* (pp. 108-114). IEEE.
45. Latchoumi, T. P., & Parthiban, L. (2017). Abnormality detection using weighed particle swarm optimization and smooth support vector machine. *Biomedical Research*, *28*(11), 4749-4751.