International Journal of Information Retrieval Research
Volume 12 • Issue 2

Accuracy Enhancement for Breast Cancer Detection Using Classification and Feature Selection

Somil Jain, Mody University of Science and Technology, India*
Puneet Kumar, Mody University of Science and Technology, India

ABSTRACT

Chronic diseases like kidney failure, heart disease, cancer, etc. are the major causes of deaths worldwide. The most dangerous type of disease from which the women of every age group are suffering is breast cancer. To detect this type of disease at an early stage is a challenging task. In order to predict the breast cancer at an early stage, a classification algorithm of high accuracy and a lower error rate is desirable. In this research work, the authors have used four classification algorithms—K-NN, J48, logistic regression, and Bayes net—for building the predictive model. Also, the wrapper method of feature selection is used to enhance the accuracy rate and reduce the error rate of the used classifiers. To carry out this research, the authors have used the Wisconsin Diagnostic Breast Cancer dataset, which contains 569 instances along with 32 attributes and a class attribute, which will predict the type of cancer.

KEYWORDS

Breast Cancer, Classifiers, Feature Selection, Wrapper Approach

1. INTRODUCTION

Breast cancer is the most widely occurring cancers in women across the globe (Adnan et.al., 2020); it is accountable for 25% of the total type of cancers. It is estimated that it will reach around 2.1 million till 2030, especially in the young age women of the developing nations (Khairunnahar et.al., 2019; Oskouei et.al., 2017). Generally, there are two type of stages for breast cancer i.e. Benign and Malignant. Benign stage is the initial stage in which the tumor occurs in one part of the body and it can be diagnosed easily. On the other hand, malignant stage is quite dangerous for the patient suffering from breast cancer, as the cancerous cells tends to move to other parts of the body and creates a form of contact with other tissues in the body (Heymach et.al., 2017). It is important to diagnose this type of cancer at the early stage in order to reduce the mortality rate (Napoleon & Pavalakodi, 2011; Lahoura et.al., 2021). Here machine learning plays an important role, as it enables various stakeholders to predict the outcome of the disease at an early stage (Islam et.al., 2020). Feature selection helps in creation of the subset from the main dataset for identifying the important features, which improves the computation process and provides better results, it has an advantage that it can work with the classification algorithms in two ways i.e. dependent or independent (Fatima et.al., 2020). Mainly there are two types of learning methods exists (Kourou et.al., 2015; Yildrim, 2015):

DOI: 10.4018/IJIRR.299931 *Corresponding Author

This article published as an Open Access article distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/4.0/) which permits unrestricted use, distribution, and production in any medium, provided the author of the original work and original publication source are properly credited.

(I) SUPERVISED LEARNING

Supervised machine learning is the construction of algorithms capable of creating general patterns and hypotheses by using externally distributed instances to predict future instances. Various algorithms for supervised machine learning aim to categorize data on the basis of prior knowledge. The process of Classification and Regression comes under the category of supervised learning methods.

(II) UNSUPERVISED LEARNING

Unsupervised learning is used to draw inferences from datasets that consist of unlabeled input data. The most popular unsupervised form of learning is cluster analysis, which is used to identify hidden patterns or to group in data for exploratory data analysis.

This paper is structured in the following manner:

i) Section 1 provides the general introduction and overview of the paper.
ii) Section 2 provides Literature review.
iii) Section 3 provides the methodology of the work to be done along with a brief introduction of the used methods.
iv) Section 4 this section shows the experimental results.
v) Section 5 provides the conclusion of the study.

2. RELATED WORKS

(Khairunnahara et.al., 2019) has pointed the importance of feature selection with classification, they have used logistic regression as their classifier with a dynamic weighted approach and they have attained the accuracy of 96.83% on the Wisconsin Diagnostic Breast Cancer Dataset (WDBC). (Yildirim, 2015) has used feature selection along with J48, Naïve Bayes, IBK, Decision Tree as their classifiers on the evaluation parameters like precision, recall, and root mean square error (RMSE) and obtained best results from Naïve Bayes using subset evaluator with an RMSE of 0.3446% and they have suggested that feature selection is capable of improving the performance of the classification algorithms. (Shaikh & Ali, 2019) used the Naïve Bayes, KNN, SVM as there classifiers along with wrapper subset evaluator of feature selection on the Wisconsin Diagnostic Breast Cancer Dataset (WDBC) and Breast Cancer Digital Repository (BCDR) and there results shows a significant form of increment in the prediction accuracy rate of the classifiers using WDBC it ranges between 2-5% and for BCDR it ranges between 2-10%. (Pritom et.al., 2016) stated that, breast cancer is the major cause of deaths especially amongst the middle aged women's and it is the second most dangerous cause of deaths, in order to improve the accuracy the authors has used ranker method of feature selection along with Info gain attribute Evaluator for Naïve Bayes & C4.5 and for SVM they have used SVM attribute evaluator and reported a incremented growth as 1.5% from SVM, 2.53% from C4.5 and 9.09% from Naïve Bayes. (Darzi et.al., 2011) has applied wrapper-based approach using WDBC dataset based on genetic algorithm and case-based reasoning. They reported the accuracy improvement from 94.74% to 97.97%. (Chokka & Rani, 2019) have used PCA with Adaboost using 6 different datasets and merging them as one dataset and suggested that feature selection plays an important role in enhancing the performance of the base classifier. (Kumari & Swarnkar, 2011) have pointed out that selection of appropriate feature and removing least important features is a challenging task, as it leads to less accurate results, and also increase in the computational cost. (Dasgupta et.al., 2019) have used Naïve Bayes, Neural Networks, Decision Tree, Logistic Regression as their base classifiers over WDBC dataset and selected 8-9 feature out of total 32 feature and achieved the accuracy of 98.4% with Naïve Bayes, 97.4% with Logistic Regression, 97.5% with Neural Networks and 93.10 with Decision Tree. (Aalaei et.al., 2016) conducted their study using wrapper approach of

feature selection along with ANN, PS-classifier and GA- classifier, they have applied this approach over 3 Wisconsin datasets i.e. WDBC, WPBC, WBC. The result of their study shows a significant form of increment in the accuracy rate of the classifiers when used with feature selection. (Yang & Xu, 2019) they have emphasized on the utility of feature selection, they have used PCA with SVM and reported an increment of 5% with GA SVM and 2% with DESVM which turns to be quite beneficial while diagnosing the breast cancer. (Diwakar et.al., 2013) used feature extraction technique for extracting the color information of any object, also they have extracted the texture information in order to represent the object in proper manner.

3. METHODOLOGY

In this approach K-NN, J48, Logistic Regression, and Bayes Net were used as the classifiers and wrapper approach of feature selection is used to improve the accuracy rate of the classifiers and to reduce the error rate, BFS searching method is also used for searching the relevant attribute with the wrapper approach. Firstly, the accuracy rate of all the classifiers has been recorded by applying them without feature selection and in the second stage the classifiers used in this approach were integrated with the wrapper approach and the improved results obtained were recorded in terms of accuracy and various other parameters. Table 1 shows the features selected for the individual classifier out of total 32 features from the dataset; these features are obtained by applying wrapper approach with each classifier individually.

Table 1. Selected features of classifiers

Classifier	Features Selected
KNN	1,2,9,10,13,16,20,21,22,23,24,26,30,32
J48	2,3,4,7,10,11,20,23,27,31,32
Logistic Regression	2,4,9,19,22,23,24,25,27
Bayes Net	1,2,4,7,10,16,23,24,27,29,31

3.1 Feature Selection

Feature selection is the process of identifying important features from a particular dataset it is also known as subset selection (Saeys et.al., 2007). The main motive behind identification of the relevant features is to improve the accuracy and to reduce the computational cost & time. It is the technique which has its wide applicability in the area of statistical analysis, medical, Image or Pattern Recognition.

Mainly feature selection is categorized into two parts:

i) Filtering Approach

In this approach ranking method is used by which all the attributes are ranked as per there relevance to the classification tasks, selection of attributes depends on the rank of the attribute. This approach is classifier independent as only one-time feature selection is performed which will selects the features that can be tested with different classifiers (Mwadulo, 2016; Pervez & Farid, 2015; Lu et.al., 2012).

ii) Wrapper Approach

This approach follows an iterative approach, here the learning process is specific with every classification algorithm with a believe that different classification algorithm can perform better with different set of features. This approach is classifier dependent as for each classifier there will be a different set of features (Ruckstieb et.al., 2013; Kabi et.al., 2011; Aljawarneh et.al. 2019).

3.2 Classification Algorithms

i) J48

It is a type of classifier which is having a tree like structure, where the node can be a leaf node showing the target attributes value and the class or a decision node which specifies that some testing is to be performed on a single-valued attribute. J48 has the ability to handle both categorical and continuous attributes, which are helpful in building a decision tree. It can also handle the missing values. It uses the information entropy to build the decision tree from training data which is labeled. It is based on the fact that each data attribute is useful in decision making and this is done by splitting it into smaller subsets. Here the attributes with highest gain information are selected first to make the decision and then the algorithm proceeds further towards smaller subsets. Pessimistic pruning is applicable to J48 for the removal of unnecessary branching in the tree and for improving the classification. This classifier uses a greedy divide-and -conquer method to recursively induce decision trees that contain attributes of a dataset for doing further classification. (Harshvardhan et.al. 2016). Equation 1 and 2 depicts the formulation of entropy and information gain respectively.

$$Entropy\left(D\right) = -\sum_{i=1}^{m} p_i log_2\left(p_i\right) \tag{1}$$

Here, 'p_i' is the probability of a random attribute in the dataset.

$$Gain\left(A\right) = Entropy\left(D\right) - Entropy_A\left(D\right) \tag{2}$$

Here, $Entropy_A\left(D\right)$ is the expected information required for performing classification based on some attribute 'A' which holds some distinct values.

ii) Logistic Regression

It is a statistical regression classifier and has a capacity to form a means of relationship between the categorical variable which is dependent and between one or more independent variables. It has a wide field of application which includes health, social science, bioinformatics, etc. It is generally of two types Binomial and Multinomial. If the dependent variable has the value 0 or 1then it is said to be binary logistic regression analysis and this is quite helpful in disease prediction. On the other hand, multinomial regression uses two or more categories of the dependent variable, it uses maximum likelihood estimation to examine the probability (Li-Yu, 2016). The Logistic Regression can be stated formally as:

$$Log\left[p\,/\,1-p\right] = \beta_0 + \beta_1 X_1 + ... + \beta_n X_{n^2} \tag{3}$$

Where β_0 is the constant and $\beta_1, \beta_2 ..... \beta_n$ are the coefficient of regression predictor variables i.e. $X_1, X_2 ..... X_n$

iii) K-NN

It is a supervised learning classifier and has the ability to identify the features of every individual category well in advance. K-NN can be used for classifying the input points to obtain outcomes in discrete form. Classification in K-NN is rely upon a distance metric. It uses voting technique where the samples are classified on the basis of majority votes from its neighbors. To avoid a curse of dimensionality, it requires various dimensionality reduction techniques like wrapping, feature selection etc. which will help in reducing the complexity. It fetches the characteristics after examining the total distance between the new features which is to be classified and the ones who have been already classified, afterward the nearest K point is used. K-NN uses Euclidean distance to calculate the distance between point P and Q, which will be represented as feature vectors like P= (a_1, $a_{2.....}$ a_n) and Q= (b_1, b_2,…..b_n), here 'n' represents the dimensionality of the feature space (Sabancı & Koklu, 2015; Zhou, 2009; Atoui, 2019). This can be stated formally as:

$$dist\left(P,Q\right)=\sqrt{\frac{\sum_{i=1}^{n}\left(P_i-Q_i\right)^2}{n}} \tag{4}$$

iv) Bayes Net

Bayes Net is a type of graphical model which shows a relationship between the variables based upon probabilistic nature. It is also called Direct Acyclic Graph (DAG) which consists of nodes that represent the variables of random nature and the arcs which show the connection between there variables in the form of G=(V,E), where 'V' represents the vertices and 'E' represents the edges of the graph.. In Bayes Net, a local probability distribution is associated with each node which is to be decided from the parent's state. Within a Bayesian network, given a value of its parents within G, each variable is independent of its non-descendants. Such independence encoded in G reduces the number of parameters that are necessary to explain a joint distribution, so that it can be effectively dependent after distribution. This can be formulated as:

$$P\left(X_{1,}X_2.....X_i\right)=\prod_{i=1}^{n}P(X_i\mid Pa\left(X_i\right)) \tag{5}$$

Here, Pa (X_i) represents the set of parent nodes in graph G.

3.3 About the Dataset

Wisconsin Diagnostic Breast Cancer (WDBC) dataset is used to carry out this research work which is available at machine learning repository of UCI. This dataset has total 569 instances and 32 attributes along with class attribute which will be used for predicting the stage of cancer whether it is benign or malignant. 10 real valued attributes of this dataset are shown in Table 2.

Table 2. Details of the Attribute

Attribute Number	Description of attribute	Range		
		Mean	Standard Error	Worst
1	Radius	6.98-28.11	0.11-2.87	7.93-36.04
2	Texture	9.71-39.28	0.36-4.88	12.02-49.54
3	Perimeter	43.79-188.5	0.76-21.98	50.41-251.20
4	Area	143.5-2501	6.80-542.20	185.20-4254
5	Smoothness	0.05-0.16	0.00-0.003	0.07-0.22
6	Compactness	0.02-0.35	0.00-0.135	0.027-1.058
7	Concavity	0.00-0.427	0.00-0.396	0.00-1.25
8	Concave points	0.00-0.20	0.00-0.053	0.00-0.29
9	Symmetry	0.10-0.30	0.008-0.079	0.157-0.66
10	Fractal Dimension	0.05-0.097	0.001-0.03	0.05-0.20

4. EXPERIMENT AND RESULTS

In this section a detailed view of obtained results of this study has been provided. Performance of the classifiers is measured on different parameters of efficiency and effectiveness by applying 10-fold cross validation technique in order to get more precise results.

4.1 Effectiveness

Effectiveness of the classifier can be obtained on various parameters like accuracy, correctly classified instances, and incorrectly classified instances. Accuracy can be derived as:

$$Accuracy = (TP + TN) / (TP + TN + FP + FN) \quad (6)$$

Where TP denotes True Positive, FP denotes False Positive, TN denotes True Negative and FN denotes False Negative. Table 3 and Figure 1, Table 4 and Figure 2 shows the performance of the classifiers with and without feature selection.

Table 3. Results for effectiveness without feature selection

Evaluation parameters	J48	KNN	Logistic Regression	Bayesian Network
Correctly classified instances	530	547	538	542
Incorrectly classified instances	39	22	31	27
Accuracy (%)	93.14	96.13	94.55	95.25

Training and simulation parameters also play an important role in order to assess the classifiers performance. Various parameters of training and simulation with and without feature selection are shown in Table 5 and Table 6, Figure 3, Figure 4, Figure 5 and Figure 6.

Figure 1. Comparison Graph on the Parameters of Table 3

Table 4. Results for effectiveness with feature selection

Evaluation parameters	J48	KNN	Logistic Regression	Bayesian Network
Correctly classified instances	552	554	556	551
Incorrectly classified instances	17	15	13	18
Accuracy (%)	97.01	97.36	97.71	96.83

Figure 2. Comparison Graph on the Parameters of Table 4

Table 5. Results for Training and Simulation without feature selection

Evaluation parameters	J48	KNN	Logistic Regression	Bayesian Network
Kappa Statistic(numeric)	0.85	0.91	0.88	0.89
Mean Absolute Error(numeric)	0.07	0.040	0.05	0.05
Root Mean Square Error in (%)	0.25	0.19	0.23	0.21
Relative Absolute Error in (%)	15.83	8.65	11.85	11.53
Root Relative Square Error in (%)	53.23	40.59	48.35	45.08

Figure 3. Comparative graph of the classifiers on the basis of Kappa Statistic, Mean Absolute Error and Root Mean Square Error without feature selection.

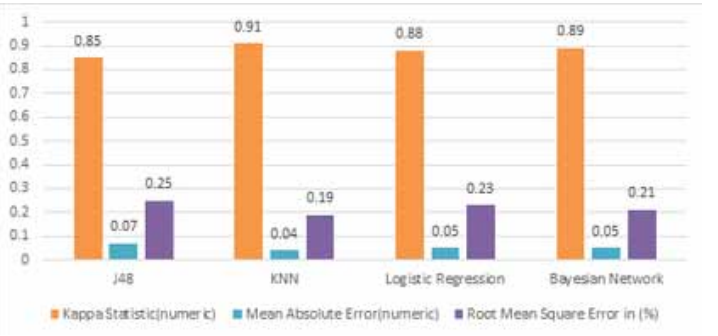


Figure 4. Comparative graph of the classifiers on the basis of Relative Absolute Error, Root Relative Square Error without feature selection.

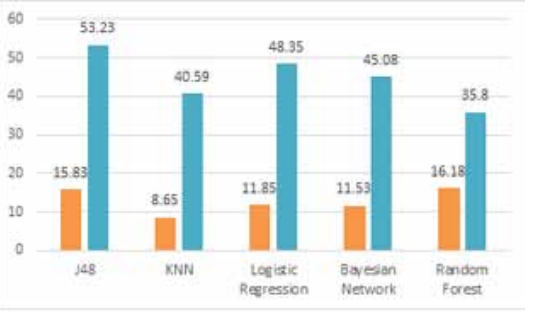


Table 6. Results for Training and Simulation with feature selection

Evaluation parameters	J48	KNN	Logistic Regression	Bayesian Network
Kappa Statistic(numeric)	0.93	0.94	0.95	0.93
Mean Absolute Error(numeric)	0.04	0.02	0.04	0.03
Root Mean Square Error in (%)	0.17	0.16	0.14	0.15
Relative Absolute Error in (%)	8.81	6.03	9.39	8.22
Root Relative Square Error in (%)	35.57	33.51	30.08	33.08

Figure 5. Comparative graph of the classifiers on the basis of Kappa Statistic, Mean Absolute Error and Root Mean Square Error with feature selection.

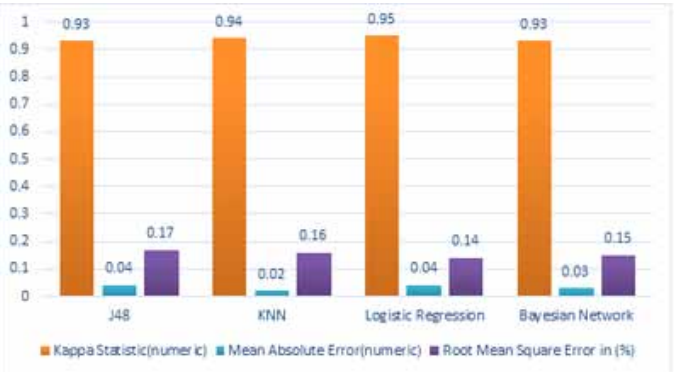
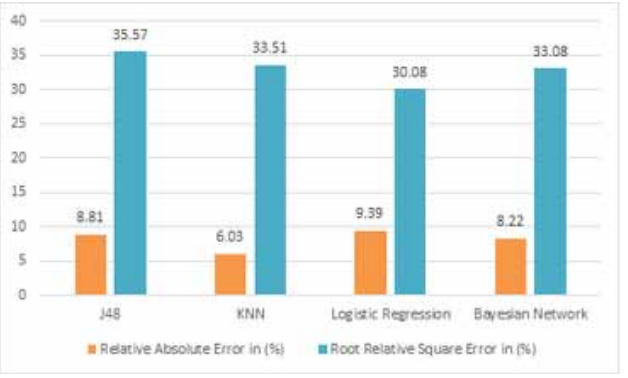


Figure 6. Comparative graph of the classifiers on the basis of Relative Absolute Error, Root Relative Square Error with feature selection.



The output from the training and simulation parameters reveals that there is significant effect of feature selection over the performance of the classifiers in terms of reduction in the error rate.

Table 7 reveals the results of confusion matrix for the classification algorithms without feature selection that we have used in our study. Confusion matrix is helpful in providing an abstract view of results which was predicted by the classification algorithm. In table 7 the prediction results were obtained on the basis of two classes class a and b which specifies the Malignant and Benign stages of breast cancer. Here for J48 there are 195 cases of True Positive, 17 of False Negative, 22 of False Positive and 335 of True Negative. Similarly, for K-NN there are 200 cases of True Positive, 12 cases of false Negative, 10 cases of False Positive and 347 of True Negative. For Logistic Regression there are 200 cases of True Positive, 12 of False Negative, 19 of False Positive and 338 of True Negative. For Bayes Net 198 cases of True Positive, 14 cases of False Negative, 13 of False Positive and 344 of True Negative.

Table 7. Confusion Matrix without feature selection

Classifier	a	B	Class
J48	195	17	a= M
	22	335	b= B
KNN	200	12	a= M
	10	347	b= B
Logistic Regression	200	12	a= M
	19	338	b= B
Bayesian Network	198	14	a= M
	13	344	b= B

Table 8 reveals the results of confusion matrix for the classification algorithms with feature selection that we have used in our study. Here for J48 there are 203 cases of True Positive, 9 of False Negative, 8 of False Positive and 349 of True Negative. Similarly, for K-NN there are 204 cases of True Positive, 8 cases of false Negative, 7 cases of False Positive and 350 of True Negative. For Logistic Regression there are 204 cases of True Positive, 8 of False Negative, 5 of False Positive and 352 of True Negative. For Bayes Net 201 cases of True Positive, 11 cases of False Negative, 7 of False Positive and 350 of True Negative.

Table 8. Confusion Matrix with feature selection

Classifier	a	B	Class
J48	203	9	a= M
	8	349	b= B
KNN	204	8	a= M
	7	350	b= B
Logistic Regression	204	8	a= M
	5	352	b= B
Bayesian Network	201	11	a= M
	7	350	b= B

4.2 Efficiency

Once the predictive model is built its efficiency is to be measured in terms of True Positive Rate (TPR), False Positive Rate (FPR), Recall, Precision and False-Measure and Area under the Receiver operating characteristics (AUROC). In order to assess the efficiency based on these parameters prior analysis of confusion matrix is required which was discussed in Table 9 and Table 10 shows the results for the parameters of efficiency with and without feature selection.

$$Sensitivity = TP \,/\, (TP + FN)\,. \tag{7}$$

$$Specificity = TN \,/\, (FP + TN) \tag{8}$$

$$\mathrm{Re}\,call = TP \,/\, (TP + FN) \tag{9}$$

$$\mathrm{Pr}\,ecision = TP \,/\, (TP + FN) \tag{10}$$

$$F - Measure = 2(P * R) \,/\, P + R \tag{11}$$

Here P= Precision, R= Recall, F= False

Table 9. Efficiency of classifiers without feature selection

Evaluation parameters	J48	KNN	Logistic Regression	Bayesian Network	Class
TPR	0.92	0.943	0.943	0.934	M
	0.93	0.972	0.947	0.964	B
FPR	0.03	0.028	0.053	0.036	M
	0.08	0.057	0.057	0.066	B
Precision	0.89	0.952	0.913	0.938	M
	0.95	0.967	0.966	0.961	B
Recall	0.92	0.943	0.943	0.934	M
	0.93	0.972	0.947	0.964	B
F-Measure	0.90	0.948	0.928	0.936	M
	0.94	0.969	0.956	0.962	B
AUROC	0.92	0.956	0.971	0.986	M
	0.92	0.956	0.973	0.983	B

Table 10. Efficiency of classifiers with feature selection

Evaluation parameters	J48	KNN	Logistic Regression	Bayesian Network	Class
TPR	0.958	0.962	0.962	0.948	M
	0.978	0.980	0.986	0.980	B
FPR	0.02	0.02	0.014	0.020	M
	0.04	0.03	0.038	0.052	B
Precision	0.962	0.967	0.976	0.966	M
	0.975	0.978	0.978	0.970	B
Recall	0.958	0.962	0.962	0.948	M
	0.978	0.980	0.986	0.980	B
F-Measure	0.960	0.965	0.969	0.957	M
	0.976	0.979	0.982	0.975	B
AUROC	0.951	0.968	0.991	0.992	M
	0.951	0.968	0.991	0.992	B

4.3 Comparative Analysis

This section reveals the results of the existing work done by the researchers for predicting the breast cancer using various machine learning techniques. Table 11, shows the comparison of the results obtained from this research work with the existing works.

Table 11. Comparative Analysis with existing work

Authors	Year	Algorithms	Accuracy
(Islam et.al., 2020)	2020	SVM, KNN, RF ANN, LR	97.14%, 97.14%, 95.71%, 98.57%, 95.71%.
(Khairunnahara et.al., 2019)	2019	Logistic Regression with dynamic weighted approach	96.83%.
(Khan et.al., 2020)	2020	BCP-SVM	97.06%
(Banu et.al., 2018)	2018	Boosted Augmented Naïve Bayes, Tree augmented Naïve Bayes and Bayes Belief Network.	91.7%, 94.11%, 97.11%.
(Israni, 2019)	2019	SVM, KNN, LR	92.78%, 92.23%, 92.10%.
(Bayrak et.al., 2019)	2019	ANN, SVM	95.4%, 96.9%
(Bharati et.al., 2018)	2018	Naïve Bayes, RF, LR, MLP, KNN	71.65%, 69.5%, 68.8%, 64.6%, 72.37%.
(Khourdifi & Bahaj, 2018)	2018	KNN, SVM, RF, Naïve Bayes	96.1%, 97.9%, 96%, 92.6%.
Our Study		J48, KNN, LR, Bayesian Network	97.01%, 97.36%, 97.71%, 96.83%.

5. SUGGESTION

This research work is an attempt to improve the accuracy rate of the classification algorithms *viz.* J48, KNN, Logistic Regression and Bayesian Networks and to showcase the importance of feature selection. Here wrapper approach of feature selection is used, which in turns produced a good form of enhanced accuracy. Furthermore, various other methods of feature selection like Principle Component Analysis (PCA), Gain Ratio, Information Gain etc. can be investigated on various other datasets of breast cancer available in the public domain.

6. CONCLUSION

In this research work 4 most relevant machine learning algorithms are used for developing the prediction model along with the wrapper method of feature selection and an attempt is made to improve the prediction accuracy rate (PAR) of the classifiers. Results of the study reveals that there is a significant form of increase in the prediction accuracy rate (PAR) ranges between 2-4% and reduction in the error rate between 2-9%. Hence from the results obtained we have noticed that feature selection techniques play an important role improving the accuracy rate and reducing the error rate, which is useful for various stakeholders of the medical domain for the diagnosis of breast cancer.

FUNDING AGENCY

Publisher has waived the Open Access publishing fee.

REFERENCES

Aalaei, S., Shahraki, H., Rowhanimanesh, A., & Eslami, S. (2016). Feature selection using genetic algorithm for breast cancer diagnosis: Experiment on three different datasets. *Iranian Journal of Basic Medical Sciences.*, *19*(5), 476.

Adnan, N., Liu, Z., Huang, T. H., & Ruan, J. (2020). Comparative evaluation of network features for the prediction of breast cancer metastasis. *BMC Medical Genomics*, *13*(S5), 1–10. doi:10.1186/s12920-020-0676-3 PMID:32241278

Aljawarneh, S., Yassein, M. B., & Aljundi, M. (2019). An enhanced J48 classification algorithm for the anomaly intrusion detection systems. *Cluster Computing*, *22*(5), 10549–10565.

Atoui, M. A., Cohen, A., Verron, S., & Kobi, A. (2019). A single Bayesian network classifier for monitoring with unknown classes. *Engineering Applications of Artificial Intelligence*, *85*, 681–690.

Bayrak, E. A., Kırcı, P., & Ensari, T. (2019, April). Comparison of machine learning methods for breast cancer diagnosis. In 2019 Scientific Meeting on Electrical-Electronics & Biomedical Engineering and Computer Science (EBBT) (pp. 1-3). IEEE.

Bazila Banu, A., & Thirumalaikolundusubramanian, P. (2018). Comparison of Bayes classifiers for breast cancer classification. *Asian Pacific Journal of Cancer Prevention*, *19*(10), 2917.

Bharati, S., Rahman, M. A., & Podder, P. (2018, September). Breast cancer prediction applying different classification algorithm with comparative analysis using WEKA. In *2018 4th International Conference on Electrical Engineering and Information & Communication Technology (iCEEiCT)* (pp. 581-584). IEEE.

Chokka, A., & Rani, K. S. (2019). AdaBoost with Feature Selection Using IoT to Bring the Paths for Somatic Mutations Evaluation in Cancer. In *Internet of Things and Personalized Healthcare Systems* (pp. 51–63). Springer.

Darzi, M., AsgharLiaei, A., & Hosseini, M. (2011). Feature selection for breast cancer diagnosis: A case-based wrapper approach. *International Journal of Biomedical and Biological Engineering*, *5*(5), 220–223.

Dasgupta, S., Chaudhuri, R., & Purushotham, S. (2019). Feature Selection for Breast Cancer Detection using Machine Learning Algorithms. *International Journal of Innovative Technology and Exploring Engineering*, *8*(9), 2080–2083.

Diwakar, M., Patel, P. K., Gupta, K., & Chauhan, C. (2013, December). Object tracking using joint enhanced color-texture histogram. In *2013 IEEE Second International Conference on Image Information Processing (ICIIP-2013)* (pp. 160-165). IEEE.

Fatima, N., Liu, L., Hong, S., & Ahmed, H. (2020). Prediction of Breast Cancer, Comparative Review of Machine Learning Techniques, and Their Analysis. *IEEE Access: Practical Innovations, Open Solutions*, *8*, 150360–150376. doi:10.1109/ACCESS.2020.3016715

Harshvardhan, G., Venkateswaran, N., & Padmapriya, N. (2016, March). Assessment of Glaucoma with ocular thermal images using GLCM techniques and Logistic Regression classifier. In *2016 International Conference on Wireless Communications, Signal Processing and Networking (WiSPNET)* (pp. 1534-1537). IEEE.

Heymach, J., Krilov, L., Alberg, A., Baxter, N., Chang, S. M., Corcoran, R. B., & Burstein, H. (2018). Clinical cancer advances 2018: Annual report on progress against cancer from the American Society of Clinical Oncology. *Journal of Clinical Oncology*, *36*(10), 1020–1044.

Hu, L. Y., Huang, M. W., Ke, S. W., & Tsai, C. F. (2016). The distance function effect on k-nearest neighbor classification for medical datasets. *SpringerPlus*, *5*(1), 1–9.

Islam, M. M., Haque, M. R., Iqbal, H., Hasan, M. M., Hasan, M., & Kabir, M. N. (2020). Breast cancer prediction: A comparative study using machine learning techniques. *SN Computer Science*, *1*(5), 1–14. doi:10.1007/s42979-020-00305-w PMID:33063053

Israni, P. (2019). Breast cancer diagnosis (BCD) model using machine learning. *International Journal of Innovative Technology and Exploring Engineering*, *8*(10), 4456–4463. doi:10.35940/ijitee.J9973.0881019

Khairunnahar, L., Hasib, M. A., Rezanur, R. H. B., Islam, M. R., & Hosain, M. K. (2019). Classification of malignant and benign tissue with logistic regression. *Informatics in Medicine Unlocked*, *16*, 100189. doi:10.1016/j.imu.2019.100189

Khan, F., Khan, M. A., Abbas, S., Athar, A., Siddiqui, S. Y., Khan, A. H., & Hussain, M. (2020). Cloud-based breast cancer prediction empowered with soft computing approaches. *Journal of Healthcare Engineering*, *2020*, 2020. doi:10.1155/2020/8017496 PMID:32509260

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.

Kourou, K., Exarchos, T. P., Exarchos, K. P., Karamouzis, M. V., & Fotiadis, D. I. (2015). Machine learning applications in cancer prognosis and prediction. *Computational and Structural Biotechnology Journal*, *13*, 8–17.

Kumari, B., & Swarnkar, T. (2011). Filter versus Wrapper Feature Subset Selection in Large Dimensionality Micro array: A Review. *International Journal of Computer Science and Information Technologies*, *2*(3), 1048–1053.

Lahoura, V., Singh, H., Aggarwal, A., Sharma, B., Mohammed, M. A., Damaševičius, R., & Cengiz, K. (2021). Cloud Computing-Based Framework for Breast Cancer Diagnosis Using Extreme Learning Machine. *Diagnostics (Basel)*, *11*(2), 241. doi:10.3390/diagnostics11020241 PMID:33557132

Lu, X., Peng, X., Liu, P., Deng, Y., Feng, B., & Liao, B. (2012, August). A novel feature selection method based on CFS in cancer recognition. In *2012 IEEE 6th International Conference on Systems Biology (ISB)* (pp. 226-231). IEEE.

Machine Learning Repository, U. C. I. (2019). *Breast Cancer Wisconsin (Diagnostic)* [Data Set]. https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+%28Diagnostic%29

Mwadulo, M. W. (2016). A Review on Feature Selection Methods For Classification Tasks. *International Journal of Computer Applications Technology and Research*, *5*(6), 395–416.

Napoleon, D., & Pavalakodi, S. (2011). A new method for dimensionality reduction using k-means clustering algorithm for high dimensional data set. *International Journal of Computers and Applications*, *13*(7), 41–46.

Oskouei, R. J., Kor, N. M., & Maleki, S. A. (2017). Data mining and medical world: Breast cancers' diagnosis, treatment, prognosis and challenges. *American Journal of Cancer Research*, *7*(3), 610–627.

Pervez, M. S., & Farid, D. M. (2015). Literature review of feature selection for mining tasks. *International Journal of Computers and Applications*, *116*(21).

Pritom, A. I., Munshi, M. A. R., Sabab, S. A., & Shihab, S. (2016, December). Predicting breast cancer recurrence using effective classification and feature selection technique. In *2016 19th International Conference on Computer and Information Technology (ICCIT)* (pp. 310-314). IEEE.

Rückstieß, T., Osendorfer, C., & van der Smagt, P. (2013). Minimizing data consumption with sequential online feature selection. *International Journal of Machine Learning and Cybernetics*, *4*(3), 235–243.

Sabancı, K., & Koklu, M. (2015). The Classification of Eye State by Using KNN and MLP Classification Models According to the EEG Signals. *International Journal of Intelligent Systems and Applications in Engineering*, *3*(4), 127–130.

Saeys, Y., Inza, I., & Larrañaga, P. (2007). A review of feature selection techniques in bioinformatics. *Bioinformatics, 23*(19), 2507-2517.

Shaikh, T. A., & Ali, R. (2019). Applying machine learning algorithms for early diagnosis and prediction of breast cancer risk. In *Proceedings of 2nd International Conference on Communication, Computing and Networking* (pp. 589-598). Springer.

Yang, L., & Xu, Z. (2019). Feature extraction by PCA and diagnosis of breast tumors using SVM with DE-based parameter tuning. *International Journal of Machine Learning and Cybernetics*, *10*(3), 591–601.

Yildirim, P. (2015). Filter based feature selection methods for prediction of risks in hepatitis disease. *International Journal of Machine Learning and Computing*, *5*(4), 258.