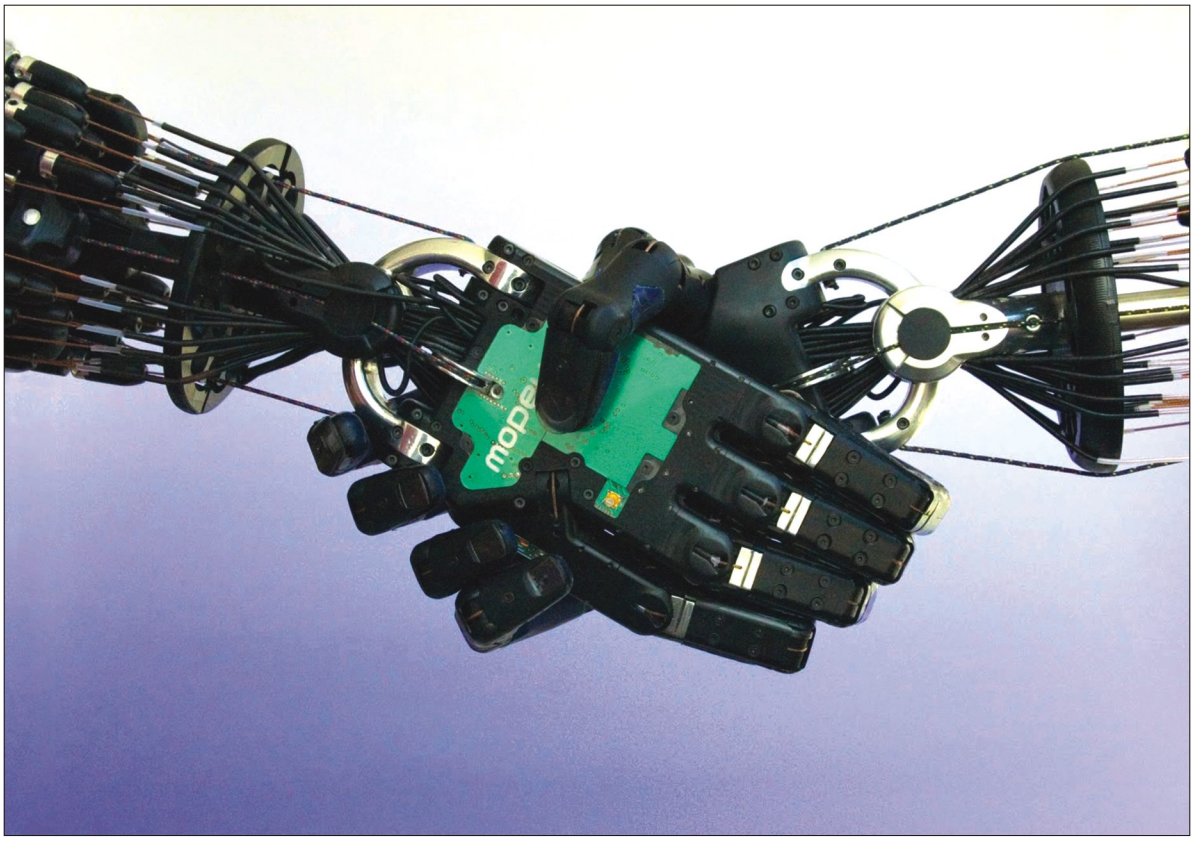

Journal of Artificial Intelligence

ISSN 1994-5450

science alert

ANSInet
an open access publisher
http://ansinet.com

OPEN ACCESS Journal of Artificial Intelligence

ISSN 1994-5450
DOI: 10.3923/jai.2018.55.64

Mini Review
Feature Selection from Biological Database for Breast Cancer Prediction and Detection Using Machine Learning Classifier

Abhineet Gupta and Baij Nath Kaushik

School of Computer Science and Engineering, Shri Mata Vaishno Devi University, Katra-182320, J and K, India

Abstract

Cancer is one of the diagnostic threats appearing to the mankind in this century and among various cancers, breast cancer is the major death causing disease which occurs mainly in women belonging to age between 45 and 60. Early detection and its appropriate treatment can significantly reduce the chances of their death. The objective of this review paper was to study the current systems to develop models with higher classification accuracy for prediction of breast cancer symptoms, their chances of recurrence at the early stage and also their chances of survivability. Here investigation was also done to verify whether comparable accuracy can be achieved even with lesser number of features or not. Initially the feature set is reduced to avoid the over fitting problem and then various machine learning techniques are applied. Here, three different types of feature selection techniques and various machine learning classifiers have been discussed. Further, the comparative analysis among feature selection methods has been done based on their accuracy, computational speed and their dependency on machine learning classifiers. Moreover, the advantages and disadvantages of various classifiers are also discussed. A study of different results from past years have been compared based on the applied classifier, feature selection technique, number of features used and different performance measures like accuracy, sensitivity etc. From different research studies, it is found that comparable accuracy can be achieved even with lesser number of features, which overall reduces the computational complexity of the model. It have discovered that different researchers have found the optimal number of features by hit and trial method which is a very difficult task and to overcome this difficulty, the future scope has been discussed.

Key words: Cross validation, K-nearest neighbour, naive bayes, over fitting, random forest, support vector machine

Citation: Abhineet Gupta and Baij Nath Kaushik, 2018. Feature selection from biological database for breast cancer prediction and detection using machine learning classifier. J. Artif. Intel., 11: 55-64.

Corresponding Author: Baij Nath Kaushik, School of Computer Science and Engineering, Shri Mata Vaishno Devi University, Katra-182320, J and K, India Tel: +91-9654482709

Competing Interest: The authors have declared that no competing interest exists.

Data Availability: All relevant data are within the paper and its supporting information files.

INTRODUCTION

Cancer is a disease in which rapidly cells divide and multiply out of control which causes the formation of a mass of extra tissue[1]. These masses are termed as tumours. Tumours are either malignant (cancerous) or benign (non cancerous). Malignant tumour spreads rapidly and cause damage to its surrounding tissues. A cancer is generally named after the body part where it gets originated. So the breast cancer refers to the malignant tumour due to multiplication of cells in the breast. The main symptoms[2] of breast cancer includes-increase in size or change in shape of the breast, breast pain, swelling of all or part of the breast, differences in the color of the breast skin, a lump in the under arm area etc. According to the World Health organization[3], every year there are about 1.2 million cases of women that are diagnosed with breast cancer. In USA one out of every 8 women is diagnosed with breast cancer.

The physician can also identify the breast cancer manually but it is a difficult process as they have to remember all the information that they require for every particular circumstance which results in low accuracy. Breast cancer deaths can be reduced if it is detected at early stage[4,5]. There are conventional methods for breast cancer detection but machine learning classifiers are getting importance due to its higher accuracy. Now various machine learning techniques are used for its early detection and also to check its recurrence. Some important machine learning techniques are Support vector machine (SVM), Artificial neural network, Naive bayes, Decision trees, Relevance vector machine, K-nearest neighbour, K-means, Random forests etc. The use of these machine learning techniques in building classification systems is getting importance for medical diagnosis. These classification systems can help both experienced and inexperienced experts in minimizing possible errors and also provide the medical data to be examined in short span of time with high accuracy. One of the limitation for effective machine learning classification can be the high dimensionality of the data set. The quality of data and careful feature selection are the important parameters for effective machine learning. Feature selection is NP-Hard problem which is used to select a subset of relevant features from the original feature set.

Once a machine learning classification model is built then the classifiers performance is measured in terms of sensitivity, specificity, accuracy and area under the curve (AUC). The term sensitivity indicates the proportion of true positives that are correctly identified by the classifier[6] whereas specificity indicates the proportion of true negatives that are correctly identified by the classifier. Area under the curve gives us the measure of models performance which is dependent on ROC curve. The ROC curve[7] is a graph that gives the summary of the classifiers performance over all possible thresholds.

The objective of this review is to study the current systems to develop systems with high accuracy to predict breast cancer symptoms, their chances of recurrence at the early stage and also their chances of survivability. Here investigation is also done to verify whether comparable accuracy can be achieved even with less number of features or not for prediction of breast cancer.

FEATURE SELECTION TECHNIQUES

The aim of the classification model is to predict the breast cancer occurrence with high precision and accuracy. When the number of features are large in number then it causes over fitting problem. So, the feature selection algorithms are used to remove the redundant and irrelevant features from the original feature set which will avoid over fitting[8] and hence causes an improvement in the accuracy of the classification model. Moreover, this feature selection will reduce the complexity of the classification model both in terms of time and space[9]. Following are the different feature selection methods that are used to identify the contribution of each feature:

- Filter methods
- Wrapper methods
- Embedded methods

Comparison among three feature selection methods based on their accuracy, computational speed and their dependence on learning classifier is shown in Table 1.

Table 1: Comparison between different feature selection methods based on their accuracy, computational speed and their dependence on learning classifier

Model	Accuracy	Computational speed	Dependence on learning classifier	Examples
Filter	Comparatively low	High (due to the use of some mathematical evaluation function like correlation)	No	Information gain, euclidean distance, chi-square, correlation based feature selection, t-test
Wrapper	Medium	Medium(due to repeated learning and cross validation)	Yes	Genetic algorithms, sequential feature selection, sequential backward selection, randomized hill climbing
Embedded	High	Low	Yes	Decision trees, weighted naive bayes, weighted vector of support vector machine

DIFFERENT MACHINE LEARNING TECHNIQUES

Machine learning is considered as a branch of artificial intelligence where a variety of probabilistic, statistical and optimization tools are employed that learn from past examples and then that prior training is used for classification of new data or for identification of new patterns. There are mainly three different types of learning:

- **Supervised learning[10]:** In this type of learning there are input variables, output variables and algorithm that learns the mapping function from input variable to the output variable. Here it involves predefined output classes. If the output variable is expressed in terms of some classes, then it is called classification problem (when output is category like disease and no disease). Alternatively, if the output variable expressed is continuous then it is called regression problem (like weight). Various examples of supervised learning are K-nearest neighbors, decision tree, support vector machine, naive bayes etc
- **Unsupervised learning[11]:** In this type of learning there are input variables available but not output variables. There are no predefined output classes and the system has to discover pattern or output classes on its own. Examples of unsupervised learning includes k means, K medoids for clustering problems
- **Reinforcement learning:** Here the agent interacts with the environment to maximize the reward as each agent's action is associated with some reward or punishment. Here reward is given for right action and punishment is given for wrong action
- There are some factors that are to be considered before the selection of a particular machine learning algorithm-dimension of the features, number of training samples, over fitting can take place or not, features are independent or not, processing speed, accuracy in terms of performance, memory usage etc

Important machine learning techniques include support vector machine, artificial neural network, naive bayes, decision trees, K-nearest neighbour, random forests etc:

- **Artificial neural network:** The ANN is a network of non linear self adaptable, parallel computing neurons which are used to simulate the computing functionalities of the human brain[12-16]. Here each connection is associated with some weight. Processing of records is done on the training data using the weights and functions of the hidden layer and then the comparison is done between the desired output and the resulting output[17]. Back propagation of errors is done iteratively and then finally the weights are adjusted for the next input record. Interpretation of knowledge that is acquired in the form of network of units is connected via weighted links, which ultimately makes it a difficult task. Machine learned internal decision structure is difficult to understand by humans (black box structure). The ANN may suffer from over fitting problem because of its tendency to adapt themselves too much of data
- **Naive Bayes:** This classification technique is also known by other names-Bayesian belief network, probabilistic network and causal network. It is a probabilistic classifier based on applying Bayes theorem and here it is assumed that the attributes are statistically independent i.e., for a given class tuple, effect of one attribute value is independent of the values of the other attributes, which ultimately simplifies the computation

  However, there exists dependencies or conditional probabilities that have predecessors. Here the node of a graph represents the variable and the arc represents the probabilistic relationship among the variables. So because of their graphical representation, they are easy to interpret. In Bayesian network probability values between nodes reflect the degree of dependence between nodes. They are used in medical domain where the symptoms have dependency among them. For example higher the obesity, higher the chances of various diseases
- **Support vector machine:** It is one of the most powerful machine learning classification technique in terms of its accuracy and has the ability to model complex non linear boundaries. It is less prone to over fitting and by the use of appropriate kernel, they are considered to work well even when the data is not linearly separable. This method tries to find a hyper plane that separates the outcomes of two classes along with the aim of finding maximum distance to the closest point of two output classes[18]. The SVMs are widely used in bioinformatics and text classification problems
- **Decision tree:** It is a tree structure where each non leaf node represents a test on an attribute, each branch represents an outcome of the test and each leaf node represents an output class. The prerequisite condition is that they must have mutually exclusive classes. ID3, C4.5,

C5 and CART are some of the important decision trees which acquire a greedy non backtracking technique[19] in which decision trees construction follows top-down recursive divide and conquer strategy for improving the prediction accuracy. When a decision tree is built many of its branches may reflect outliers or noise in the training input data then tree pruning is used to remove such branches after its identification which will ultimately lead to its improved classification accuracy on the unseen data. Here the attribute selection measures like gain ratio, information gain and gini index for the selection of that attribute that discriminates the given tuples in least amount of time by using least number of splits. The main advantage of these decision trees is that they are very easy to interpret

The main disadvantage of decision tree is that easily tend to over fit, so it gives rise to new class called ensemble methods[20] like Random Forests, Bagging and Boosting where they avoid over fitting. The random forest is considered even better than SVM in terms of its speed and scalability. When compared with decision trees, random forests also have low classification error. They work well even when they have data with missing variables. However, advantages and disadvantages of different machine learning techniques (Table 2).

- Some of the other machine learning techniques are logistic regression, relevance vector machine, K-means, KNN, extreme learning machine etc

Based on different number of features, feature selection technique, classifier data set, the results have been concluded in Table 3. Different results have shown that comparable accuracy can be achieved even with lesser number of features for prediction of breast cancer in lesser computational time.

Comparison of the accuracy of three different classification techniques namely Naive Bayes C4.5 SVM and decision tree is done for prediction of cancer recurrence[21]. In order to remove the redundant and irrelevant attributes, information gain attribute eval is selected for feature selection for c4.5 decision tree and naive bayes whereas SVM attribute eval (attributes are ranked by the square of the weight assigned by the SVM) is used as feature selection for SVM classifier. Various results have shown that SVM is better than other two classification techniques both after and before feature selection. Moreover, it is found that maximum efficiency is achieved when the best 11 features are selected for SVM, 10 attributes for C4.5 decision tree and best 8 attributes for Naive bayes.

In another work[22] four different classification techniques-C4.5 decision tree, SVM, k-NN and Naive bayes are compared based on their accuracies and time to build the model for breast cancer detection. Based on results, it is found that SVM provides the highest accuracy with least error but at the cost of highest computational cost of 0.7s to build the model whereas k-NN takes only 0.1 s to build the model.

The SVM classification for breast cancer detection achieves accuracy of 99.51 when it is applied on selected 5 features based on F-Score feature selection technique[23].

The accuracies of 8 different classification techniques namely C5.0 decision tree, SVM, naive bayes, KNN, fuzzy c means, PAM, K means and EM (Expectation Maximization) are compared to predict the cancer recurrence[24]. Various results have shown that decision tree C5.0 and SVM are the best predictors with accuracy is 81% whereas fuzzy c means gives the lowest accuracy with 37%.

In another paper, a comparison of three machine learning techniques-random forests, SVM and Bayesian networks is done for breast cancer detection[25]. Results have shown that random forests gives the optimum ROC performance and in terms of recall and precision, Bayesian network performs better.

Three machine learning classification techniques-ANN, C.5 decision tree and one statistical method called logistic regression are used for prediction of breast cancer survivability[26]. The results have confirmed that C.5 decision tree predicts with the highest accuracy of 93.6%, then ANN with 91.2% and logistic regression with 89.2% accuracy. All these three classification techniques have used 10-fold cross validation. Also to know the contribution of each variable in cancer survivability, sensitive analysis on ANN model is conducted.

The author has used initially PCA feature selection technique that has reduced the number of variables from 14-5, that captures about 98% variance of the original data and then logistic regression is used as classification technique for breast cancer detection[1,27]. The accuracy measured is 92.9% when all the 14 variables are considered for breast cancer detection and 92.4% when only 5 variables are considered which indicates that no significant difference exists in their accuracies.

Table 2: Advantages and disadvantages of various machine learning techniques

Classification techniques	Advantages	Disadvantages
Artificial neural network	• Ability to tolerate noisy input due to data generalization • Results comparatively good when little knowledge of data set • Can be used when there exists a complex, non linear pattern relationship between input and output • Reduces error that is associated with human error	• Convergence of network not possible when there is not enough training data (as the learning process is not complete) • Network structure difficult to understand by humans (black box) • Over fitting can be caused due to use of many attributes • Training takes a lot of time and training data • Finding the correct topology is difficult
Naive Bayes	• Simple representation • Ability to work with incomplete data • Provide explanations for their decisions • Simple model (easy to understand because of its graphical representation) • Works better even with lesser data • Order of instances has no effect on training • Robust to over fitting	• Redundancy of attributes will mislead classification
Support vector machine	• High accuracy • With the help of non linear kernel it can model non linear class boundaries • Results better when there is high dimensional data • Less prone to over fitting	• Network structure difficult to understand by humans (black box) • Raining is slow as compared to naive bayes and decision tree • Finding the correct kernel is a challenge
Decision tree	• Easy to interpret	• Easily over fit • Depend on order of attribute selection • Computational intensive (because of the training data set needs to be traversed repeatedly)
K-Nearest neighbor	• Works well for small number of dimensions • Robust to noisy training data (especially when we use inverse square of weighted distance as the distance measure) • Works well when the training data set is large	• Consumes lot of memory for execution (because of the storage of all the instances) • Performance gets degrade when high number of dimensions are used • Value of K needs to be determined (i.e., number of nearest neighbors) • High computation cost because of it's needs to compute distance for every query instance for all the training sample data. Also indexing like K-D tree may reduces this computation

Table 3: Recent research trends for prediction of breast cancer detection based on number of features used

Publication (Year)	Title of the paper	Technologies used	Number of features used	Results	Conclusion	Dataset used
Applied soft computing, SCI. 2018	"WCBA: Weighted classification based on association rules algorithm for breast cancer disease"	Weighted classification based on association rules algorithm (the features are ranked by experts)	9	Confidence is set to 0.5 for all three cases when support=0.1, accuracy is 69.77 for recurrence and 97.4 for diagnosis. When support = 0.2, accuracy is 73.26 for recurrence and 97.4 for diagnosis. When support = 0.3, accuracy is 70.93 for recurrence and 96.8 for diagnosis	WCBA gives better accuracy when compared with other association classification algorithms like-Classification Based on Associations (CBA), Classification based on Multiple Association Rules (CMAR), Multi-class classification based on association rule (MCAR), Fast Associative Classification Algorithm(FACA) and (Enhancement classification based on Association rule) ECBA	Wisconsin Diagnostic Breast Cancer (WDBC) for Recurrence and Diagnosis
Egyptian informatics journal, Elsevier. 2017	"Gene expression based cancer classification"	Backward elimination hilbert-schmidt independence Criterion (BAHSIC), Extreme value distribution (EVD) and singular value decomposition entropy are used for feature selection KNN as classifier	50, 5, 25 are number of input genes for BAHSIC1, BAHSIC2, BAHSIC3, respectively	Area under curve 0.99 for colon data set,1.00 for leukemia and 1.00 for breast cancer	Ensemble of these techniques performs better than individual systems	Leukemia, colon dataset, breast cancer dataset
Telematics and informatics, 2017	"A knowledge-based system for breast cancer classification using fuzzy logic method"	Hybrid of expectation Maximization, principle Component analysis, classification and regression trees and fuzzy rule-based	All	Accuracy 93.2% for WDBC and 94.1% for Mammographic mass datasets	EM (Expectation Maximization) -PCA (principle component analysis) -classification and regression trees (CART)-Fuzzy Rule-based has higher accuracy than principle component Analysis -K nearest neighbour, principle component analysis-support vector machine and decision tree	Wisconsin diagnostic breast cancer (WDBC) and Mammographic mass datasets
Procedia computer science, Elsevier. 2016	"Using machine learning algorithms for breast cancer risk prediction and diagnosis"	Support vector machine, decision tree (C4.5), Naive bayes and k nearest neighbour	All	Accuracy of SVM is 97.13%, decision tree (C4.5) is 95.13%, Naive Bayes is accuracy 95.99% and of k-NN is 95.27%	Support vector machine gives the highest accuracy (97.13%) with lowest error rate. Support vector machine takes about 0.07 s to build its model unlike that takes just 0.01 s	Wisconsin diagnostic breast cancer (WDBC)
19th International Conference on Computer and Information Technology, 2016	"Predicting breast cancer recurrence using effective classification and feature selection technique"	Ranker algorithm for feature selection, Naive Bayes, C4.5 decision tree and support vector machine as classification algorithms	11 for sequential minimal optimization, 8 for Naïve bayes and 10 for C4.5 decision tree	Support vector machine prediction accuracy = 75.75%, Naïve Bayes prediction accuracy (67.17%) and C4.5 prediction accuracy (73.73%)	Results have shown that support vector machine (with ranker algorithm) has higher prediction accuracy than Naïve Bayes (with ranker algorithm) and C4.5 (with ranker algorithm)	Wisconsin diagnostic breast cancer (WDBC)
Expert systems with applications, elsevier. 2015	"Breast cancer classification using deep belief networks"	Back-propagation neural network with Liebenberg Marquardt learning function while weights are initialized from the deep belief network path (DBN-NN)	All	Accuracy of deep belief network -neural network 99.68% with 100% sensitivity and 99.47% specificity	The classifier gives an accuracy of 99.68% indicating promising results over previously published studies	Wisconsin diagnostic breast cancer (WDBC)
Expert systems with applications, elsevier. 2014	"Breast cancer diagnosis based on feature extraction using a hybrid of K-means and support vector machine algorithms"	Hybrid of K-Means and support vector machine	6	Accuracy 97.38%, CPU time (in sec) for K-support vector machine (with 6 features) is 0.0039 and for support vector machine (with 30 features) is 15.8913	K-support vector machine with 6 features gives higher accuracy when compared with Ant colony optimization -support vector machine with 15 features, genetic algorithm-support vector machine with 18 features and Particle Swarm optimization-support vector machine with 17 features	Wisconsin diagnostic breast cancer (WDBC)

Table 3: Continue

Publication (Year)	Title of the paper	Technologies used	Number of features used	Results	Conclusion	Dataset used
Journal teknologi, 2013	"Cancer detection using artificial neural network and support vector machine: A comparative study"	Artificial neural network, support vector machine	All	Breast cancer (support vector machine accuracy = 99.51, artificial neural network accuracy = 98.54) liver cancer (SVM accuracy = 63.11, Artificial neural network accuracy = 57.28) Prostate cancer (support vector machine accuracy = 78.35, Artificial Neural Network accuracy = 82.47) ovarian cancer (Support vector machine accuracy = 64.47, Artificial Neural Network accuracy = 78.95)	Artificial neural network classifier can obtain good classification performance in the dataset with larger number of input features (prostate and ovarian cancer dataset)	Breast cancer and liver cancer dataset (UCI machine library database) Prostate cancer dataset and ovarian cancer dataset (JNCI Data)
Expert systems with applications, elsevier. 2009	"Support vector machines combined with feature selection for breast cancer diagnosis"	F-score, support vector machine	5	Support vector machine accuracy (99.51%)	Highest classification accuracy (99.51%) is obtained for the support vector machine model that contains five features	Wisconsin diagnostic breast cancer (WDBC)
IEEE, 2009	"Machine learning techniques to diagnose breast cancer"	Signal-to-noiseratio feature ranking, sequential forward selection-based feature selection and principle component analysis support vector machine, K Nearest neighbour and probabilistic neural networks classifiers are used	10	Without feature selection accuracy support vector machine -Poly-97.09%, Support Vector Machine -radial basis Function-98.80%, K Nearest Neighbour-93.37%, probabilistic neural network-97.23% With feature selection accuracy, support vector machine Poly-95.00%, support vector machine radial basis function -96.33%, K nearest neighbour-88.45%, probabilistic neural network -93.39%	The best overall accuracy for breast cancer diagnosis is achieved equal to 98.80% by support vector machine -radial basis function (without feature selection) and 96.33% by support vector machine - radial basis function (with feature selection) respectively	Wisconsin diagnostic breast cancer (WDBC)
Artificial intelligence in medicine, elsevier. 2005	"Predicting breast cancer survivability: a comparison of three data mining methods"	Artificial neural network, decision trees and logistic regression	All	Accuracy for decision tree (C5) is 93.6% Accuracy for artificial neural network is 91.2% Accuracy for logistic regression is 89.2%	The results indicated that the decision tree (C5) is the best predictor with 93.6% accuracy on the holdout sample (this prediction accuracy is better than any reported in the literature), artificial neural networks came out to be the second with 91.2% accuracy and the logistic regression models came out to be the worst of the three with 89.2% accuracy	Breast cancer web site (http://www. seer.cancer.-gov)

Different results have confirmed that after feature reduction RVM which has low computational cost gives better results than other machine learning techniques like Naive Bayes, neural networks and fuzzy[28].

The author has used SVM-RBF, SVM-Poly, K-Nearest Neighbour and probabilistic neural network (PNN) are used along with PCA (where feature set reduced to 10 features from 70 features) and sequential forward selection (feature set reduced to 25 from 70 features) for breast cancer detection[29]. Various results have indicated that SVM-RBF is better than all other three classification techniques in all the cases i.e., in case when no feature selection is used (accuracy of 98.80%), in case of PCA (accuracy of 95.01%) or in case of sequential forward selection (accuracy of 96.33%).

Extreme learning machine for breast cancer detection is used and the results have indicated that ELM that has 20 nodes gives better accuracy than SVM with lesser computational cost[30]. Results also have indicated that with lesser number of hidden nodes in ELM, the average success rate is supposed to be very low but there is one advantage of lesser computational cost. As the number of nodes are increased average success rate starts increasing and when number of nodes reaches 20, it gives the best accuracy of 93%. Five different feature selection techniques are used to select genes[31]. The first three are based on BAHSIC algorithm which takes 50, 5, 25 number of input genes respectively and the other two are EVD and SVD Entropy. Here KNN is used as classifier which gives AUC 0.99 for colon data set,1.00 for leukemia and 1.00 for breast cancer.

In another work[32] back-propagation neural network is used with Liebenberg Marquardt learning function and weights are initialized from the deep belief network. The classifier gives an accuracy of 99.68% which is better than previously published results. It also gives 100% sensitivity and 99.47% specificity.

Nine features are ranked by experts by giving different weights and then based on different association rules classification is done[33]. This procedure gives better accuracy when compared with other association classification algorithms like-Classification Based on Associations (CBA), Classification based on Multiple Association Rules (CMAR), Multi-class classification based on association rule (MCAR), Fast Associative Classification Algorithm (FACA) and (Enhancement classification based on Association rule) ECBA. Confidence is set to 0.5 for all cases. In first case when support is set to 0.1, accuracy is 69.77 for recurrence and 97.4 for diagnosis. In second case when support = 0.2, accuracy is 73.26 for recurrence and 97.4 for diagnosis and when support = 0.3, accuracy is 70.93 for recurrence and 96.8 for diagnosis.

Here hybrid of EM (Expectation Maximization), PCA (Principle Component Analysis), Classification and Regression Trees (CART) and Fuzzy Rule-Based is used for classification[34]. Various results show that the PCA-EM-CART-Fuzzy Rule-Based has greater accuracy than PCA-KNN, PCA-SVM and Decision Tree. Hybrid of EM-PCA-CART-Fuzzy gives accuracy of 93.2% for WDBC and 94.1% for mammographic mass datasets.

CONCLUSION AND FUTURE SCOPE

From different research studies, it was found that comparable accuracy can be achieved even with less number of features for prediction of breast cancer. Only the features that are selected by a particular feature selection technique will be the input for machine learning classifier, which overall reduces the computational complexity of the model. So future study includes the investigation to check whether number of features to be selected depends on factors like data set, standard deviation, correlation etc. or not. Moreover as a future direction, proposed to use some hybrid machine learning classifiers based on deep learning and extreme learning classifiers to compare and show the effectiveness of proposed algorithms. Further, proposed to use nature inspired algorithm for feature reductions and smooth identification of cancer from biological database.

SIGNIFICANCE STATEMENT

It is a fact that machine learning classification models cannot replace doctors but these models would help in minimizing possible errors which may be committed by the inexperienced doctors. These classification systems examine the detailed medical data in lesser time. Current review compared the feature selection methods which help in removing the redundant attributes thereby reducing the computational cost. This study discovered that different researchers have found the optimal number of features by hit and trial method which is very difficult task. Fixing the number of features in advance may further degrade the performance.

REFERENCES

1. Rull, G., 2017. What is cancer/causes/diagnosis/types. https://patient.info/health/cancer
2. Breast Cancer Care, 2018. Signs and symptoms of breast cancer. https://www.breastcancercare.org.uk/information-support/have-i-got-breast-cancer/signs-symptoms-breast-cancer
3. Breastcancer.ORG, 2018. U.S. breast cancer statistics. https://www.breastcancer. org/symptoms/ understand_bc/statistics

4. World Health Organisation, 2018. Breast cancer: Prevention and control. http://www.who.int/cancer/detection/breastcancer/en/index1.html
5. National Breast Cancer Foundation, 2018. Breast cancer facts. https://www.nationalbreastcancer.org/breast-cancer-facts
6. Kourou, K., T.P. Exarchos, K.P. Exarchos, M.V. Karamouzis and D.I. Fotiadis, 2015. Machine learning applications in cancer prognosis and prediction. Comput. Struct. Biotechnol. J., 13: 8-17.
7. Fawcett, T., 2006. An introduction to ROC analysis. Pattern Recognit. Lett., 27: 861-874.
8. Wang, J.H., J.H. Jiang and R.Q. Yu, 1996. Robust back propagation algorithm as a chemometric tool to prevent the overfitting to outliers. Chemometrics Intell. Lab. Syst., 34: 109-115.
9. Vanaja, S. and K.R. Kumar, 2014. Analysis of feature selection algorithms on classification: A survey. Int. J. Comput. Applic., 96: 29-35.
10. Feng, D., F. Chen and W. Xu, 2014. Supervised feature subset selection with ordinal optimization. Knowledge-Based Syst., 56: 123-140.
11. Elghazel, H. and A. Aussem, 2015. Unsupervised feature selection with ensemble learning. Mach. Learn., 98: 157-180.
12. Kaushik, B. and H. Banka, 2015. Performance evaluation of Approximated Artificial Neural Network (AANN) algorithm for reliability improvement. Applied Soft Comput., 26: 303-314.
13. Kaushik, B. and H. Banka, 2014. Approach for improving reliability in optimal network design. Int. J. Adv. Intell. Paradigms, 6: 157-175.
14. Kaushik, B., N. Kaur and A.K. Kohli, 2013. Achieving maximum reliability in fault tolerant network design for variable networks. Applied Soft Comput., 13: 3211-3224.
15. Kaushik, B., N. Kaur and A.K. Kohli, 2013. Improved approach for maximizing reliability in fault tolerant networks. J. Adv. Comput. Intell. Intell. Inform., 17: 27-41.
16. Kaushik, B., N. Kaur and A.K. Kohli, 2015. Improved neural approach in maximising reliability for increasing networks. Int. J. Comput. Sci. Eng., 11: 176-185.
17. Ubaidillah, S.H.S.A., R. Sallehuddin and N.A. Ali, 2013. Cancer detection using artificial neural network and support vector machine: A comparative study. J. Teknol., 65: 73-81.
18. Zheng, B., S.W. Yoon and S.S. Lam, 2014. Breast cancer diagnosis based on feature extraction using a hybrid of K-means and support vector machine algorithms. Expert Syst. Applic., 41: 1476-1482.
19. Han, J., M. Kamber and J. Pei, 2014. Classification: Basic Concepts. In: Data Mining Concepts and Techniques, Han, J., M. Kamber and J. Pei (Eds.). 3rd Edn., Elsevier, Waltham, USA., pp: 332-335.
20. Sujatha, G. and K. Usha Rani, 2013. An experimental study on ensemble of decision tree classifiers. Int. J. Applic. Innov. Eng. Manage., 2: 300-306.
21. Pritom, A.I., M.A.R. Munshi, S.A. Sabab and S. Shihab, 2016. Predicting breast cancer recurrence using effective classification and feature selection technique. Proceedings of the 19th International Conference on Computer and Information Technology, December 18-20, 2016, Dhaka, Bangladesh, pp: 310-314.
22. Asri, H., H. Mousannif, H. Al Moatassime and T. Noel, 2016. Using machine learning algorithms for breast cancer risk prediction and diagnosis. Procedia Comput. Sci., 83: 1064-1069.
23. Akay, M.F., 2009. Support vector machines combined with feature selection for breast cancer diagnosis. Expert Syst. Applic., 36: 3240-3247.
24. Ojha, U. and S. Goel, 2017. A study on prediction of breast cancer recurrence using data mining techniques. Proceedings of the 7th International Conference on Cloud Computing, Data Science & Engineering-Confluence, January 12-13, 2017, Noida, India, pp: 527-530.
25. Bazazeh, D. and R. Shubair, 2016. Comparative study of machine learning algorithms for breast cancer detection and diagnosis. Proceedings of the 5th International Conference on Electronic Devices, Systems and Applications, December 6-8, 2016, Ras Al Khaimah, United Arab Emirates, pp: 1-4.
26. Delen, D., G. Walker and A. Kadam, 2005. Predicting breast cancer survivability: A comparison of three data mining methods. Artif. Intell. Med., 34: 113-127.
27. Hussain, S., N.Z. Quazilbash, S. Bai and S. Khoja, 2015. Reduction of variables for predicting breast cancer survivability using principal component analysis. Proceedings of the IEEE 28th International Symposium on Computer-Based Medical Systems, June 22-25, 2015, Sao Carlos, Brazil, pp: 131-134.
28. Gayathri, B.M. and C.P. Sumathi, 2016. Comparative study of relevance vector machine with various machine learning techniques used for detecting breast cancer. Proceedings of the IEEE International Conference on Computational Intelligence and Computing Research, December 15-17, 2016, Chennai, India.
29. Osareh, A. and B. Shadgar, 2010. Machine learning techniques to diagnose breast cancer. Proceedings of the 5th International Symposium on Health Informatics and Bioinformatics, April 20-22, 2010, Antalya, Turkey, pp: 114-120.
30. Malik, A. and J. Iqbal, 2015. Extreme learning machine based approach for diagnosis and analysis of breast cancer. J. Chin. Inst. Eng., 39: 74-78.

31. Tarek, S., R.A. Elwahab and M. Shoman, 2017. Gene expression based cancer classification. Egypt. Inform. J., 18: 151-159.
32. Abdel-Zaher, A.M. and A.M. Eldeib, 2016. Breast cancer classification using deep belief networks. Expert Syst. Applic., 46: 139-144.
33. Alwidian, J., B.H. Hammo and N. Obeid, 2018. WCBA: Weighted classification based on association rules algorithm for breast cancer disease. Applied Soft Comput., 62: 536-549.
34. Nilashi, M., O. Ibrahim, H. Ahmadi and L. Shahmoradi, 2017. A knowledge-based system for breast cancer classification using fuzzy logic method. Telematics Inform., 34: 133-144.