A Cost Sensitive SVM and Neural Network Ensemble Model for Breast Cancer Classification

Tina Elizabeth Mathew
Department of Computer Science, Government College Kariavattom, India

Article Info

Article history:

Received Jun 18, 2022
Revised Apr 17, 2023
Accepted Apr 28, 2023

Keywords:

Breast cancer
Classification
Cost Sensitive Learning
Neural Networks
Support Vector Machines

ABSTRACT

Breast Cancer has surpassed all categories of cancer in incidence and is the most prevalent form of cancer in women worldwide. The global incidence rate is seen to be highest in the country of Belgium as per statistics of WHO. In the case of developing countries specifically, India, it has overtaken other cancers and stands first in incidence and mortality. Major factors identified as impacting the prognosis and survival in the country is chiefly the late diagnosis of the disease and diverse situations prevailing in different parts of the country including lack of diagnostic facilities, lack of awareness, fear of undergoing existing procedures and so on. This is also true for many other countries in the world. Early diagnosis is a vital factor for survival. The implementation of machine learning techniques in cancer prediction, diagnosis and classification can assist medical practitioners as a supplementary diagnostic tool. In this work, an ensemble model of a polynomial kernel-based Support Vector machines and Gradient Descent with Momentum Back Propagation Artificial Neural Networks for Breast Cancer Classification is proposed. Feature selection is applied using Genetic Search for identifying the best feature set and data sampling techniques such as combination of oversampling and undersampling and cost senstivke learning are applied on the individual Neural Network and Support Vector Machine classifiers to deal with issues related with class imbalance. The ensemble model is seen to show superior performance in comparison with other models producing an accuracy of 99.12%.

Corresponding Author:

Tina Elizabeth Mathew,
Department of Computer Science,
Government College Kariavattom,
Thiruvananthapuram 695581, Kerala State, India
Email: tinamathew04@gmail.com

1. INTRODUCTION

Cancer is considered the deadliest among diseases occurring in humankind. Cancers which are growths, can occur in any portion of the body. Certain cancers like breast cancer or cervical cancer are prevalent specifically in women, even though the former can occur with low likelihood, also in men. Various studies show that they are the leading diseases in most countries like India, with breast cancer showing an upward trend in incidence [1] Breast Cancer manifests chiefly as lumps in the breast which occur mainly due to the uncontrolled growth of cells in the lining, lobules or ducts of the breast tissue. As per the statistics of the World Health Organization it is considered as the most incident and prevalent cancer in women universally [2]. The incidence and mortality rates vary from country, region and ethnicity and the risk of contracting the disease increases as age increases. The most vulnerable and risk prone amongst women are the older population of women. In 2020, 2.3 million women were diagnosed with breast cancer and 685 000 deaths were reported globally [3]]. The Survival rate of breast cancer for at least 5 years after diagnosis also varies from highly developed to lowly developed countries. It ranges from more than 90% in high-income countries, to 66% in India and 40% in South Africa. Breast cancer has surpassed all other cancers prevailing in women in India [4].

The incidence of the disease in India is reported as 25.8 per 100,000 women and mortality at 12.7 per 100,000 women [4] with highest incidence recorded in the state of Kerala. India is also home to the most aggressive form of breast cancer, Triple Negative Breast Cancer [5]. The Indian Council of Medical Research also reports that the death rate due to various cancers is relatively higher in Indian women when compared to their male counterparts[6]. Global studies show that one third of the global breast cancer burden is contributed by US, India and China, collectively[7]

The morbidity and mortality is more when metastasis of the disease sets in and hence rapid and early diagnosis is the key strategy for better prognosis as well as survival[4]. With the advent of state of art medical facilities mortality rate has been much reduced, yet more techniques need to be utilized to aid medical professionals in the rapid identification and treatment of the disease. Besides the medical modalities additional techniques can be used to assist medical practitioners in diagnosis and classification of the disease[8]

Machine learning techniques has been in the fray and has been applied for the past two decades in various classifications tasks in severals domains [9], specifically, the medical domain and for several disease classifications like breast cancer [10], [11]. Many methods like decision trees[8] , k-NN[12], logistic regression[13], and many more have been applied for classification problems. A broad classification of machine learning techniques based on the mode of learning is supervised and unsupervised learning techniques. Classification is an umbrella term that comes under the aegis of supervised techniques and it comprises of a plethora of techniques. Two other categories of supervised classifiers usable in the classification process are Support Vector Machines and Artificial Neural Networks. Literature shows that the application of these techniques to disease classification provides promising results[14]. Neural Networks need large amounts of data for training while SVMs do not. In the case of accuracy, studies show that training using SVMs have smaller standard errors compared to that of Neural Networks. Hence, both can be considered as complementary methods and this concept is made in use for the breast cancer classification problem in this study. The task of the study is to employ an ensemble model comprising of SVM and NN for breast cancer classification. The two classifierss, SVM and ANNs are combined into a voting ensemble model for breast cancer classification and the resultant model is seen to produce better performance with lesser misclassification when compared with the individual models. Besides to improve performance of individual classifiers ensembling or using hybtid models can be a solution. Even though studies in literature show good performance in disease classification by SVMs and ANNs better methods are needed to boost the issues seen. ANNs need a lot of training time. Methods that help to reduce training time as well as boost performance is necessary.

This section provides an overview of some of the related works from the latest scientific articles. SVM is seen to be a most popular as well as robust classifier [15] and artificial neural network is a most frequently used classifier besides being also a very robust classifier. SVMs were used by [10] to identify regions of interest in mammograms and it was performed with 80% accuracy. They concluded that SVMs were very accurate in classifying breast cancer. In their work [16] used various classifiers like k-nn, random forest, ANN and SVM for breast cancer detection from histopathological images and SVM was seen to provide an accuracy of 90%. Besides using simple SVMs, ensemble of SVMS can be used. In their work,[17] used SVM and SVM ensembles for a comparative study and they concluded that linear kernel based SVM ensembles with bagging method and RBF kernel based SVM ensembles with the boosting method are better choices for small scale datasets, with feature selection in the data pre-processing stage, whereas, for large-scale datasets, RBF kernel based SVM ensembles based on boosting were seen to perform better. [18] proposed a cost sensitive SVM ensemble with feature selection and it provided promising results with an accuracy of 98%. [19] used SVM with RBF kernel and Random forests to evaluate breast cancer classification performance with the Boruta feature selection technique and svms were seen to outperform random forests with an accuracy of 95%. In the work proposed by [20] they compared NEAT and backpropagation ANN and obtained an accuracy of 95.8% for breast cancer classification. In their work [21] used scaled conjugate backpropagation ANN and obtained an accuracy of 97.47% for breast cancer classification. Similarly, [22] used conjugate gradient back propagation for breast cancer classification and obtained an accuracy of 97.6%. In all these studies data imbalances was not considered. SVM and NNs do not handle class imbalance well. In the proposed work data imbalance is also considered besides parameter optimization and feature selection.

The paper is organized as follows, Section 1 presents an introduction to the topic, alongwith literature available, while Section 2 deals with the methods applied and materials used. The results and discussions are given in section 3, and Section 4 gives the conclusion followed by references

2. RESEARCH METHOD

The aim of this study is to produce a model with better performance and lesser misclassification of instances. The classifiers used for producing the proposed ensemble model is Support Vector machnies with a polynomial kernel and Neural Networks using Gradient Descent with Momentum Back Propagation method.

2.1. Dataset Used

The Wisconsin breast cancer data set is used in this study. The dataset has 699 instances and 11 attributes, of which the first attribute, the id number, holds no relevance in the work and hence is discarded. The last attribute is the class or target variable which categorizes the instances into two – benign and malignant tumours. The remaining 9 attributes are taken for the study. 16 instances were seen to have missing values hence they were also discarded.

2.2. SVM

Support Vector machines are supervised machine learning methods proposed by Vapnik which now plays a major role with applications primarily in classification and regression. An advantage of SVMs is that they are memory efficient as they use a subset of the training data denoted as support vectors. Literature suggests them as good classifiers in binary classification problems. To function well with non -linear problems they make use of kernel functions. There are many categories of kernel functions, like RBF, linear, sigmoid, and polynomial. In this study, the polynomial kernel used. with SVM was seen to outperform the performance depicted by various other kernels used for breast cancer classification.[23].

2.3. ANN

An artificial neural network functions like human brain. Applications implementing ANNs have increased and it has become the most active research area with extensions into deep learning [24]. ANNs have achieved good performance for classification and diagnosis of breast cancer at early-stage A basic structure of an ANN model consists of 3 layers the input, hidden and output layer. Each layer is interconnected with neurons and each contains an activation function that helps improving the ability to implementand solve nonlinear problems. The working of the model commences from the input layer passing through the hidden layers, finally to the output layer. The final classification result is depicted at the output layer. The number of iterations involved while working varies with the structure and nature of the problem involved.

Back Propagation algorithm is a set of methods that efficiently train artificial neural networks by following a gradient descent approach which are being used in various domains. These are considered as fundamental building blocks of ANNs. There a various categories of backpropagation algorithms, BPNNs make use of a large training time slowing the working of ANNs. Researchers have developed different methods which produce better outcomes. In this study Gradient Descent with momentum method is chosen. This helps the algorithm and the ANN to function faster. This method uses a momentum factor with the Gradient Descent method. The Momentum factor allows the network to aptly respond to the local gradient and latest condition prevailing on the error surface. The advantage of this method is that it provides faster convergence as well as helps the network to ignore and disregard small features in the error surface and henceforth, also prevents the network from getting trapped in a shallow local minimum. The method has a few parameters and identifying their optimal values are of great significance to the performance of the artificial neural network, [25].

2.4. Genetic Search

Genetic Search algorithm is a metaheuristic search algorithm that belongs to the larger class of evolutionary algorithms. This optimization technique is based on the Natural evolution theory of Darwin. It mimics three major biological processes natural selection, gene crossover and mutation. It maintains a set of chromosomes which are considered equivalent to the potential solutions of the investigated problem. The concept of GA lies in obtaining and arriving at the optimal solution, undergoing few steps in evolution, after a few generations. Genetic algorithms are seen to be good optimizers for ANNs, [26], [27], in his work suggested that implementing feature selection with ANNs help to enhance its performance.

The Genetic search process begins with initialization, and an initial population consisting of randomly generated bit strings of candidate solutions are used. The fitness function or the objective function is a significant element that is to be defined. The objective function used here is the difference between the predicted and actual values. At each step, a pool of parents is chosen from the parent population based on the calculated fitness value of each individual. This is done by the selection mechanism. In this work the roulette wheel selection method is applied. The selected parents or individuals have a greater probability to pass on genetic material to the subsequent generations by the way of crossover and mutations. From the selected parents a child population is created and this constitutes the next generation or population.

2.5. Data Imbalance

A major problem and challenge involved in machine learning datasets in several domains is the imbalance of data. Data imbalance implies that the classes available in the dataset do not have a uniform distribution of instances. The number of instances in one class mostly the positive class or the class which represents the disease or the malignant class in this case will be very less compared to the other class which is the negative class or class without breast cancer in this study. The data imbalance can be problematic [28] and this has to be resolved. Many techniques can be applied to solve it. Two broad classifications that can be applied are data-based methods or algorithmic based methods. In this study cost sensitive learning approach, which comes under the category of data sampling, is implemented. It uses a cost matrix. Each instance is given a misclassification cost and for each incorrect classification it is penalized 'n' times the misclassification cost. The objective lies in minimizing the misclassification cost. SVMs are seen to work fine with various datasets but performance deteriorates when data is imbalanced[29]. Hence data balancing is necessary for SVMs. Here a cost sensitive SVM using a cost matrix that penalizes twice for a misclassification is proposed to deal with data imbalance.

2.6. Working of Proposed Model

The proposed model works in three parts.Initially,data preprocessing techniques are applied and then the dataset is partitioned into training and testing datasets. Once this is done parameter optimization, feature selection and data sampling is done in the second part. Resampling techniques are applied for balancing the training data. A combination of oversampling of the minority class and undersampling of the majority class.is applied on the dataset. The balanced data obtained is used with the NN. Feature selection and extraction of the most relevant features are also done. This is done by applying the genetic search algorithm and it selects the best and relevant set of features. The reduced feature set is provided to the artificial neural network using the gradient descent with momentum backpropagation algorithm. For SVMs the cost senstivie learning is applied and a cost matrix as in Table 1 is used. While training with svm, for every misclassifcation of the minority class or in this case, the positive class, that occurs, the classifiers is penalized twice. This is implememted using the cost matrix. The working of the model is depicted in Figure 1

The final part is the voting ensemble of classifiers.The voting ensemble produces the majority label of the predictions as output of the model. When all the classifiers of a voting classifier are independent, and use different methods for training better results are obtained [30] The parameters for the svm polynomial kernel are selected using the grid search technique. Literature suggets that parameter optimization helped classification performance. Hence parameter optimization is incorporated.The best SVM parameters C and gamma obtained by using grid search method and for each parameter a value 1 is seen appropriate. The grid values were vareid from 0.001 to 100. Metaheuristic algorithms are seen to help in better performance during the classification process [31], hence, GA is used for feature selection.. To avoid overestimation of results a training and testing partitionof 80-20 is used for training and testing purposes. The parameters and their values used in the proposed model is shown in table 1.

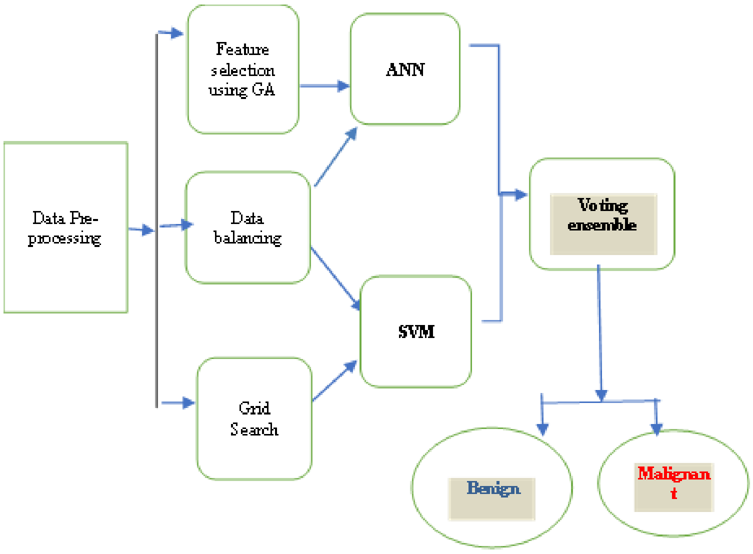


Figure 1. Working of Model

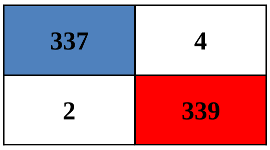
The studies done is [34-36] also suggest using optimized SVM parameters as well as hybrid models for better SVM performance. Using the best optimized paramters is seen to improve performance. Voting is of two types Soft Voting and Hard Voting. Here hard voting is selected. The model with highest votes for performance gets selected when ensembling is done.SVM helps in reducing the time of ANNs which use more time for training models.

Table 1. Parameters and Values

Parameters	Values
Learning rate	0.7
Momentum constant	0.5
No of Hidden neurons	10
C parameter	1
Gamma parameter	1
Cost Matrix used	0 1 2 0
Population size- 20	20
Crossover Probability	0.6
Mutation Probability	0.033

3. RESULTS AND DISCUSSION

The prime objective of the study is to produce a model with better classification accuracy and lesser misclassification of instances. The classification qualityof the experimentation done is measured using various performance measures like ROC-AUC, P-R AUC, Kappa Statistic, F Measure, Recall, Precision, MCC, and FPR. The confusion matrix is computed from True Negatives, True Positives, False Positives and False Negatives. The time taken to build the model was seen to be comparatively lesser than that of the individual neural network. The combined model of SVM and NN helps to counterbalance the issues that each of the individual model have. The performance measures of the proposed model are given in Table 2. The model produced an accuracy of 99,12%. The ROC value obtained is 1. F measure value of 0.991 was obtained. The Matthews Correlation Coefficient (MCC) was 0.982, thus proving the efficiency of the proposed model. The confusion matrix is shown in Fig 2 and misclassification of the true positives was much reduced to 2 and that of the true negatives to 4.



337	4
2	339

Figure 2. Confusion Matrix of Proposed Model

Table 2. Performance Measures of Proposed Model

Method	Proposed -Svm+ Bpnn Model
Accuracy	**99.12**
Kappa	**0.9824**
ROC-AUC	**1**
FPR	**0.009**
F- Measure	**0.991**
MCC	**0.982**
Recall	**0.991**
Precision	**0.991**
Time Taken to Build the Model (Secs)	**13.5**

Figures 3-6 represent the ROC and P-R curve of the two classes malignant and benign class each respectively.The blue lines represent benign class and orange lines denote the malignant class in all diagrams.

The ROC curves for the malignant and benign class of the proposed model are displayed in Fig 3and Fig 4 respectively. The figures of the ROC curve are seen to be along the leftmost and top portion of the graph and it shows the superior performance of the proposed model.

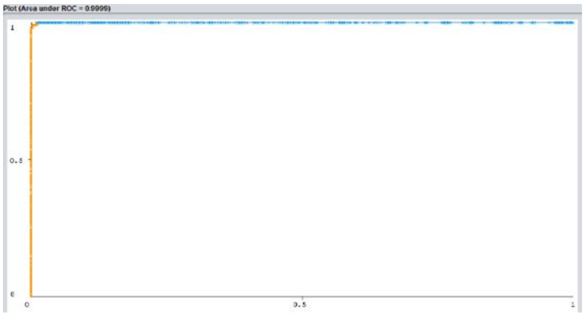


Figure 3. ROC of malignant class

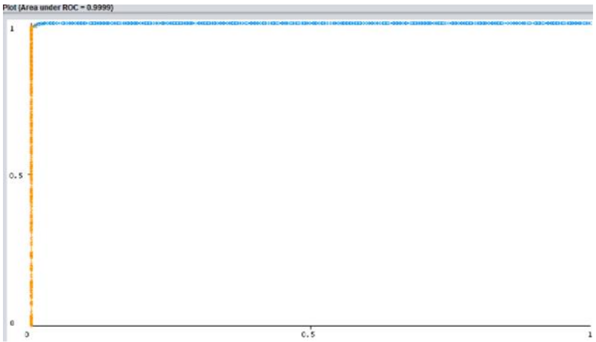


Figure 4. ROC of class benign

The ROC curves can be seen to be closer to the left top edge of the axes and from this it can be perceived that the classifier is illustrating a superior performance. Area under the curve of the ROC helps to measure the overall performance of binary classifier [32]

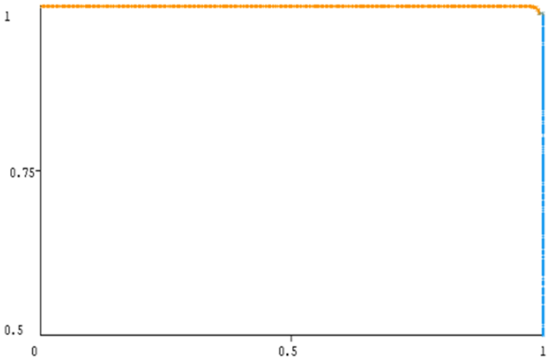


Figure 5 P-R AUC of class malignant

The Precision – Recall Area under the curve of the malignant and benign class of the proposed model is shown in Fig 5 and Fig 6. The curves are seen to be along the top right hand cormner of the graph which represents a good classification performance. The obtained curves highlight the superior performance of the model.

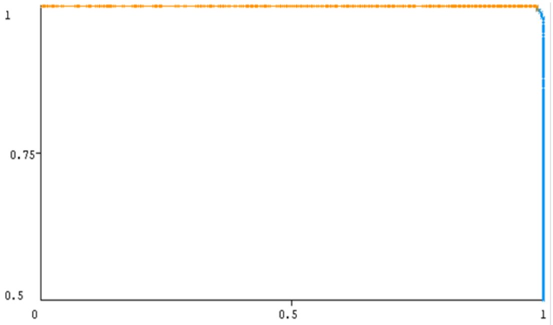


Figure 6 P-R AUC of class benign

A comparative analysis of the proposed ensemble is done with the individual SVM and ANN classifiers and the comparison of accuracy is shown in Fig 7. The ensemble voting model illustrated significant improvement of performance over the individual models. The advantage of the proposed model is that it helped to improve performance of the SVM classifier and also that of the NN classifier.

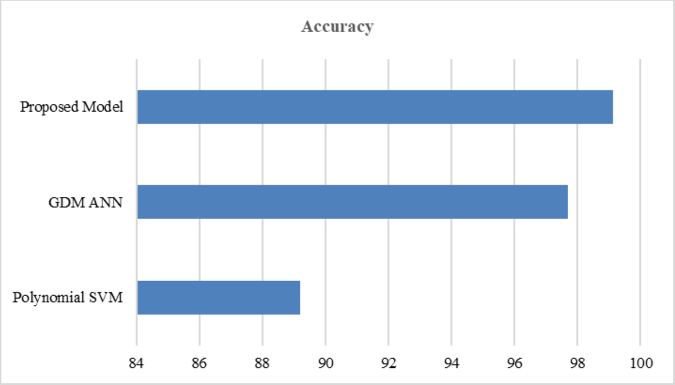


Figure 7. Comparison of Accuracy

The F1 scores of the proposed model with individual SVM and ANN classifiers are also compared and is displayed in Fig **8.** The F1 score of the proposed model is superior to that of the individual models.

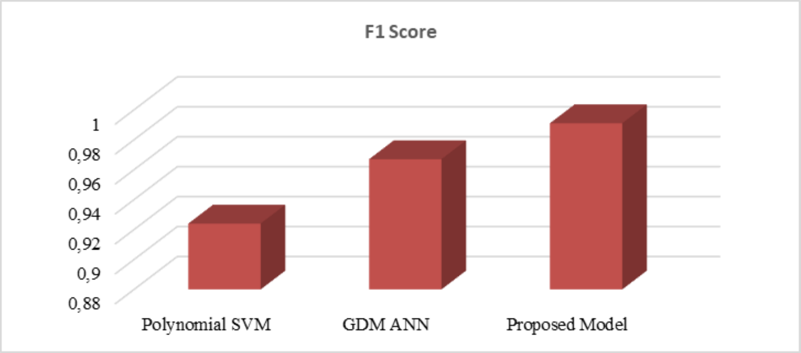


Figure 8. Comparison of F1 score

The proposed model is also compared with two other models found in literature a class balanced Hoeffding Tree model and Random Forest Model using a cost matrix and the proposed model displayed superior performance. Fig 9 displays the comparison of these models. The two models from literature display an accuracy of 97.5% and 97.9% respectively. The proposed model obtained an accuracy of 99.12%.

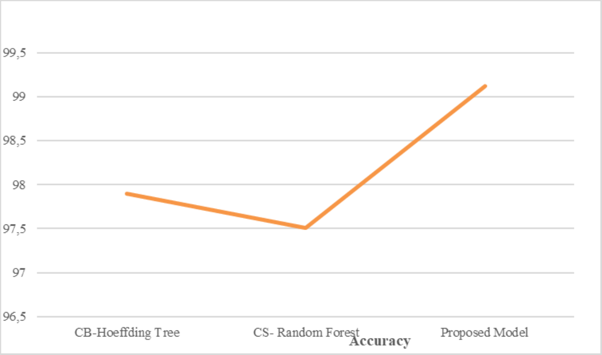


Figure 9. Comparison of Models

Hence, the significant findings of the study are that the proposed model has illustrated high accuracy. The misclassification of the classes has reduced significantly. The AUC of ROC obtained was 1 which shows the performance. High recall precision rates were illustrated by the model. The F measure reveals the sensitivity of the methods employed. The proposed model is compared with SVM and NN individual models [33] by varying kernels of SVM and different training functions for neural networks. Here the accuracy of the models varied within the range of 88%- 92%. [34] in their work highlighted the fact that SVM ensembles performed better than SVM individual models. The proposed model supports this fact as its performance was better than that of individual SVM models.

4. CONCLUSION

The study proposed an ensemble model for breast cancer classification using a voting ensemble of SVM and ANN. along with optimization, feature selection and sampling techniques The model was seen to give superior performance when compared to a few other models and produced an accuracy of 99.12%. Even though the model produced a high accuracy the time taken by the ensemble model is seen to be high. This has to be taken care of. Better methods are to be used to speed up training of ANNs. Besides this, a future work will be to test the performance of the model and consistency of the model across various datasets.

ACKNOWLEDGMENTS

The author thanks Dr William H Wohlberg of the University of Wisconsin Hospitals for the dataset used in the study.

REFERENCES

[1] K. Sathishkumar *et al.*, “Trends in breast and cervical cancer in India under National Cancer Registry Programme: an age-period-cohort analysis,” *Cancer Epidemiol.*, vol. 74, p. 101982, 2021.

[2] F. Bray, J. Ferlay, I. Soerjomataram, R. L. Siegel, L. A. Torre, and A. Jemal, “Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries,” *CA. Cancer J. Clin.*, vol. 68, no. 6, pp. 394–424, 2018.

[3] H. Sung *et al.*, “Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries,” *CA. Cancer J. Clin.*, vol. 71, no. 3, pp. 209–249, 2021.

[4] S. Malvia, S. A. Bagadi, U. S. Dubey, and S. Saxena, “Epidemiology of breast cancer in Indian women,” *Asia Pac. J. Clin. Oncol.*, vol. 13, no. 4, pp. 289–295, 2017.

[5] K. K. Thakur, D. Bordoloi, and A. B. Kunnumakkara, “Alarming burden of triple-negative breast cancer in India,” *Clin. Breast Cancer*, vol. 18, no. 3, pp. e393–e399, 2018.

[6] P. K. Dhillon *et al.*, “The burden of cancers and their variations across the states of India: the Global Burden of Disease Study 1990–2016,” *Lancet Oncol.*, vol. 19, no. 10, pp. 1289–1306, 2018.

[7] P. Priyadarshini, V. Hemavathy, and S. Sarathi, “RISING INCIDENCE OF BREAST CANCER IN INDIA,” *NVEO-Nat. VOLATILES Essent. OILS J. NVEO*, pp. 2284–2288, 2021.

[8] Kumar, N., Narayan Das, N., Gupta, D., Gupta, K., & Bindra, J. (2021). Efficient automated disease diagnosis using machine learning models. *Journal of Healthcare Engineering*, *2021*.

[9] K. Kourou, T. P. Exarchos, K. P. Exarchos, M. V. Karamouzis, and D. I. Fotiadis, “Machine learning applications in cancer prognosis and prediction,” *Comput. Struct. Biotechnol. J.*, vol. 13, pp. 8–17, 2015.

[10] T. E. Mathew and K. A. Kumar, “A Logistic Regression Based Hybrid Model For Breast Cancer Classification,” *Indian J. Comput. Sci. Eng.*, vol. 11, no. 6, pp. 899–906, 2020, doi: DOI : 10.21817/indjcse/2020/v11i6/201106201.

[11] Uddin, S., Khan, A., Hossain, M. *et al.* Comparing different supervised machine learning algorithms for disease prediction. *BMC Med Inform Decis Mak* **19**, 281 (2019). https://doi.org/10.1186/s12911-019-1004-8

[12] Hatem, M.Q. Skin lesion classification system using a K-nearest neighbor algorithm. *Vis. Comput. Ind. Biomed. Art* **5**, 7 (2022). https://doi.org/10.1186/s42492-022-00103-6

[13] T. E. Mathew, “A logistic regression with recursive feature elimination model for breast cancer diagnosis,” *Int. J. Emerg. Technol.*, vol. 10, no. 3, pp. 55–63, 2019.

[14] M. Islam, M. Haque, H. Iqbal, M. Hasan, M. Hasan, and M. N. Kabir, “Breast cancer prediction: a comparative study using machine learning techniques,” *SN Comput. Sci.*, vol. 1, no. 5, pp. 1–14, 2020.

[15] F. J. M. Shamrat, M. A. Raihan, A. S. Rahman, I. Mahmud, and R. Akter, “An analysis on breast disease prediction using machine learning approaches,” *Int. J. Sci. Technol. Res.*, vol. 9, no. 02, pp. 2450–2455, 2020.

[16] M. A. Aswathy and M. Jagannath, “An SVM approach towards breast cancer classification from H&E-stained histopathology images based on integrated features,” *Med. Biol. Eng. Comput.*, vol. 59, no. 9, pp. 1773–1783, 2021.

[17] M.-W. Huang, C.-W. Chen, W.-C. Lin, S.-W. Ke, and C.-F. Tsai, “SVM and SVM ensembles in breast cancer prediction,” *PloS One*, vol. 12, no. 1, p. e0161501, 2017.

[18] N. Liu, J. Shen, M. Xu, D. Gan, E.-S. Qi, and B. Gao, “Improved cost-sensitive support vector machine classifier for breast cancer diagnosis,” *Math. Probl. Eng.*, vol. 2018, 2018.

[19] C. Aroef, Y. Rivan, and Z. Rustam, “Comparing random forest and support vector machines for breast cancer classification,” *Telkomnika*, vol. 18, no. 2, pp. 815–821, 2020.

[20] H. Turabieh, “Comparison of NEAT and Backpropagation Neural Network on Breast Cancer Diagnosis.,” *Int. J. Comput. Appl.*, vol. 139, no. 8, pp. 40–44, 2016.

[21] S. Singh, H. Sushmitha, J. Harini, and B. R. Surabhi, “An efficient neural network based system for diagnosis of breast cancer,” *Breast Cancer*, vol. 8, no. 10, p. 12, 2014.

[22] K. Kaushik and A. Arora, “Breast cancer diagnosis using artificial neural network,” *Int. J. Latest Trends Eng. Technol. IJLTET*, vol. 7, pp. 41–48, 2016.

[23] T. E. Mathew, “A comparative study of the performance of different Support Vector machine Kernels in Breast Cancer Diagnosis,” *Int. J. Inf. Comput. Sci.*, vol. 6, no. 6, pp. 432–441, 2019.

[24] L. Wang, Z. Wang, G. Wei, and F. E. Alsaadi, “Finite-time state estimation for recurrent delayed neural networks with component-based event-triggering protocol,” *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 29, no. 4, pp. 1046–1057, 2017.

[25] M. M. Saritas and A. Yasar, “Performance analysis of ANN and Naive Bayes classification algorithm for data classification,” *Int. J. Intell. Syst. Appl. Eng.*, vol. 7, no. 2, pp. 88–91, 2019.

[26] A. Alzubaidi, G. Cosma, D. Brown, and A. G. Pockley, “Breast cancer diagnosis using a hybrid genetic algorithm for feature selection based on mutual information,” in *2016 International Conference on Interactive Technologies and Games (ITAG)*, 2016, pp. 70–76.

[27] M. A. Rahman and R. C. Muniyandi, “An enhancement in cancer classification accuracy using a two-step feature selection method based on artificial neural networks with 15 neurons,” *Symmetry*, vol. 12, no. 2, p. 271, 2020.

[28] M. Kumar and H. S. Sheshadri, “On the classification of imbalanced datasets,” *Int. J. Comput. Appl.*, vol. 44, no. 8, pp. 1–7, 2012.

[29] R. Akbani, S. Kwek, and N. Japkowicz, “Applying support vector machines to imbalanced datasets,” in *European conference on machine learning*, 2004, pp. 39–50.

[30] S. Chand, “A comparative study of breast cancer tumor classification by classical machine learning methods and deep learning method,” *Mach. Vis. Appl.*, vol. 31, no. 6, pp. 1–10, 2020.

[31] Kaur, S., Kumar, Y., Koul, A. et al. A Systematic Review on Metaheuristic Optimization Techniques for Feature Selections in Disease Diagnosis: Open Issues and Challenges. Arch Computat Methods Eng 30, 1863–1895 (2023). https://doi.org/10.1007/s11831-022-09853-1

[32] Hanley, J. A., & McNeil, B. J. (1982). The meaning and use of the area under a receiver operating characteristic (ROC) curve. Radiology, 143(1), 29-36.

[33] Ali, E. E. E., & Feng, W. Z. (2016). Breast cancer classification using support vector machine and neural network. International Journal of Science and Research, 5(3), 1-6.

[34] Huang M-W, Chen C-W, Lin W-C, Ke S-W, Tsai C-F (2017) SVM and SVM Ensembles in Breast Cancer Prediction. PLoS ONE 12(1): e0161501. https://doi.org/10.1371/journal.pone.0161501

[35] Abdar, M., & Makarenkov, V. (2019). CWV-BANN-SVM ensemble learning classifier for an accurate diagnosis of breast cancer. Measurement, 146, 557-570.

[36] Wang, H., Zheng, B., Yoon, S. W., & Ko, H. S. (2018). A support vector machine-based ensemble algorithm for breast cancer diagnosis. European Journal of Operational Research, 267(2), 687-699.

BIOGRAPHY OF AUTHOR

Tina Elizabeth Mathew, MSc, PhD is currently working as an Associate Professor in Government College Kariavattom, Thiruvananthapuram Kerala State, India She has 19 years of UG teaching experience. Her areas of specialization are Data Mining, Machine Learning, Cyber Security, Cyber Forensics and has 15 publications to her name.
Email: tinamathew04@gmail.com