Machine Learning for Breast Cancer Classification With ANN and Decision Tree

Reetodeep Hazra*, Megha Banerjee*, and Leonardo Badia†
* Electronics and Communication Engineering, Techno International New Town, Kolkata, India
† Department of Information Engineering, University of Padova, Padova, Italy
Email: {reetodeep.hazra.2017,megha.banerjee.2017}@ece.tint.edu.in, badia@dei.unipd.it

***Abstract*—Breast cancer is one of the commonest cause of cancer deaths in women. It starts developing when threatening bumps start forming from the breast cells, and unfortunately most diagnoses happen in later stages, thus resulting in low chances of survival for the patient. So for early detection and prognosis, it is necessary to detect the benign or threatening nature of the bumps. In this paper, Artificial Neural Networks (ANN) and Decision Tree (DT) classifiers are used to develop a machine learning (ML) model using the Wisconsin diagnostic breast cancer (WDBC) dataset, in order to evaluate the attributes of a breast cancer development at beginning phases and classify it as malignant or benign. In the proposed scheme, feature selection and feature extraction are done to extract statistical features from the dataset and comparison between the models is provided based on their performance to identify the most suitable approach for diagnosis. The dataset apportioned into various arrangements of train-test split. The presentation of the framework is estimated, depending on accuracy, sensitivity, specificity, precision, and recall. The binary classification problem achieved a maximum accuracy of 98.55%.**

***Index Terms*—Artificial neural networks; Breast cancer; Biomedical imaging; Decision tree classifier; Machine learning.**

I. Introduction

THE FAST evolution of artificial intelligence (AI) and particularly deep learning (DL) keeps on powering the enthusiasm of the clinical imaging community towards applying these methods to improve cancer screening. Breast cancer is the commonest of tumor for women after skin cancer. As of 2020, the data related only to the United States population predict 276480 to suffer from invasive breast malignancy, and 48530 to have non-invasive (in situ) breast cancer. While much less frequently, also men can be affected, and the expected number for the United States is for 2620 men to have invasive breast cancer. Accordingly, nearly 42,690 deaths which includes 42,170 women and 520 men, will occur in the year of 2020 from breast cancer [1]. The main types of breast cancers are ductal carcinoma in situ (DCIS) and invasive (or infiltrating) breast cancer (ILC) [2]. Based on these types, breast cancer tumors can be determined and classified into benign or malignant. Benign tumors are considered as noncancerous, that is, non-dangerous whereas malignant tumors begins from an abnormal cell development which eventually spreads into its surrounding tissues. The nuclei of the malignant tissue is much greater than in benign tissue, which can be dangerous in future stages.

978-1-7281-8416-6/20/$31.00 © 2020 IEEE

Several research works are available that focuses on classification of breast cancers. Nahid *et al* [2] used Convolutional neural network (CNN) for breast image classification. The work also involved other algorithms like Random forest (RF) and Support vector machines (SVM) to predict mammographical images. In [3], Kourou *et al.* proposed ML applications for prognosis of cancers and predictions on it. Likewise, a variety of breast cancer prediction models based on different machine learning (ML) algorithms have also been proposed. Amutha *et al.* [4] proposed Decision Tree (DT), SVM and Sequential Minimal Optimization (SMO) for early diagnosis of growth of breast cancers. Ambrane *et al.* [5] used two ML algorithms namely Naive Bayes (NB) classifier and k-nearest neighbor (KNN) and performed cross-validation to improve the overall accuracy. Other novel algorithms like breast cancer recurrence prediction based on SVM (BCRSVM), Coxproportional hazard regression model, back-propogation neural network (BPNN) etc. have also been used. Kim *et al.* [6] used BCRSVM. They also contrasted the SVM with Artificial neural network (ANN) and Coxproportional hazard regression model and achieved an overall sensitivity of 0.89 and specificity of 0.73. Paulin *et al.* [7] built a framework with BPNN and used Levenberg-Marquardt calculation to achieve a higher overall accuracy.

In this work, we approach a binary classification problem for classifying breast cancers based on different anomalies present in breast tumors. Two frequently adopted ML algorithm frameworks, namely ANN and DT, are utilized to classify breast cancer and compared in terms of their respective performance. In ANN, label encoder is used by which the levels of categorical features are encoded into numeric values of 0 and 1. In DT, three methods of feature scaling is applied to the dataset for statistical scaling of features. After feature scaling and feature extraction, the data is fed as inputs to the ML model. The dataset utilized in this work is acquired from the openly accessible dataset of Wisconsin diagnostic breast cancer (WDBC) [8]. The effects of various factors from feature extraction such as radius, textures, area of breasts are being isolated to determine how they affect the model performance. The maximum accuracy achieved is from the ANN algorithm used here which reports an accuracy

of 98.55% which is better in comparison of state-of-the-art facilities. The contribution of the present paper can be summarized as follows:

- We structured an ANN with low latency to amplify the responsiveness.
- We performed several feature scaling process which resulted in a higher accuracy.
- We compared two most preferred algorithm for breast cancer classification - ANN and DT.

The rest of this paper is explained as follows: Section II portrays the related works of ML applications in the field of clinical imagining in particular reference with breast cancer. This is trailed by the description of our methodology in Section III. Section IV clarifies the performance evaluation which is followed by results and discussions in Section V. Conclusion of our work is referenced in Section VI.

II. Related Works

The subject of breast cancer classification via ML is widely studied, and many datasets are available, to apply these methodologies to different kinds of breast tumors. Here, we take a general approach where we categorize the literature from the standpoint of the classification techniques used, further discussing advantages and disadvantages of each technique.

A. Classification using Data Mining techniques

Gupta *et al.* [9] provided a broad review of breast cancer researches done with the help of data mining techniques. They concluded that by applying data mining techniques, breast cancer can be analyzed at an early stage. Their proposed approach investigates the data successfully with respect to previous models: they accomplished 98.1% accuracy over random split. Majali *et al.* [11] also used data mining techniques for determination and diagnosis of breast cancer. They utilized techniques like FP Growth calculation and ID3 calculation to identify malignancy in its beginning phases. Mohammed *et al.* [12] used data mining techniques and classified breast cancers as benign and malignant. They used three ML algorithms namely DT, NB, and SMO. The results concluded that utilizing a resample channel upgrades the classifier's presentation and concluded that SMO performs better than others in the WDBC dataset whereas J48 is better than state-of-the-art facilities in the breast cancer dataset.

B. Classification on the basis of Mammograms

Nahid *et al* [2] gave a special emphasis on CNN for breast image classification. The work also described the involvements of other classifiers such as RF, SVM and other supervised and unsupervised methods, which have been used in the classification of breast image. Kathale *et al.* [14] presented a diagnosis process to detect the cancerous region. On the basis of this detection, they classified normal and cancer patients. Initially pre-processing was applied to the mammogram images and the undesirable parts were removed. They used RF classifiers and acheived an accuracy of 95%.

C. Other classification techiques

Ghosh *et al.* [15] made a NFS which is neuro-fuzzy-based breast cancer classification system using datasets of WDBC and mammographic mass. Here a multilayer perceptron model is utilized for breast cancer classification. The dataset is fuzzified utilizing sigmoidal membership works and processes level of membership for individual patterns to different classes. Lastly defuzzification is used and the NFS system achieved an accuracy of 97.8%.

Mumin *et al.* [16] made a comparative study of a few classification algorithms for breast malignant growth determination utilizing data collection from the estimations of a radio wire with a 10-overlap cross-validation technique. RF performed the best during 10-fold cross-validation and the model achieved an accuracy of 92.2%. Tuba *et al.* [17] proposed a statistical neural network-based breast cancer diagnosis. In the model, they utilized radial basis network (RBF), general relapse neural system (GRNN), probabilistic neural network (PNN) and factual neural system structures on WDBC dataset. The framework acquired 98.8% on 50–50 apportioning split.

Nayak *et al.* [18] used a framework which uses adaptive resonance theory (ART-1) network for classification purpose. They contrasted ART-1 with PSO-MLP and PSO-BBO calculations and concluded that ART is the best among other two classifiers. They partitioned the dataset into 70-30 ratio for preparing and testing the information. Nikita *et al.* [19] proposed a comparison of six ML algorithms namely RF, NB, KNN, ANN, SVM and DT on the WDBC dataset. They compared these algorithms and classified the cancers as benign and malignant.

For achieving a higher accuracy, pre-processing is needed. So we applied pre-processing like feature extraction and feature scaling techniques. The accuracy achieved was over 98%.

III. Methodology

This section describes the binary classification problem and the ML algorithms used: ANN and DT for the task of classification of breast cancers. The proposed methodology is divided into four sections. The first section is the data assortment and source which is trailed by feature extraction and feature selection. Feature extraction is a process that increases the accuracy of the learned model by extracting features from the input data. It aims to lessen the quantity of features in a dataset. Feature selection is the process where the features are consequently or manually chose which contribute most of the prediction variable in which we are interested in. It aims at creating an accurate predictive model. Then comes the main section which describes the applications of ML algorithms used and finally performance evaluation is reported in the last section.

A. Data collection and source

The data set used in this paper is the WDBC data set [8]. This dataset consists class division of breast cancer diseases as malignant and benign. Features are processed from a digitized picture of a fine needle aspirate (FNA) of a breast mass. There were ten real-valued features which were computed for each cell nucleus. They were: perimeter, area, radius, texture, smoothness, concave points, symmetry, fractal dimension, concavity and compactness.

B. Feature extraction and selection

The proposed model is tested on the basis of two ML algorithms: ANN and DT. For ANN, we have used LabelEncoder for extraction of features from the existing data set. Here, LabelEncoding is performed to encode our variable to numbers. It refers to converting the labels to numerical form for making it machine readable. ML algorithms can then decide in a better way on how those labels must be operated. The approach worked reasonably well with the ANN model.

For DT, we used and compared three methods for feature selection. The three methods are listed below:

(i) No Feature Selection - Initially we evaluated the model with no feature selection to see how the model performs and calcutated the mean and worst of the ten features composed for each cell nucleus. The heatmap of the no feature selection model is shown in Fig. 1.

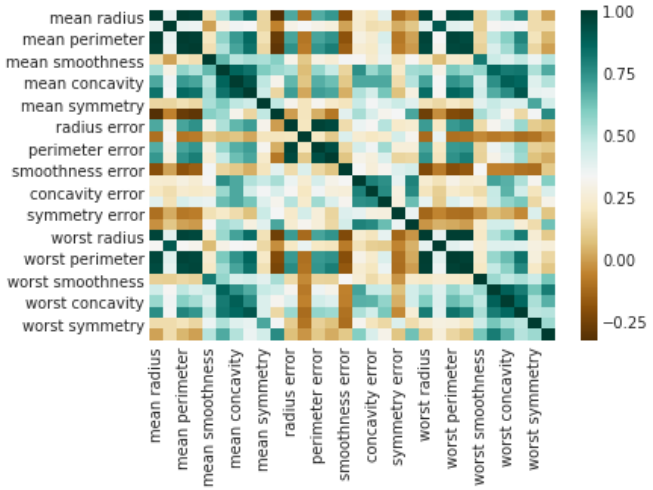

Figure 1: Heatmap of No Feature Selection

(ii) Features that are not correlated - During no feature selection, we found that there were many features that were correlated. In the model, features namely radius, compactness, concavity, smoothness, concave points, perimeter and fractal dimensions were found to be correlated. So these features were eliminated. The heatmap after eliminating these features is shown in Fig. 2.

(iii) PCA transformation -Principal Component Analysis (PCA) change, is a dimensionality decrease strategy that is frequently used to lessen the dimensionality of huge data collections, by changing a huge set of variables into a more modest one that actually contains the greater part of the information in the enormous set. In this model we used PCA transformation to select features and reduce feature correlation. The heatmap is shown in Fig. 3.

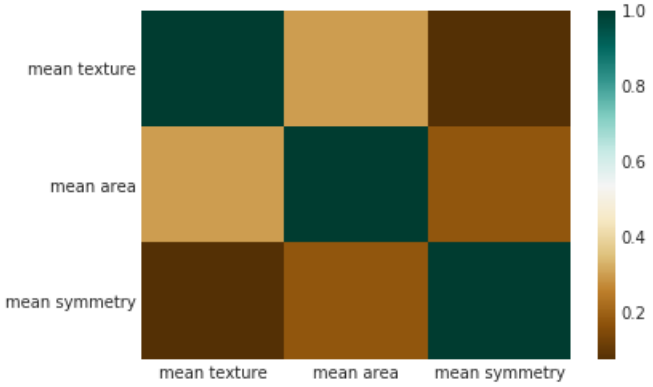

Figure 2: Heatmap of Non-correlated Features

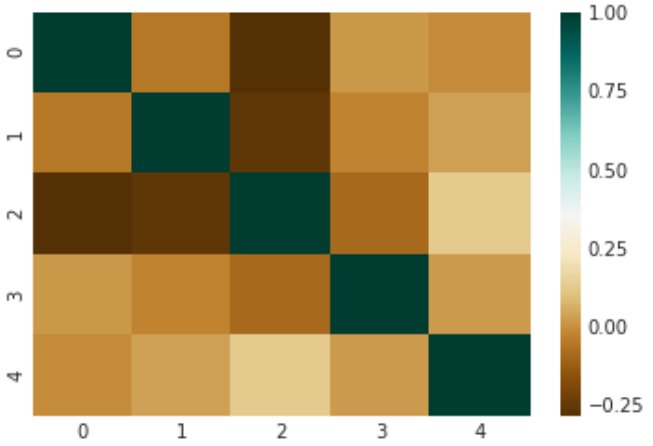

Figure 3: Heatmap of Feature correlation after PCA

C. Application of ML Algorithms

In the proposed model, two ML algorithms were used: ANN and DT.

(i) Artificial Neural Network (ANN) -ANN is a computational model dependent on the basis of structural elements of a biological neural network. ANNs can be used in binary classification problems where only a single output neuron using the logistic activation function: the output will be a binary number where the estimated probability of the positive class can be interpreted. A single neuron, known as *perceptron*, consists of a layer of inputs (corresponding to columns of a dataframe), where each input has a weight which controls its magnitude for a weighted summation, which is in turn fed to the activation function.

In our classifier, we utilized densely connected neural network of four layers with one input output layer and two

hidden layers with a Rectified linear circuit (ReLu) activation. The classifier is made on a sequential basis. This is the most straightforward keras model for neural networks. We include a dense hidden layer with 16 neurons. Each dense layer deals with its own weight matrix and contains all the association weights of the neurons and their sources. It also aditionally deals with a vector of bias terms (one per neuron). The initiation work ReLu just potrays the positive part of the contention as the negative part of the contention is zero. This model has low latency as it involves least layers and least channels per layer. Input dimension portrays the quantity of nodes in the input layer. The output dimension of each hidden layer recoils as we continue further in the network. As it is a binary classification, we have utilized sigmoid activation in the last hidden layer. Results go improved by setting units to 16 at the input layer and reducing the units in the hidden layers. As the classification is binary, so binary crossentropy loss function is used with softmax activation. RMSprop optimizer is being utilized to prepare the classifier as it confines the oscillation in a vertical way and calculation could make bigger strides in the horizontal direction converging quicker. A total of 785 parameters are tested which gives the model a lot of flexibility to fit the training data.

Table I summarizes the dimensional and operational information of ANN architecture proposed.

Table I: Architecture of ANN

Layer	Type	Units	Output shape
1	Input	16	(1, 16)
2	Hidden layer 1	8	(1, 8)
3	Hidden layer 2	6	(1, 6)
4	Sigmoid	1	(1, 1)

(ii) Decision Tree (DT) - DT is a fundamental component of RF. DT uses a layered splitting process, where at each layer the information data is split into two or more groups so that elements of the same group are homogenous to each other. The root node of the DT considers whether the mean area is smaller than 696.25 at depth 0, which would imply that the class is benign. There can be two possibilities: True or False. If it is true, then DT moves downside to the root's left child node. Here in the same manner, it checks the mean symmetry is lesser than 0.202 and the class is benign. Similarly if the parent node is false then the DT moves downside to the root's right child node. A node's sample property checks the number of training samples it applies to. In the proposed mode, from the parent node it is seen that there are 455 samples which has a mean area of less than or equal to 696.25. Out of these 455 samples, 133 training instances have a mean area of greater than 692.5. Here 0 applies to benign and 1 applies to malignant. In the same manner, the total structure of DT is formed. Finally, a node's Gini attribute [21] measures it's impurity. A node is pure (Gini score equal to 0) is all training instances it applies belongs to the same class. The equation stated below shows how the training algorithm computes the Gini score G_i of the i^{th} node.

$$G_i = 1 - \sum_{k=1}^{n} P_i \ , k^2$$

In this equation,

P_i, k is the ratio of class k instances among the training instances in the i^{th} node.

IV. Performance Evaluation

A. Metrics

Confusion matrix is used to classify results where the output is of two classes. The confusion matrix are given below in the Table II.

Table II: Confusion matrix

			Actual data	
			Benign	Malignant
Predicted data	ANN	Benign	39	1
		Malignant	0	29
	DT	Benign	52	1
		Malignant	1	89

V. Results and Discussion

A. Accuracy

The ANN classifier was trained for 150 epochs with a batch size of 128 utilizing RMSprop optimizer whereas the DT classifier was trained with Adam optimizer achieving an accuracy of 96%. The other parameters utilized for proposed model are referenced in Tables III and IV.

Table III: Accuracy and Other parameters (ANN)

Optimizer	Batch size	Epochs	Accuracy
RMSprop	128	150	98.5%

Table IV: Accuracy (DT)

Optimizer	Accuracy
Adam	96%

B. Classification Results

The complete analysis result for both the algorithms are presented in Tables V and VI. To address the performance of these two investigations, precision, accuracy, and F1-score were considered as the examination standards.

Table V: Classification results for ANN

	Precision	Recall	F1-score
Benign	1.00	0.97	0.99
Malignant	0.97	1.00	0.98

Table VI: Classification results for DT

	Precision	Recall	F1 score
Benign	0.94	0.94	0.94
Malignant	0.97	0.97	0.97

C. Receiver Operating Characteristic (ROC) Curve

The ROC is a tool used for binary classification and is plotted with respect to TFR and FPR which signifies true positive rate and false positive rate. The FPR can be said as 1 - the true negative rate (TNR). TNR is also termed as specificity. In the proposed model, the ROC curve area for ANN came 99.87 whereas the ROC curve area for DT came 99.47. The ROC curves for both the algorithms, ANN and DT, are shown in Fig. 5 and Fig. 6.

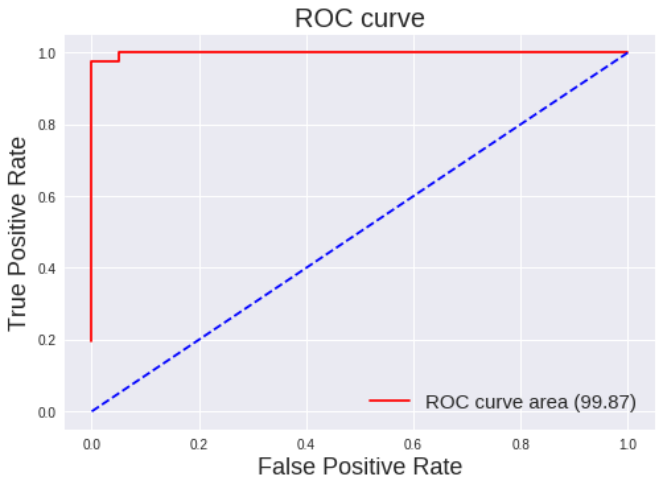

Figure 4: ROC curve for ANN

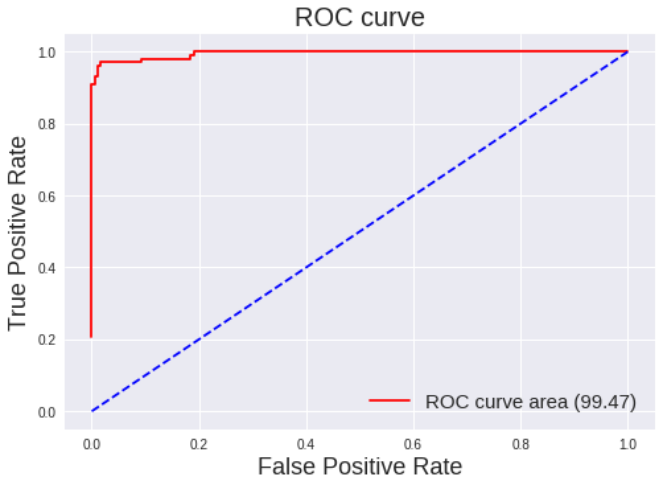

Figure 5: ROC curve for DT

Table VII: Related Works And Their Accuracy

Model	Reference	Accuracy
ANN	(Proposed Model)	0.98
DT	(Proposed Model)	0.94
DT (Feature correlation)	(Proposed Model)	0.86
DT (PCA Transformation)	(Proposed Model)	0.96
KNN	[6]	0.97
NB	[6]	0.96
BCRSVM	[7]	0.85
MPANN	[10]	0.97
ROI (Region Based Features)	[13]	0.91
SVM	[20]	0.96
SVM (LDA)	[20]	0.98

D. Comparison and Discussion

Table VII describes the related works and the accuracies of the related works. The accuracy of the model presented here consisting of two ML algorithms ANN and DT is placed at the top of the table. Between ANN and DT, ANN reported the highest accuracy of 98.55%. In the proposed model, ANN approach took a little more time in comparison with DT but DT is more complex. ANN model can be implemented easily; the only drawback is that it is a little more time consuming than DT. Ambrane *et al.* [18] performed classification on KNN and NB. KNN achieved the highest accuracy of 97% with the lowest error rate whereas our proposed model achieved a maximum accuracy of over 98%. Kim *et al.* [7] used BCRSVM which is a novel idea and achieved an accuracy of 85% which is much lesser for a completely numerical dataset. Abbass *et al.* [10] used MPANN to classify breast cancers and our accuracy outperforms their accuracy. Our proposed model classified breast cancers on several parameters and the accuracy achieved is higher than them. Kashyap *et al.* [13] used mammographic images to classify breast cancers. The work only used ROI to detect the malignancies in the mammograms but classification is not present so any conclusion cannot be inferred. Omondiagbe *et al.* in [20], used SVM and SVM (LDA) for classification and the overall accuracy came as 98.82% which is slightly greater than the proposed work. The area under ROC came as 99.94 whereas our ANN reported an ROC curve area of 99.87. So the results proposed by Omandiagbe *et al.* is somewhat similar to ours.

It is seen from the proposed works that boundaries like dimensionality reduction, include extraction and scaling assumes an essential part in the characterization models. The primary motivation behind performing feature extraction is to improve the prediction performance and guarantee quicker forecast. The accuracy detailed in the proposed work is superior to the cutting edge facilities.

VI. Conclusion

In this paper, WDBC data set [8] is being utilized to classify breast cancer by utilizing two well-known ML frameworks - ANN and DL. In the proposed classifiers, feature selections are done to remove statistical features from the data set and comparison between the models is given dependent on their

performance to determine the most appropriate methodology for conclusion. In ANN, label encoder is utilized, according to which the levels of categorical features are encoded into numeric values of 0 and 1. In DT, three strategies for feature scaling is applied to the data set for statistical scaling of features. In this two calculations, ANN outperformed DT by accomplishing accuracy of 98.55%.

We utilized feature extraction technique to improve the prediction performance and ensure faster predictions. Also, in ANN, RMSprop optimizer is being used in place of traditional Adam optimizer which provides a greater learning rate and allows the algorithm to take greater strides in the horizontal direction converging faster. Future work can be coordinated towards forming the chosen approach into a likely practical strategy for supporting and helping specialists with brisk assessment in diagnosing breast cancer.

REFERENCES

[1] Cancer.Net, *Breast Cancer: Statistics*, May 2020, Accessed on September 7, 2020 [Online], Available: https://www.cancer.net/cancer-types/breast-cancer/statistics

[2] A. Al Nahid, Y. Kong, "Involvement of machine learning for breast cancer image classification: A Survey," *Hindawi Comput. Math. Meth. Medicine*, vol. 2017, pp. 1-29, 2017.

[3] 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.*, vol. 13, pp. 8-17, Nov 2014.

[4] R. Amutha and M. Savithri, " Diagnosis and prognosis of breast cancer using data mining techniques," *Paripex-Indian J. Res.*, vol. 4, pp. 6-8, March 2015.

[5] M. Amrane, S. Oukid, I. Gagaoua, and T. Ensar, "Breast cancer classification using machine learning," *Proc. Electric Electronics, Computer Science, Biomedical Engineerings' Meeting (EBBT), IEEE,*, 2018, pp. 1-4.

[6] W. Kim, K. S. Kim, J. E. Lee, D.-Y. Noh, S.-W. Kim, Y. S. Jung, M. Y. Park, and R. W. Park,"Development of novel breast cancer recurrence prediction model using support vector machine," *J. Breast Cancer*, vol. 15, no. 2, pp. 230-238, June 2012.

[7] F. Paulin and A. Santhakumaran, "Classification of breast cancer by comparing backpropagation training algorithms," *Int. J. Comp. Sc. Engin.*, vol. 3, no. 1, pp. 327-332, 2011.

[8] W. H. Wolberg, W. N. Street, and O. L. Mangasarian, "Breast Cancer Wisconsin (Diagnostic) Data Set," 1995, https://archive.ics.uci.edu/ml/index.php

[9] S. Gupta and A. Sharma, "Data mining classification techniques applied for breast cancer diagnosis and prognosis," *Ind. J. Comp. Sc. Engin.*, vol. 2, no. 2, pp. 188-195, 2011.

[10] H. A. Abbass, "An evolutionary artificial neural networks approach for breast cancer diagnosis," *Artif. Intell. Med.*, vol. 25, no. 3, pp. 265-281, 2002.

[11] J. Majali, R. Niranjan, V. Phatak, and O. Tadakhe, "Data mining techniques for diagnosis and prognosis of cancer," *Int. J. Adv. Res. Comp. Commun. Engin.*, vol. 4, no. 3, pp. 613-616, March 2015.

[12] S. A. Mohammed, S. Darrab, S. A. Noaman, G. Saake, "Analysis of breast cancer detection using different machine learning techniques," *Proc. Int. Conf. Data Mining and Big Data (DMBD)*, 2020, pp. 108-117.

[13] K. L. Kashyap, M. K. Bajpai and P. Khanna, "Breast cancer detection in digital mammograms," *Proc. IEEE Int. Conf. Imaging Syst. Techniques (IST)*, 2015, pp. 1-6.

[14] P. Kathale and S. Thorat, "Breast cancer detection and classification," *Int. Conf. Emerging Trends Inf. Tech. Engin. (ic-ETITE)*, 2020, pp. 1-5.

[15] S. Ghosh, S. Biswas, D. C. Sarkar, P. P. Sarkar, "Breast cancer detection using a Neuro-fuzzy based classification method," *Ind. J. Sc. Tech.*, vol. 9, no. 14, pp. 1-15, May 2016.

[16] M. Kaya Keleş "Breast cancer prediction and detection using data mining classification algorithms: A comparative study," *Tehnički vjesnik*, vol. 26, no. 1, pp. 149-155, 2019.

[17] T. Kiyan and T. Yildirim, "Breast cancer diagnosis using statistical neural networks," *Istanbul Univ. J. Elec. Electron. Eng.*, vol. 4, no. 2, pp. 1149-1153, 2004.

[18] S. Nayak and D. Gope, "Comparison of supervised learning algorithms for RF-based breast cancer detection," *Proc. Comput. Electromag. Int. Works. (CEM)*, 2017, pp. 13-14.

[19] S. Jean, R. Nikita, K. Rucha, and D. Sulochana, "Breast cancer classification and prediction using machine learning," *Int. J. Eng. Res.*, vol. 9, no. 2, pp. 576-580, Feb 2020.

[20] O. David, V. Shanmugam, and S. Amandeep, "Machine learning classification techniques for breast cancer diagnosis," *Proc. CUTSE*, IOP Conf. Series: Mat. Sci. Eng., vol. 495, 2018.

[21] N. Suneetha, Ch. V. M. K. Hari, K. V. Sunil, "Modified Gini index classification: a case study of heart disease dataset," *Int. J. Comp. Sc. Eng.*, vol. 2, no. 6, pp. 1959-1965, October 2010.

[22] E. Jackson and R. Agrawal, "Performance evaluation of different feature encoding schemes on cybersecurity logs," *SoutheastCon*, pp. 1-9, 2019.

[23] P. Vidnerova and R. Neruda, "Evolving keras architectures for sensor data analysis," *Federated Conference on Computer Science and Information*, pp. 109-112, 2017.