Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 218 (2023) 429–437

Procedia Computer Science

www.elsevier.com/locate/procedia

International Conference on Machine Learning and Data Engineering

Diagnosis of Breast Cancer Using Random Forests

Manas Minnoor[a], Veeky Baths[b]

[a]Department of Computer Science and Information Systems, BITS Pilani KK Birla Goa Campus, Goa 403726, India
[b]Cognitive Neuroscience Lab, BITS Pilani KK Birla Goa Campus, Goa 403726, India

Abstract

Breast cancer was the most diagnosed form of cancer in 2020. Early diagnosis of breast cancer results in a significant improvement in long-term survival rates. Current methods require consultation with experts, which is expensive and time-consuming and thus may not be accessible to all. This paper seeks to train and evaluate supervised machine learning models for the accurate and efficient detection of breast cancer. The Wisconsin Breast Cancer Database dataset describes 30 attributes of cell nuclei, including, but not limited to, their radius, texture, and concavity. It contains 569 instances, 212 of which are malignant tumors. The Random Forest algorithm outperforms other algorithms in classifying breast tumors as either malignant or benign and is thus selected as our primary model. It is trained on two different subsets of the dataset having 16 and 8 features, respectively, identified with the help of multiple feature selection methods. The Random Forest models are tested post hyperparameter tuning on a holdout set, and accuracies of 100% and 99.30% respectively. The models are also compared with four other machine learning classification algorithms: Support Vector Machine (SVM), Decision Tree, Multilayer Perceptron, and K-Nearest Neighbors. The results confirm that Random Forest is the superior method for breast cancer diagnosis.

Keywords: Breast cancer; Supervised learning; Classification; Random forests; Hyperparameter tuning;

1. Introduction

Breast cancer is cancer that develops in the cells of the breast. It may manifest as a lump, change in shape, or fluid, among other symptoms. Risk factors may be genetic, such as the BRCA1/2 mutations [4], or lifestyle habits, including alcohol and tobacco consumption. However, only a quarter of cases may be prevented through lifestyle changes.

The number of breast cancer cases diagnoses overtook those of lung cancer in 2020 to become the most common form of cancer [1]. A dearth of qualified experts, as well as the high costs associated with consultation, result in many women being unable to access the care they need. Developing countries like India also suffer from an acute

10.1016/j.procs.2023.01.025

shortage of medical professionals. Thus, the introduction of automated clinical decision systems may help alleviate this problem. The lack of widespread knowledge about the function of machine learning models in medicine poses an obstacle to the adoption of such systems, however. This paper aims to help overcome this issue by providing results that support the case for including these models in the clinical workflow.

Undiagnosed breast cancer is often fatal, while early diagnosis results in better outcomes [2]. Thus, an automated system to diagnose breast cancer may help drastically bring down the number of fatalities. Furthermore, such a system may help complement clinicians in the field by confirming their diagnoses. The existing method of manually reviewing mammogram results fails to scale in large populations, such as that of India.

This paper seeks to develop such a system with the help of supervised machine learning techniques. A classification model must be designed to differentiate malignant tumors, as these are the cases that are likely to be fatal. Benign tumor treatment is limited to removing the cancerous cells and preventing a reoccurrence. However, malignant tumors may metastasize to other parts of the body, and thus it is of utmost importance to identify such patients in the earliest stages of the disease.

Five different supervised machine learning algorithms are explored for this use case. Out of these, the Random Forest model is chosen as our primary model due to its superior out-of-the-box performance in diagnosing malignant tumors. This model's hyperparameters are then tuned to create the final model that may be used in the medical field. Furthermore, a similar model is also developed on a reduced dataset containing just eight dataset features. This model performs on par with the model trained on the full dataset. This reduced dataset is chosen to optimize the model in terms of computational power and data required for diagnosis.

The rest of this paper is organized as follows. Section 2 details work related to this paper's objective. Section 3 presents the methodology of this paper. Section 4 describes the dataset and its preprocessing. Section 5 sheds light on the feature selection utilized to obtain the final datasets. Section 6 and 7 elucidate the model development and evaluation, respectively. The results and discussion are contained in section 8. Section 9 briefly describes the possible future work in this avenue and concludes the paper.

2. Related Work

As can be seen in [5]-[9], the Random Forest algorithm consistently provides superior models for the diagnosis of various diseases. Jackins et al. [5] achieve an accuracy of 83.85% while diagnosing coronary heart disease using Random Forests. They also portray the superior performance of Random Forests at detecting breast cancer. Sarica et al. [6] conclude that Random Forests outperform existing machine learning methods for classifying neuroimaging data in Alzheimer's disease.

A review of [10]-[13] supports the application of machine learning in breast cancer diagnosis. Naji et al. [12] provide evidence that the Support Vector Machine and Random Forest algorithms achieve accuracies of over 96% while detecting malignant tumors. Vaka et al. [11] find that a Deep Neural Network surpasses standard supervised models such as K-Nearest Neighbors and Decision Tree, producing an accuracy of 97.21%. However, neural networks are prone to overfitting on smaller datasets and thus are avoided for the chosen dataset.

As shown by [14]-[17] and [24], various other supervised machine learning techniques also prove to be proficient at detecting malignant breast tumors. Azar and El-Metwally [14] elicit accuracies of over 95% while diagnosing breast cancer using various Decision Tree classifiers. Desai and Shah [15] deploy a Multilayer Perceptron (MLP) classifier and observe an accuracy of 91.9% on the Wisconsin breast cancer dataset. Polat and Güneş [16] employ a variation of the Support Vector Machine, namely the Least Square Support Vector Machine (LS-SVM). They state a final accuracy of 98.53%, cementing the feasibility of utilizing SVMs in tumor detection. Finally, Sarkar and Leong [17] apply K-Nearest Neighbors (KNN) to this problem. They claim to achieve an improvement of 1.17% in classification results as compared to existing models.

Based on the numerous papers reviewed in this section, we may observe that a multitude of supervised machine learning methods may be effectively trained to detect and diagnose malignant breast tumors.

3. Methodology

Fig. 1 describes the implemented machine learning workflow.

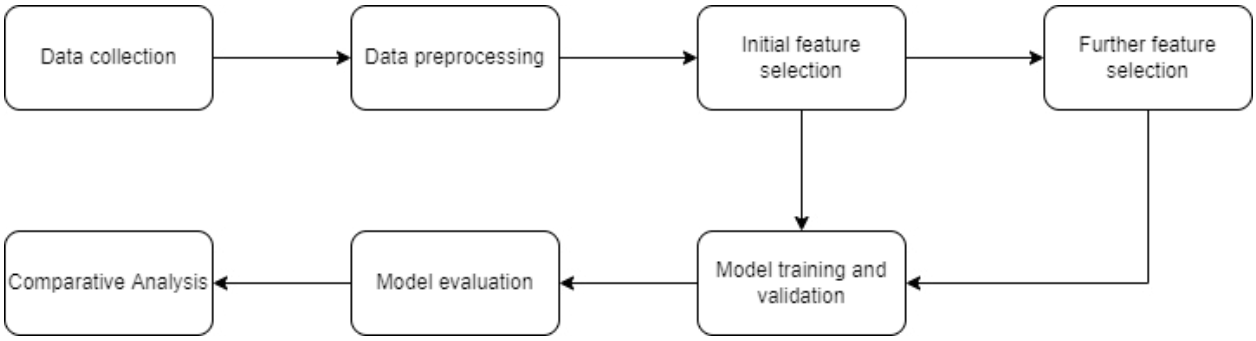

Fig. 1. Machine Learning Workflow

The experiment begins with the retrieval of the selected dataset from the UCI Machine Learning repository. The data is then cleaned and preprocessed, to ensure only relevant data remains. Feature selection is initially carried out using correlation coefficients, yielding a dataset with 16 features. This is done to remove the redundant features. The five machine learning models are then trained on this data. The Random Forest model is further tuned as well, by modifying its hyperparameters. Following this, three different feature selection methods are harnessed to further reduce the dataset to just eight attributes. A similar procedure for training the models is followed on this minimal dataset as well. Finally, the models are tested on a holdout set, and their performances are analyzed.

4. Dataset

4.1. Dataset Description

The UC Irvine Machine Learning Repository's Wisconsin Breast Cancer Diagnostic dataset is chosen for the experiment [3]. The dataset contains 569 instances of tumors, 212 of which are malignant and 357 of which are benign growths. The dataset is insufficiently balanced, with less than 40% of the instances belonging to the positive class. Thus, some form of upscaling is necessary for the chosen dataset. The attributes are detailed in Table 1. There are ten attributes describing measurements of each cell nucleus, calculated from a computerized image of a fine needle aspirate of a tumor. The mean, standard error (se), and worst values are included for each attribute in the table, resulting in 10*3 = 30 features. The final attribute is the diagnosis- benign or malignant. The features are hereafter referred to by their name followed by their value type (mean/se/worst).

Table 1. Attribute Description

Number	Attribute	Description
1	Radius	Radius of the cell
2	Texture	Standard Deviation of greyscale values
3	Perimeter	Perimeter of the cell
4	Area	Area of the cell
5	Smoothness	Local variation in radii
6	Compactness	$(Perimeter^2/Area) - 1.0$
7	Concavity	Severity of concave portions of contour
8	Concave Points	Number of concave points of contour
9	Symmetry	Symmetry of cell
10	Fractal Dimension	"Coastline approximation" – 1
11	Diagnosis	Malignancy of tumor

4.2. Data Preprocessing

The dataset contains 32 columns, out of which the arbitrary ID column is dropped as it has no relevance to the experiment. The diagnosis column further becomes our target variable. None of the instances contain any erroneous or missing data, and thus all 596 entries are included in the training data. The Diagnosis feature is converted to binary values using one-hot encoding. The rest of the features are decimal values with four significant digits.

As mentioned before, the dataset suffers from a class imbalance problem. The minority class is upsampled to match the majority dataset count to resolve this. The final dataset contains 714 instances. This helps prevent the models from blindly predicting the majority class. The features have widely varying ranges, which may affect the performance of some supervised learning methods such as K-Nearest Neighbors. A MinMax scaler is also employed to bring the features of the dataset to a uniform range between 0 and 1.

Finally, the data is split into an 80/20 training-test split to maximize the training data. As there are not many outliers in the data, the 20% test set is sufficient to evaluate the final models objectively.

5. Feature Selection

5.1. Initial Feature Selection

The initial stage of feature selection is based on the Pearson correlation coefficient. For pairs of features having a correlation coefficient higher than 0.8, one of the features is dropped to avoid multicollinearity. This process results in 14 features being removed and 16 features remaining. This dataset is used to train and evaluate the first set of models. It will be referred to as the initial dataset from now on.

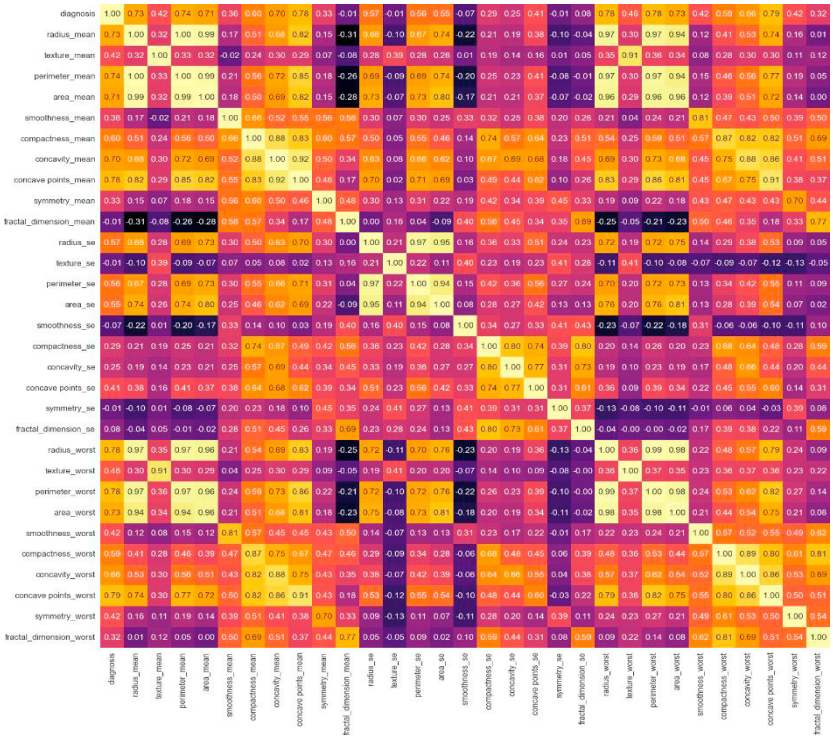

Fig. 2. Feature correlation map

The set of features contained in the initial dataset is detailed in set F_1 as follows:

F_1 = {texture_mean, area_mean, smoothness_mean, concavity_mean, symmetry_mean, fractal_dimension_mean, texture_se, area_se, smoothness_se, concavity_se, symmetry_se, fractal_dimension_se, smoothness_worst, compactness_worst, symmetry_worst, fractal_dimension_worst}

5.2. Further Feature Selection

The second stage of feature selection is based on three different methods- Recursive Feature Elimination, Logistic Regression, and Univariate Selection.

5.2.1. Recursive Feature Elimination

Recursive Feature Elimination (RFE) is a wrapper-type feature selection method that has been shown to significantly improve model performance [18]. RFE uses an external estimator- Extra Trees classifier in this case. The features are ranked by importance using the external estimator, and the least important features are removed. This procedure is then repeated recursively till eight features remain.

5.2.2. Logistic regression

In the second method, feature importance are determined using a Logistic Regression model [19]. The eight features with the highest importance are then selected.

5.2.3. Univariate Selection

Univariate selection selects the required features by performing univariate statistical tests on the data. In this case, the ANOVA F-value is used to determine the eight best features.

No feature selection method is without fault. Thus, it is naïve to base the final feature set on any one of the three methods. To combat the particular selection methods' weaknesses, a feature is selected if at least two of the methods above rank it in the top eight. This ensures that the feature selection is not affected by biases in the individual methods. This strategy, however, leaves us with eight features that make up the second dataset. The eight top features were shared in all three ways, implying that they are reasonable indications of the malignancy of a tumor. This dataset will be referred to as the minimal dataset.

The set of features contained in the minimal dataset is detailed in set F_2 as follows:

F_2 = {texture_mean, area_mean, concavity_mean, fractal_dimension_mean, area_se, smoothness_worst, compactness_worst, symmetry_worst}

6. Model Development

Five models are selected for a preliminary comparison- Random Forest (RF), Support Vector Machine (SVM), Decision Tree (DT), Multilayer Perceptron (MLP), and K-Nearest Neighbors (KNN). They are trained on the initial dataset with default parameters, and their results are compared. The models are 5-fold cross-validated using recall value as the scoring function. The recall value is used as it is more acceptable for a medical model to return a false positive than a false negative. The Random Forest model performs the best with a cross-validation recall score of 0.9896. Thus, the Random Forest model is selected as our primary model.

The Random Forest is an ensemble supervised machine learning technique [20]. It functions by constructing multiple decision trees during training. In classification scenarios, the class chosen by the majority of trees is returned by the classifier. This ensemble method has been shown to consistently outperform single Decision Tree classifiers [21]-[22]. The averaging built into this meta estimator prevents any possible overfitting that may occur in non-ensemble methods. Furthermore, it raises the predictive accuracy of the model as well. The model may be tuned to either use the entire dataset or a sample of the dataset to train each decision tree [23].

The four remaining models are trained using their default parameters on the initial and the minimal dataset. Random Forest's hyperparameters are precisely tuned to maximize the primary model's performance. Manual tuning of hyperparameters is time-consuming and not always accurate. Thus, an automated grid search is done over the possible hyperparameters. 5-fold cross-validation is employed to avoid any over-fitting. The data is split into five equal folds. In each of the five iterations, four of the folds are used for training and the fifth fold for evaluation. Then, the folds are shuffled, and the next iteration begins.

The final tuned hyperparameters are detailed in Table 2. Any parameter not included in the table is set to its default value.

Table 2. Optimal Hyperparameters

Hyperparameter	Value	Description
n_estimators	200	Number of trees
criterion	Gini	Function to measure split quality
max_depth	10	Max depth of a tree
max_features	log2	Number of features to consider while splitting
min_samples_split	2	Number of samples needed to split an internal node

This process is repeated for all five models for the minimal dataset as well.

7. Model Evaluation

The five models are trained and evaluated on both the initial and minimal datasets. As accuracy is not a sufficient metric in the medical field, the models are compared based on a multitude of measures. The particular metrics used are detailed below.

7.1. Basic definitions

- True Positive (TP): The count of malignant tumors classified as malignant by the model.
- True Negative (TN): The count of benign tumors classified as benign by the model.
- False Positive (FP): The count of benign tumors classified as malignant by the model.
- False Negative (FN): The count of malignant tumors classified as benign by the model.

7.2. Metrics used

- Accuracy: The percentage of tumors whose malignancy was correctly predicted.
- Precision: TP / (TP + FP)
- Recall: TP / (TP + FN)
- F1 Score: (Precision * Recall) / (Precision + Recall)
- ROC-AUC Score: Area under the Receiver Operating Characteristic Curve.

All five models are trained and evaluated using these metrics on both the initial and minimal datasets. Their performances are collated and analyzed in the Results section.

8. Results and Discussion

The performances of the models on both datasets are discussed in this section.

8.1. Initial Dataset

The Random Forest model outperforms all other models by achieving a perfect score on all the measured metrics. The Support Vector Machine model performs the second best, supporting the claims made in the related works previously analyzed in the paper. The Decision Tree, as expected, provides a performance inferior to the ensemble method. The K-Nearest Neighbors performs the worst, and we conclude that clustering methods may not be ideal for malignant tumor detection.

Table 3. Model comparison on the initial dataset

Model	Accuracy (%)	Precision	Recall	F1 Score	ROC-AUC
Random Forest	100	1.00	1.00	1.00	1.00
Support Vector Machine	98.60	0.97	1.00	0.99	N/A
Decision Tree	95.10	0.93	0.97	0.95	0.95
Multilayer Perceptron	97.20	0.97	0.97	0.97	0.99
K-Nearest Neighbors	94.41	0.93	0.95	0.94	0.99

8.2. Minimal Dataset

Table 4. Model comparison on the minimal dataset

Model	Accuracy (%)	Precision	Recall	F1 Score	ROC-AUC
Random Forest	99.30	0.99	1.00	0.99	0.99
Support Vector Machine	97.90	0.97	0.98	0.98	N/A
Decision Tree	95.80	0.93	0.98	0.96	0.96
Multilayer Perceptron	96.50	0.94	0.98	0.96	0.99
K-Nearest Neighbors	93.01	0.90	0.95	0.93	0.98

The Random Forest model continues to achieve the best metrics, with an accuracy of 99.30% and a perfect recall score. This is ideal as the model does not return any false negatives. Furthermore, the model performs at par with the model trained on the initial dataset, having used just half the number of features. A smaller dataset will allow clinicians to work more efficiently in the field. This will also reduce the computational power necessary to train and run the model. Fig. 2 and Fig. 3 depict the ROC curves for the RF model on the initial and minimal datasets. The model outperforms similarly trained models in related literature, with Naji et al. [12] achieving accuracies of just over 96% in their experiment.

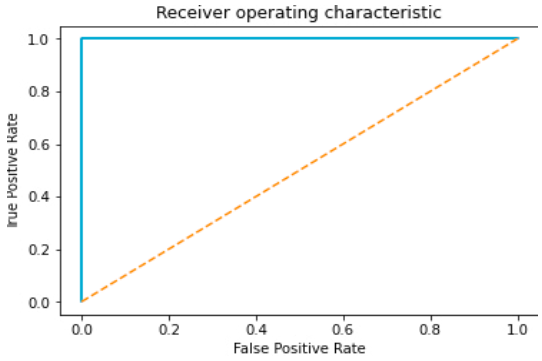

Fig. 3. ROC for RF model on Initial Dataset

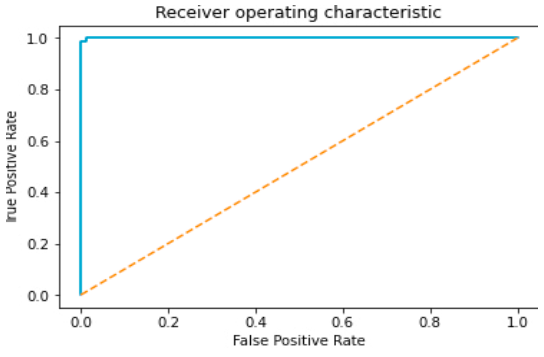

Fig. 4. ROC for RF Model on Minimal Dataset

9. Future Work and Conclusion

In this paper, the Random Forest algorithm was trained, and its hyperparameters were tuned to diagnose breast cancer efficiently and accurately. Further, these RF models are benchmarked against four other supervised learning techniques. The RF model trained on the initial dataset achieves perfect metrics, far surpassing the performances of the other four models on the same dataset. The models trained on the minimal dataset paint a similar picture- the RF model achieves near-perfect metrics while the other models lag behind. Thus, the Random Forest model may be used by oncologists in the field to confirm their diagnoses. It may also be used in rural areas to increase the reach of breast cancer awareness and treatment. This will definitely result in countless saved lives, with timely care being a scarcity in many developing countries such as India.

Future work may be found in developing image processing techniques for the detection of malignant tumors. The experiment in this paper utilized numerical measures extracted from images. Increases in efficiency may be found if it was not necessary to extract said measurements and directly feed the photos of the tumors into the model. Various deep learning algorithms are currently available for this use case and may be harnessed in future experiments.

Acknowledgment

Manas Minnoor and Veeky Baths would like to thank the UC Irvine Machine Learning Repository for providing them with the dataset for this experiment. Veeky Baths was responsible for the design of the experiment and revision of drafts. Manas Minnoor was responsible for the model training as well as the draft publication.

References

[1] Sung H, Ferlay J, Siegel RL, Laversanne M, Soerjomataram I, Jemal A, et al. Global Cancer Statistics 2020: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries. CA Cancer J Clin. 2021 May;**71**(**3**):209–49.

[2] Wang L. Early Diagnosis of Breast Cancer. Sensors (Basel). 2017 Jul 5;**17**(**7**):E1572.

[3] Dua, D. and Graff, C. (2019). UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science.

[4] Ford D, Easton DF, Stratton M, Narod S, Goldgar D, Devilee P, et al. Genetic heterogeneity and penetrance analysis of the BRCA1 and BRCA2 genes in breast cancer families. The Breast Cancer Linkage Consortium. Am J Hum Genet. 1998 Mar;**62**(**3**):676–89.

[5] Jackins V, Vimal S, Kaliappan M, Lee MY. AI-based smart prediction of clinical disease using random forest classifier and Naive Bayes. J Supercomput. 2021 May 1;**77**(**5**):5198–219.

[6] Sarica A, Cerasa A, Quattrone A. Random Forest Algorithm for the Classification of Neuroimaging Data in Alzheimer's Disease: A Systematic Review. Frontiers in Aging Neuroscience [Internet]. 2017 [cited 2022 Apr 14];9. Available from: https://www.frontiersin.org/article/10.3389/fnagi.2017.00329

[7] Byeon H. Is the Random Forest Algorithm Suitable for Predicting Parkinson's Disease with Mild Cognitive Impairment out of Parkinson's Disease with Normal Cognition? Int J Environ Res Public Health. 2020 Apr;**17**(**7**):2594.

[8] Khalilia M, Chakraborty S, Popescu M. Predicting disease risks from highly imbalanced data using random forest. BMC Medical Informatics and Decision Making. 2011 Jul 29;**11**(**1**):51.

[9] Alam MdZ, Rahman MS, Rahman MS. A Random Forest based predictor for medical data classification using feature ranking. Informatics in Medicine Unlocked. 2019 Jan 1;15:100180.

[10] Anji Reddy V, Soni B. Breast Cancer Identification and Diagnosis Techniques. In: Rout JK, Rout M, Das H, editors. Machine Learning for Intelligent Decision Science [Internet]. Singapore: Springer; 2020 [cited 2022 Apr 14]. p. 49–70. (Algorithms for Intelligent Systems). Available from: https://doi.org/10.1007/978-981-15-3689-2_3

[11] Vaka AR, Soni B, K. SR. Breast cancer detection by leveraging Machine Learning. ICT Express. 2020 Dec 1;**6**(**4**):320–4.

[12] Naji MA, Filali SE, Aarika K, Benlahmar EH, Abdelouhahid RA, Debauche O. Machine Learning Algorithms For Breast Cancer Prediction And Diagnosis. Procedia Computer Science. 2021 Jan 1;191:487–92.

[13]Priyanka KS. A Review Paper on Breast Cancer Detection Using Deep Learning. IOP Conf Ser: Mater Sci Eng. 2021 Jan;**1022**(**1**):012071.

[14] Azar AT, El-Metwally SM. Decision tree classifiers for automated medical diagnosis. Neural Comput & Applic. 2013 Dec 1;**23**(**7**):2387–403.

[15] Desai M, Shah M. An anatomization on breast cancer detection and diagnosis employing multi-layer perceptron neural network (MLP) and Convolutional neural network (CNN). Clinical eHealth. 2021 Jan 1;4:1–11.

[16] Polat K, Güneş S. Breast cancer diagnosis using least square support vector machine. Digital Signal Processing. 2007 Jul 1;**17**(**4**):694–701.

[17] Sarkar M, Leong TY. Application of K-nearest neighbors algorithm on breast cancer diagnosis problem. Proc AMIA Symp. 2000;759–63.

[18] Yan K, Zhang D. Feature selection and analysis on correlated gas sensor data with recursive feature elimination. Sensors and Actuators B: Chemical. 2015 Jun 1;212:353–63.
[19] Cheng Q, Varshney PK, Arora MK. Logistic Regression for Feature Selection and Soft Classification of Remote Sensing Data. IEEE Geoscience and Remote Sensing Letters. 2006 Oct;**3**(**4**):491–4.
[20] Qi Y. Random Forest for Bioinformatics. In: Zhang C, Ma Y, editors. Ensemble Machine Learning: Methods and Applications [Internet]. Boston, MA: Springer US; 2012 [cited 2022 Apr 14]. p. 307–23. Available from: https://doi.org/10.1007/978-1-4419-9326-7_11
[21] Ali J, Khan R, Ahmad N, Maqsood I. Random Forests and Decision Trees.
[22] T R P. A Comparative Study on Decision Tree and Random Forest Using R Tool. IJARCCE. 2015 Jan 30;196–9.
[23] Lee T-H, Ullah A, Wang R. Bootstrap Aggregating and Random Forest. In: Fuleky P, editor. Macroeconomic Forecasting in the Era of Big Data: Theory and Practice [Internet]. Cham: Springer International Publishing; 2020 [cited 2022 Apr 14]. p. 389–429. (Advanced Studies in Theoretical and Applied Econometrics). Available from: https://doi.org/10.1007/978-3-030-31150-6_13
[24] Kumari, Madhu, and Vijendra Singh. "Breast Cancer Prediction system." Procedia Computer Science, Elsevier, 132 (2018): 371-376.