RESEARCH ARTICLE OPEN ACCESS

RF-LR Ensemble Classifier for Breast Cancer Detection

Thachayani M [1], Chaitanya Sai Jangam, Kalyan T, SriManjunadh Maddukuri, Sangadi Manikanta

Dept. of Electronics and Communication Engineering, Puducherry Technological University, Puducherry, India

ABSTRACT

An ensemble learning based classifier to aid in the early diagnosis of breast cancer is presented in this paper. Four machine learning algorithms are investigated and the random forest classifier is selected as the base model based on the performance. An ensemble model is created using bagging and boosting techniques employing the base classifier. Logistic regression is applied as the meta classifier for stacking. The developed ensemble model resulted in an improved accuracy of 96.49% compared to the 92.55% accuracy of the baseline model.

Keywords — Breast cancer detection, Ensemble learning, Exploratory analysis, Logistic regression, Random forest

I. INTRODUCTION

Breast cancer is a major threat particularly for the female population which ranks second leading cause of cancer-related deaths in women [1]. Survival rate is higher in cases where the detection is done during the early stages while it is still localized and not yet spread to other parts of the body. The motivation for this work arises from the potential for machine learning to assist in early and accurate diagnosis of breast cancer which could significantly improve the recovery rates and reduce fatalities.

Several researches focusing on utilizing the power of machine learning models to enhance diagnostic accuracy is reported in literature. An extensive review of literature related to machine learning based breast cancer detection is reported in [2]. Some of the closely related work on application of machine learning and particularly ensemble based learning techniques for breast cancer detection is discussed here. Support Vector Machine (SVM), Artificial Neural Network (ANN) and Naıve Bayes Algorithms are investigated for their suitability in detecting breast cancer and it is observed that the SVM classifier outperformed the other two with an accuracy of 96.72% [3]. A stacking classifier is implemented using K-Nearest Neighbor (KNN), SVM, and Random Forest (RF) as base classifiers and Logistic Regression as meta classifier considering 20 features of the breast cancer data and achieved an accuracy of 97.20% [4]. Employed an ensemble model created with Bayesian network and Radial Basis Function and achieved a prediction accuracy of 97% [5]. The biased results due to class imbalance in the dataset is observed in the decision tree classifier and adaptive boosting is employed to address this issue. Significant improvement in accuracy is reported with boosting [6]. Applied t-distributed stochastic neighbour embedding (t-SNE) for cost optimization and dimension reduction. Then snapshot ensemble technique is used to combine the predictions from the base models leading to achieve an accuracy of 86.6% [7]. In [8], the authors investigated on the averaged perceptron model for breast cancer detection and recorded an accuracy score of 0.984 with zero false negatives. Cross validation approach is utilized for tuning the hyperparameters and its impact on accuracy performance of random forest, extra tree (ET), and support vector machine classifiers are analysed in [9]. It is observed that the tuned SVM classifier outperformed the other two with an accuracy of 97.78%. From the survey, it is observed that variants of SVM and RF outperformed several other base classifier models for breast cancer detection application. Further it is inferred that several factors such as choice of features, number of features; and processes such as cross-validation, boosting, parameter tuning and ensemble has a significant impact on the overall performance. In this paper, an ensemble model based on bagged and boosted RF base classifiers combined using a Logistic Regression (LR) meta classifier is investigated for breast cancer detection application. The following section enumerates the methodology used and Section III presents the key results and Section IV concludes the paper.

II. METHODOLOGY

Fig.1 shows the key steps involved in the process of training and testing the proposed classifier system. The process starts with the pre-processing of the data, which involves cleansing and other preliminary processing aiming to validate the data set for completeness and integrity. Then the data set is split into training and test sets. Exploratory analysis is carried-out on the training data set to visualize the relevance of various predictors to the diagnosis. The correlation between the predictors are plotted using different visualization tools to identify the relevant predictors. The ensemble model is trained with the selected predictors and cross-validated to ensure that the model is not over-fitted.

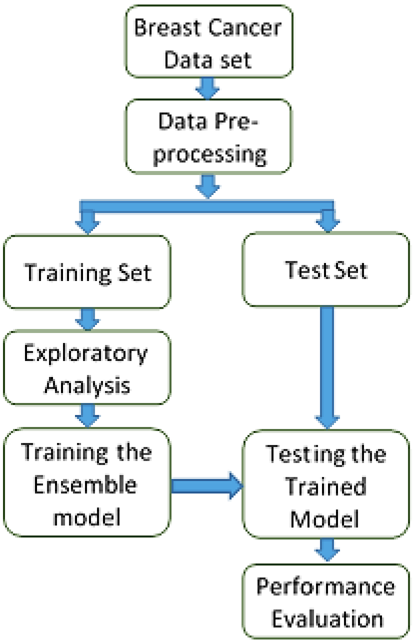


Fig.1 Methodology

The testing of the trained model is done using the test data and the performance of the proposed classifier is assessed in terms of parameters such as accuracy, precision, recall and F1-score.

III. RESULTS AND DISCUSSION

The WISCONSIN breast cancer data set which consists of 569 samples with 30 attributes derived from ten main properties of breast cell nuclei is used to train and test the classifiers [10]. This information characterizes the cell nuclei and acquired from digitized image of a fine needle aspirate (FNA) of a breast mass. Label encoding of the diagnosis parameter is done during the pre-processing. Standard scaling of data is done for uniformity. Then the data set is split into training and test sets comprising of randomly selected 70% and 30% of the original data samples. Exploratory analysis utilized various data visualization tools to explore the relationship between the predictors or features and the target parameter. It also involves analyzing correlation between the predictors. Multiple levels of analysis are carried-out to identify the relevant predictors in order to achieve optimum performance. From the first level of analysis and the literature the ten mean attributes are chosen out of the thirty attributes as the predictors and are analyzed further. The correlation matrix plot is observed for the first-level predictors and the target diagnosis. This is shown in Fig.2. From this figure, it can be observed that the fractal_dimension feature is feebly correlated to the target compared to all the other nine attributes. Hence these nine except the fractal_dimension feature is analyzed further. Pair-wise correlation plots are plotted. Some of these plots are shown in Fig.3 as samples, which depicts the attributes, radius_mean, texture_mean, perimeter_mean and area_mean. From this figure, relevance of the texture and area or perimeter features in the classification is evident. Further, the strong correlation between radius and the area as well as perimeter is also showcased.

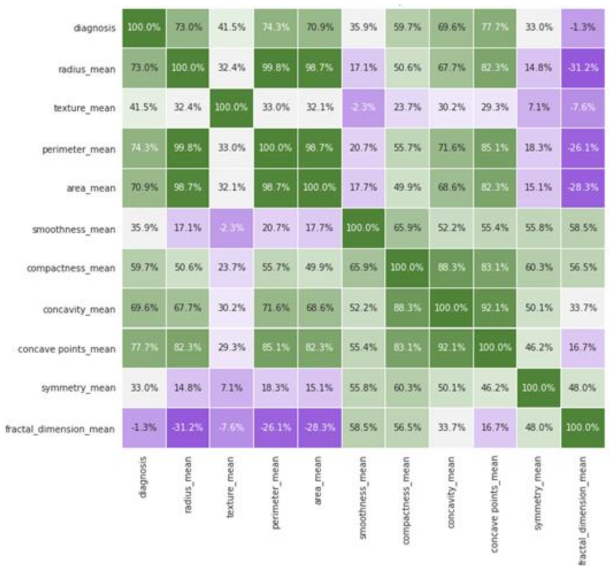


Fig. 2 Correlation matrix plot

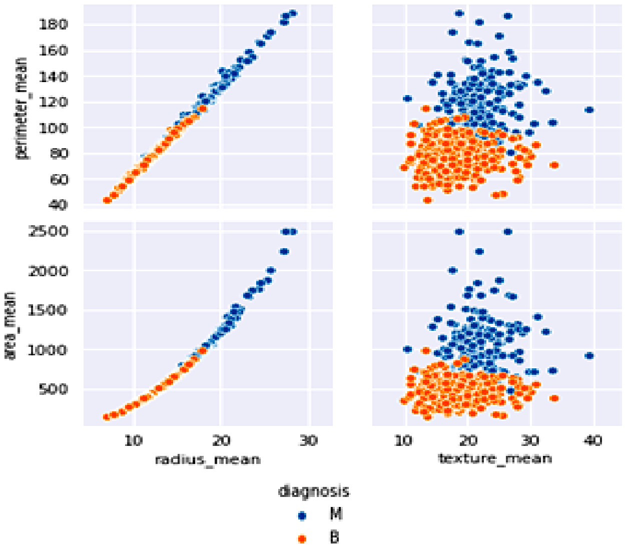


Fig. 3 Sample pair-wise plots

Grid based ten-fold cross validation is done for the base classifiers. Hyperparameters such as model complexity and training rate are optimized to enhance the performance. The models considered are LR, RF, SVM and KNN. The individual model is trained and then tested. The F1 score and accuracy score obtained are listed in Table I. This table reveals that the RF classifier outperforms the other considered classifiers and hence is selected as the base model for creating the ensemble classifier.

TABLE I
PERFORMANCE OF THE BASIC MODELS

Sl. no	Model	F1 Score	Accuracy Score
0	LR	0.916010	0.909574
1	RF	0.992126	0.925532
2	SVM	1.000000	0.909574
3	KNN	0.923885	0.914894

In bagging, multiple models of the same base classifier is trained with different data sub-sets generated using boot-strap sampling and final decision will be based on the aggregate decisions formed by voting. This process aids to improve the diversity of the classifier leading to more robust performance. Bagging and boosting techniques are used to form an ensemble of the RF model and aggregation is done using an LR meta classifier. The performance of the ensemble classifier is evaluated in terms of accuracy, precision, recall and F1-score.

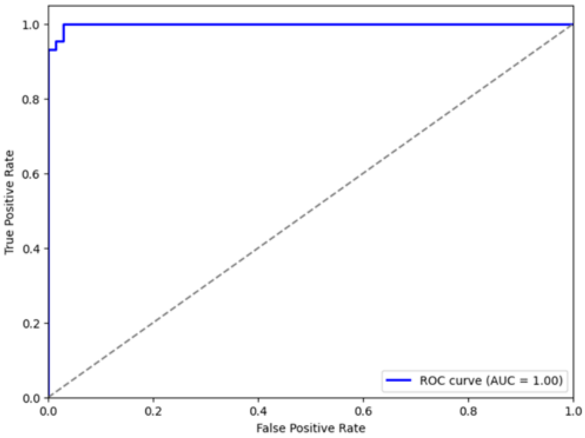
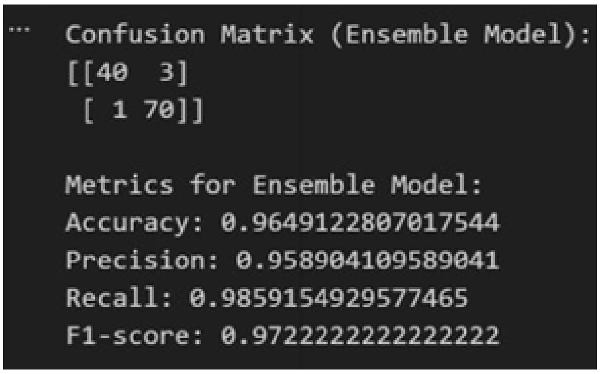


Fig. 4 Receiver operating characteristics (RoC)



```
...  Confusion Matrix (Ensemble Model):
     [[40  3]
      [ 1 70]]

     Metrics for Ensemble Model:
     Accuracy: 0.9649122807017544
     Precision: 0.958904109589041
     Recall: 0.9859154929577465
     F1-score: 0.9722222222222222
```

Fig. 5 Results of the performance evaluation

The receiver operating characteristics of the ensemble model is shown in Fig. 4 and the output screenshot showing the confusion matrix and performance metrics is presented in Fig.5. From these figures it is evident that the proposed ensemble classifier based on RF and LR classifier performs significantly better than the base RF classifier with an improved accuracy score of 0.965 compared to the original score of 0.9255.

IV. CONCLUSIONS

This paper presents an ensemble earning classifier with random forest as the base model for assisting in the accurate diagnosis of breast cancer. Using exploratory analysis nine out of the thirty predictors is chosen for classification. Four base classifiers are investigated and based on the performance, the random forest is chosen as the base model and an ensemble model is created by using bagging and boosting techniques. Logistic regression model is used for aggregating and forming the final prediction. The ensemble classifier exhibited an improved accuracy of 96.49% compared to the 92.55% accuracy of the base model.

REFERENCES

[1] Breast Cancer Statistics, available at https://www.cdc.gov/ breastcancer/statistics.

[2] Jafari, Ali, "Machine-Learning Methods in Detecting Breast Cancer and Related Therapeutic Issues: A Review." Computer Methods in Biomechanics and Biomedical Engineering: Imaging & Visualization, vol. 12, 2024.

[3] Md. I. H. Showrov, M. T. Islam, Md. D. Hossain, and Md. S. Ahmed, ''Performance comparison of three classifiers for the classification of breast cancer dataset,'' in Proc. 4th Int. Conf. Electr. Inf. Commun. Technol. (EICT), Dec. 2019, pp. 1–5.

[4] M. R. Basunia, I. A. Pervin, M. Al Mahmud, S. Saha, and M. Arifuzzaman, ''On predicting and analyzing breast cancer using data mining approach,'' in Proc. IEEE Region 10 Symp. (TENSYMP), Jun. 2020, pp. 1257–1260.

[5] M. A. Jabbar, ''Breast cancer data classification using ensemble machine learning,'' Eng. Appl. Sci. Res., vol. 48, no. 1, pp. 65–72, 2021.

[6] T. A. Assegie, R. L. Tulasi, and N. K. Kumar, ''Breast cancer prediction model with decision tree and adaptive boosting,'' IAES Int. J. Artif. Intell., vol. 10, no. 1, p. 184, 2021.

[7] N. Sharma, K. P. Sharma, M. Mangla, and R. Rani, ''Breast cancer classification using snapshot ensemble deep learning model and t-distributed stochastic neighbor embedding,'' Multimedia Tools Appl., vol. 82, no. 3, pp. 4011–4029, Jan. 2023.

[8] V. Birchha and B. Nigam, ''Performance analysis of averaged perceptron machine learning classifier for breast cancer detection,'' Proc. Comput.Sci., vol. 218, pp. 2181–2190, 2023.

[9] N. Mohd Ali, R. Besar, and N. A. A. Aziz, ''A case study of microarray breast cancer classification using machine learning algorithms with grid search cross validation,'' Bull. Electr. Eng. Informat., vol. 12, no. 2, pp. 1047–1054, Apr. 2023.

[10] Wisconsin Breast Cancer Dataset from UCI Machine Learning Repository, available at https://archive.ics.uci.edu/ml/datasets/ Breast+Cancer+Wisconsin+%28Diagnostic%29.