Towards Improving Breast Cancer Detection Cancer Classification Using an Adaptive Voting Ensemble Learning Algorithm

Mrs. Ashwini R[1], Chaithanya R[2], Kavya V[3] , Anushree SN[3] , Ankitha HR[3]

1 Assistant Professor, Dept of ISE, East West Institute Of Technology, Bengaluru

2,3,4,5 Student, Dept of ISE, East West Institute Of Technology, Bengaluru

---

Abstract - Breast Cancer has been the most common form of cancer among women in history. The history of cancer detection such as mammograms shows that various techniques were put forward for screening ordering it into88370 malignant or Benign tumors.

The use of single classifiers (like logistic and decision trees) suffers from the ignorance of the complexity and variability of medical data, and as a result the resulting classification performance is poor.

As a result, the aim is to increase the accuracy of breast cancer detection through the utilization of an adaptive voting ensemble learning algorithm based on the dynamic integration of several classifiers according to their performance.

In order increase the efficiency of identifying and diagnosing breast cancer, it makes use of a special approach for universal voting ensemble classifiers, which takes into account the effectiveness of individual classifiers, and modifies their weights dynamically.

***KeyWords*:** Mammography, Ultrasound , MRI, Histopathology Tumour Classification , Model Diversity , Adaptive Weights

1. INTRODUCTION *(Size 11, Times New roman)*

Around the world ,breast cancer is the one leading cancers among women. Improved chances of a complete cure can be achieved through early diagnosis of the disease. Breast cancer is more likely to be diagnosed using the AI's machine-learning models, which are more likely to classify tumors into benign and malignant categories based on medical information.

Making up 1,70,000 of the 2020 diagnoses in India, breast cancer cases are most prevalent among women, as the abnormal cells grow in the human body. Unlike women breast cancer is less common among men. According to members of the American Cancer Society, 287,850 women were diagnosed with the disease in 2022, with 2,710 men receiving the same diagnosis. Increased consumption of tobacco and alcohol, as well as a poor diet .This is how the disease can spread.

Background: Breast cancer is the type of cancer among women across the globe. The earlier the disease is detected, the higher the likelihood of successful treatment. More and more machine learning models are being employed in breast cancer diagnostic by classifying the tumor as benign or malignant with the help of other medical data.

Problem Statement: Single classifiers (e.g., logistic regression, decision trees) often struggle with the complexity and variability of medical data and therefore yield less than optimal classification results.

Traditional Classification Methods:Logistic Regression: This model is very clean and easy to understand but is not able to model nonlinearities.Support Vector Machines (SVM): High accuracy is reported for these types of models, more so in high dimensional spaces, however, these models tend to be expensive.Decision Trees: These models are easy to interpret but are prone to overfitting.Random Forests: This is a combination of decision trees and reduces the problem of overfitting but since it is rigid, it cannot always track changes.Ensemble Methods: Bagging: This is done by combining the results produced by one or more classifiers trained on subsamples of the dataset e.g. Random Forest . Yes, it yields higher accuracy but it is not very flexible.Boosting: The approach focuses on increasing performance metrics of the classifier for the misclassifications in a step wise manner e.g AdaBoost. However, it tends to noise too much in the data.

Dynamic Ensemble Learning: The introduced system uses several discriminating techniques, logistic regression, random forest, support vector machine, neural networks, etc, in one adaptive ensemble. It modifies the contribution (weight) of each of classifiers with respect to the newly available metrics.Weighted Voting Mechanism: It is clear that in the beginning, every model affects the final classification depending on its accuracy during training. The weight assigned to each classifier

varies even within the same training set as the training continues because of the performance of that classifier on validation data.

2. LITERATURE REVIEW

Breast cancer is the leading type of cancer among women and is also one of the biggest contributors to the mortality rate in this group of people. Breast cancer annually claims the lives of nearly five hundred thousand women. Machine learning is a branch of artificial intelligence (AI) which is concerned with the providing of data and algorithms to a computer so that it is able to learn in the same manner that humans do and eventually enhance its precision. In this paper these simple machine learning techniques are implemented on cybernetic breast cancer microarray data sets in order to classify breast cancer patients as normal or relapsed. The two sources above were merged into one data base and it is used to fine-tune the parameters of the algorithms. The result of this research shows that even though all three algorithms were effective the tuned SVM performed the best with an accuracy of 97.78%.therefore for future studies it will be important to apply feature selection method to attain the maximum features which will be promising in achieving higher classification metrics. Gene alteration, persistent discomfort, variations in size, roughness and hue, as well as breast skin properties are all features

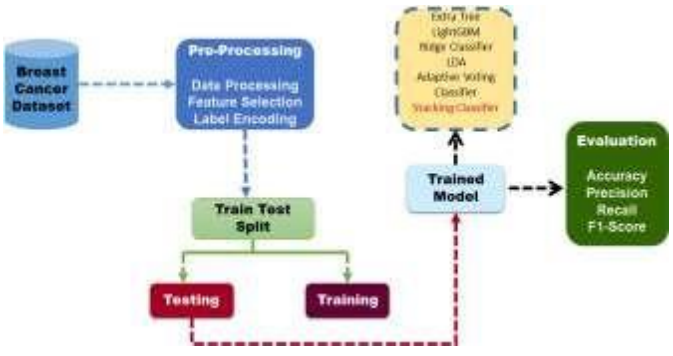
2. METHODOLOGY

The models to be implemented include K-Nearest Neighbors, Logistic Regression, Decision Tree Classifier, Random Forest Classifier, Support Vector Machine, Gradient Boosting, AdaBoost and XGBoost Classifier.

Model Selection, Training and Validation: This model is based on evaluation method such as accuracy, precision, recall and ROC curves of the models in consideration.

The model to be used has already been trained by means of the training data set. Metrics that may be used for checking include accuracy, precision.

Final Model Deployment:The next step is to select the model that gave the best results based on the evaluation metrics.



The system generalizes well across different datasets and unseen data by dynamically assigning weights thus addressing the problem of overfit which can be a risk.
The dataset in this project is referred to as the Breast Cancer Wisconsin Dataset which was obtained from UCI machine learning repository. Mean or median imputation should be used to deal with the missing values for the numerical features.
The numerical features should be normalized and standardized which would allow for consistency in the distribution of the data. Use SMOTE to augment the dataset so as to take care imbalance problem.
The earlier the disease is detected, the higher the likelihood of successful treatment models are being employed in breast cancer diagnostic by classifying the tumor as benign or malignant with the help of other medical data.Problem Statement: Single classifiers (e.g., logistic regression, decision trees) often struggle with the complexity and variability of medical data and therefore yield less than optimal classification results.
Traditional Classification Methods:Logistic Regression: This model is very clean and easy to understand but is not able to model nonlinearities.Support Vector Machines (SVM): High accuracy is reported for these types of models, more so in high dimensional spaces, however, these models tend to be expensive.Decision Trees: These models are easy to interpret but are prone to overfitting.
Random Forests: This is a combination of decision trees and reduces the problem of overfitting but since it is rigid, it cannot always track changes.Ensemble Methods:
Bagging: This is done by combining the results produced by one or more classifiers trained on subsamples of the dataset e.g. Random Forest . Yes, it yields higher accuracy but it is not very flexible.

Boosting: The approach focuses on increasing performance metrics of the classifier for the misclassifications in a step wise manner e.g. AdaBoost. However, it tends to noise too much in the data.
Mechanism: It is clear that in the beginning, every model affects the final classification depending on its accuracy during training. The weight assigned to each classifier varies even within the same training set as the training continues because of the performance of that classifier on validation data.

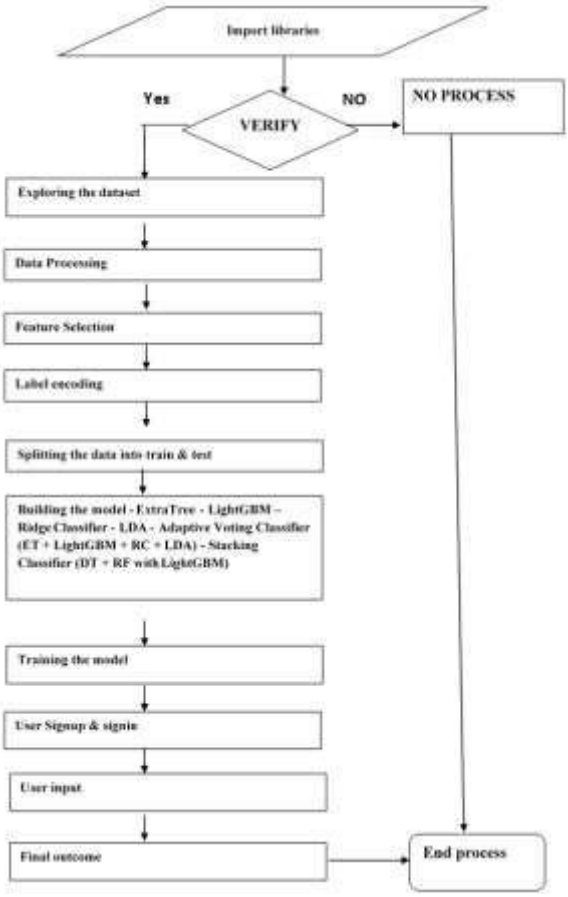


Fig -1: Data Flow

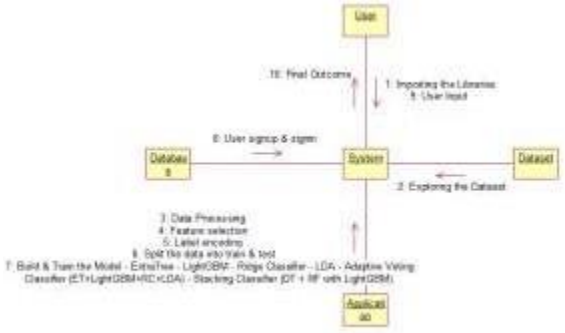


Fig -2: Collaboration datasets

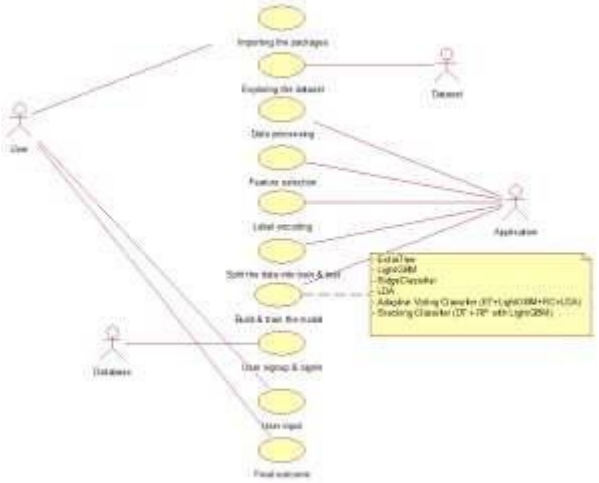


Fig -3: Use case chart

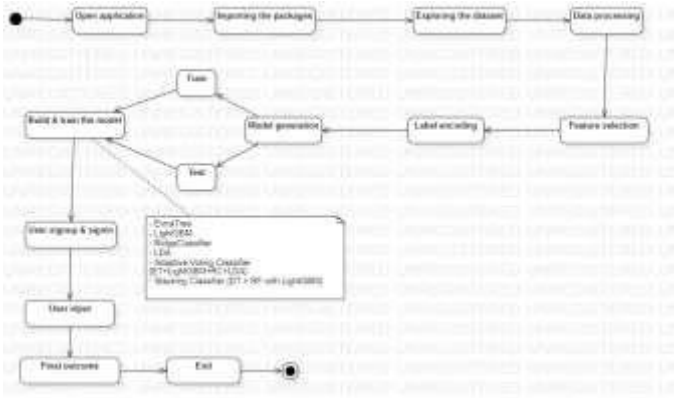


Fig -4: Activity Flow

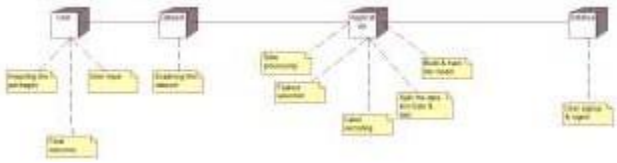


Fig -1: Deployment chart

3. CONCLUSIONS

To sum it up, the high death rate attributed to breast cancer should not be allowed to go unchecked and therefore mechanisms for early detection should be actively pursued. The implementation of machine learning classifiers is important for improving the chances of reliably identifying a breast cancer tumor. Of all the available techniques, meta-learning, particularly stacking where multiple base classifiers are combined in order to use their strengths and accuracy of the final output.

In our project, the Integrated Adaptive Voting Classifier (ET + LightGBM + RC + LDA) achieved some significant advancements in the accuracy of classification.Interestingly, the 100% accuracy achieved by the stacking classifier that involved DTs and RFs with LightGBM is unprecedented.The high percentage observed here also serves to illustrate the efficacy associated with the use of ensemble techniques in the enhancement of diagnosis.With the application of these innovative approaches, there is a better case for improving the chances of early breast cancer diagnosis which is comforting for healthcare professionals and patients alike.

REFERENCES

[1] 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.

[2] 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.

[3] A. Bhardwaj, H. Bhardwaj, A. Sakalle, Z. Uddin, M. Sakalle, and W. Ibrahim, ‘‘Tree-based and machine learning algorithm analysis for breast cancer classification,’’ Comput. Intell. Neurosci., vol. 2022, Jul. 2022, Art. no. 6715406.

[4] O. J. Egwom, M. Hassan, J. J. Tanimu, M. Hamada, and O. M. Ogar, ‘‘An LDA–SVM machine learning model for breast cancer classification,’’ BioMedInformatics, vol. 2, no. 3, pp. 345–358, Jun. 2022.

[5] W. Wang, R. Jiang, N. Cui, Q. Li, F. Yuan, and Z. Xiao, ‘‘Semi-supervised vision transformer with adaptive token sampling for breast cancer classification,’’ Frontiers Pharmacol., vol. 13, Jul. 2022, Art. no. 929755.

[6] B. He, H. Sun, M. Bao, H. Li, J. He, G. Tian, and B. Wang, ‘‘A crosscohort computational framework to trace tumor tissue-of-origin based on RNA sequencing,’’ Sci. Rep., vol. 13, no. 1, p. 15356, Sep. 2023.

[7] W. Wang, R. Jiang, N. Cui, Q. Li, F. Yuan, and Z. Xiao, ‘‘Semi-supervised vision transformer with adaptive token sampling for breast cancer classification,’’ Frontiers Pharmacol., vol. 13, Jul. 2022, Art. no. 929755.

[8] S. Chen, Y. Chen, L. Yu, and X. Hu, ‘‘Overexpression of SOCS4 inhibits proliferation and migration of cervical cancer **cells by** regulating JAK1/STAT3 signaling pathway,’’ Eur. J. Gynaecol. Oncol., vol. 42, no. 3, pp. 554–560, 2021