A NOVEL ADAPTIVE ENSEMBLE LEARNING MODEL FOR ACCURATE BREAST CANCER CLASSIFICATION

Arshia Singhal[1], Khushi Chauhan[2], Ruchira Kumar[3], Shubhi Verma[4], Astha Sharma[5], Ashwini Kumar[6]
[1,2,3,4] Department of Electronics and Communication Engineering,Indira Gandhi Delhi Technical University(IGDTUW), New Delhi, India

Abstract - Breast cancer is still among the most widespread and life-threatening diseases afflicting women on this earth. Early and accurate detection is of utmost importance as it defines treatment availability and prognosis. Hence this research presents a unique classification approach applying an adaptive voting ensemble learning algorithm targeted toward increasing diagnostic accuracy in breast cancer detection. The proposed method incorporates various decision tree-based classifiers in combination with Decision Tree and Random Forest algorithms. This ensemble model adapts dynamically to the performance of each individual classifier, by putting more weights on the classifiers which display higher correctness. Rather than depending on a fixed voting strategy, the proposed model has an adaptive mechanism where the predictive strength of each base learner modulates its contribution to the final decision, resulting in a reliable and robust classification that does not overload any specific classifier. The system performance is evaluated on the Breast Cancer Wisconsin (Diagnostic) Dataset obtained from the UCI Machine Learning Repository for training and testing. The proposed model is evaluated using standard evaluation metrics, such as accuracy, precision, recall, F1-score, and AUC-ROC. Results suggest that the adaptive ensemble approach greatly increases the model's capacity to distinguish malignant from benign cases, thereby assisting in early diagnosis and efficient treatment planning.

Keywords: Breast Cancer, Ensemble Learning, Adaptive Voting, Decision Tree, Random Forest, Classification Accuracy, Medical Diagnosis

I. INTRODUCTION

Irrespective of the modern era of research, breast cancer still continues to be the major reason for cancer death in women globally. Timely and accurate diagnosis has become even more crucial for treatment and an increase in patient survival. While traditionally efficient, most diagnostic methods can be augmented by established machine learning techniques to improve diagnostic accuracy and reliability. This project thus focuses on designing an advanced breast cancer classification system using an adaptive voting ensemble learning algorithm. The proposed breast cancer classification system combines Decision Tree and Random Forest classifiers to take advantage of their complementary strengths into the construction of a much stronger and, therefore, accurate model. The adaptive voting mechanism will dynamically change the weights assigned to each classifier as per their performance level. Therefore, more accurate models will have more influence on the final output. This adaptive approach, however, addresses the problems of overfitting, bias, and inconsistent performance usually shown by individual classifiers.

The aim of this current work is to develop a trustworthy and dependable tool that can classify breast cancer for health practitioners during early diagnosis and treatment planning. It is expected that the effect of improving diagnosis will immediately result in better outcomes for the patient and reduce mortality associated with breast cancer.

II. LITERATURE SURVEY

In medicine, ensemble learning methods applied for classification pose interesting questions of their capacity to enhance predictive performance. Zhang et al. (2018) looked into ensemble learning methods with breast cancer classifiers such as DT and RF and concluded that those multiple classifier ensemble methods performed better than single classifier settings in breast cancer classification. The authors stated that such ensemble learning methods, particularly with some mechanisms such as weighted voting, demonstrated better generalization capability with the ability to adapt dynamically according to strengths and weaknesses of the constituent classifiers [1]. This is also the point from where one might come to think of accuracy in classifying being improved by leveraging diverse models that would counterbalance their limitations. Adaptive voting beyond ensemble classification has also generated quite some interest and holds promise in using this methodology to further boost prediction reliability. Liu and Wang (2019) proposed adaptive voting in which weights of classifiers are adapted in real-time based on performance. Their classification accuracy increased significantly using the system. This shows that adjustment of the weights dynamically should therefore be tending towards being in favor of those classifiers that are most accurate. This adaptability thus becomes a key consideration for the ensemble system whereby predictions by the model are compared against those from the most trusted classifiers [2]. The advantage also overcomes the situation involved in the overfitting and underfitting of adjusting the voting mechanism based on the classifier's performance. Convolutional Neural Networks (CNNs) are yet some of the highly valued deep learning approaches in cancer diagnosis tasks. Zhang et al. (2020) used CNNs for breast cancer image classification and noted that using CNN-based methods could automatically learn hierarchical feature representation from raw medical images, thus improving the diagnostic accuracies over traditional means significantly [3]. This was in agreement with the work of Lin et al. (2021) that showed that improved classification performance on small data sets when using pre-trained CNN models (transfer learning) is particularly valuable for medical applications where data scarcity is a challenge [4]. Moreover, many of the papers proclaimed about the efficacy of hybrid models which combined CNNs with other traditional machine learning techniques. An investigation carried out by Lee et al. (2020) observed the use of CNN models with ensemble learning techniques to enhance breast cancer outcome prediction accuracy. The overall conclusion revealed that hybrid models

had superior predictive sensitivity and specificity, thus making them well suited for medical diagnoses, where reduction of both false positives and false negatives would be necessary [5]. Deep learning and classical approaches are indeed well embedded into the approach proposed in this project, whereby Decision Trees and Random Forests are blended into an adaptive voting ensemble system for classification accuracy ultimate improvement.

There is indeed increasing concern in research towards the emphasis that is to be placed on integrating ML and DL techniques. [6].

III. METHODOLOGY

A. Proposed System

Progressive system will facilitate the classification of breast cancer using various classifiers by an adaptive voting any learning algorithm. This system also includes a decision tree classifier based on a simple but effective method that splits the data on the basis of feature values into a tree structure. It includes Random Forest, an ensemble of decision trees, to provide more accurate prediction with an average of predictions from many trees to combat overfitting. The central theme of the system is Adaptive Voting Ensemble, in which every classifier is assigned weightage according to its validation performance in this dynamic voting scheme of classifiers. Thus, application of this method ensures that those models which were found to be more accurate will have more influence on the final prediction, thereby enhancing classification performance and reliability.

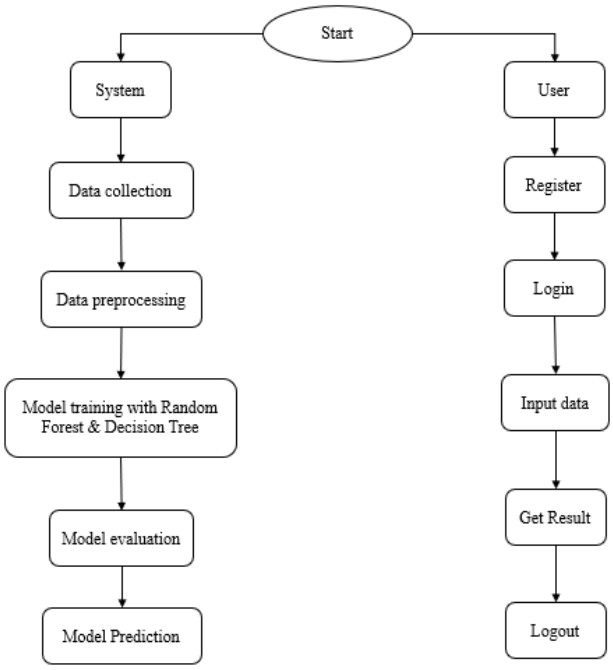

Figure 1 Project Flow

Enter the Extra Trees Classifier (ETC), an advanced method that builds many trees while averaging because of the higher demands on the prediction accuracy obtained. The Light Gradient Boosting Machine (LightGBM) is said to be a blazing fast framework for gradient boosting aimed at models based on trees to enhance predictive performance. The Ridge Classifier (RC) serves as an L2 regularised version of the linear classifier in use, in order to safeguard against any overfitting. Last but not least, Linear Discriminant Analysis (LDA) holds a statistical theory dealing with the linear combination of features that best distinguishes different classes.

IV. IMPLEMENTATION

A. Dataset

Commonly used in binary classifications, The Breast Cancer Wisconsin (Diagnostic) dataset from the UCI Machine Learning Repository has been well known for the problem of breast cancer detection. It contains 569 samples with two possible class labels-malignant (class label 1) and benign (class label 0)-and serves as a basis for modeling with a matured learning technique in medical diagnostics.

A specimen of data is described with respect to 30 numerical features derived from digitized images of FNA (fine needle aspirate) biopsies with breast tissues. These features represent several morphological characteristics of the tumor including radius, texture, perimeter, area, smoothness, compactness, concavity, and symmetry. The measurements aim at very subjective and quantitative aspects of cell nuclei, yielding critical information on tumor morphology and perhaps its malignancy potential.

The dataset holds special significance for development and evaluation of classification algorithms because it captures several real-life clinical scenarios while being small enough to support rapid experimentation and testing. Each attribute in the dataset provides a much more sophisticated description of tumor attributes by an additional statistical perspective-the mean, standard error, and worst value (maximum).

With pre-processing done already, the dataset is thus ready for immediate use in some machine learning applications, providing a benchmark for binary classification widely engaged in model training, testing, and performance comparison in various studies in breast cancer diagnostics.

B. Model Training:

For example, the learning process in machine learning is quite extensive and involves teaching the algorithm to map input data to output labels through many iterations. The training was focused on developing a classifier for determining whether the tumor is of malignant or benign types based on features extracted from Breast Cancer Wisconsin (Diagnostic) Dataset. This training is done by selecting a suitable design of the machine-learning model, splitting data into testing and training, and then training on that data. For example, seventy to eighty percenatage of the data is used for training, while 20-30% is reserved for testing the generated model. Thus, in training for example, algorithms like Decision Trees or Random Forest learn how to form some patterns and relationships with the presented data during that training. Herein, regularization in supervised learning is where the algorithm learns by feeding in examples labeled as malignant or benign; that is, the input is verified against a correct label of what should be found at the output. The trained model then continuously decreased its error, or loss function, through optimization methods, including gradient descent, through the iteration of prediction differentiated from actual results. Based on classifiers but add Adaboost and Random Forest are amateur classifiers. In a sense, Decision Trees are important to the great majority of these classifiers that are trained separately and then unified via an ensemble,

i.e., by integrating several such base classifiers with voting mechanisms. This weighted voting, however, assigns each classifier within the ensemble a weight according to its predictive capability, thereby overshadowing the influence of other, poorer classifiers on the final prediction. Hence their weighted combination enhances both the robustness and accuracy of the overall model from limitations by such individual classifiers. Hyperparameters are involved in model training, which would affect the aspects of the model design desired by the user-maximum depth of Decision Tree, number of trees in random forest, learning rates in gradient-based algorithms, etc. Such optimization of hyperparameters is usually done through grid search or random search. Different parameter values' combinations will be searched for the best configuration. Often in training, cross-validation is used to test the usefulness of the model across different subsets of training data to improve its generalization over previously unseen examples while avoiding overfitting. The performance measure stages of the whole process include model trained tested against to determine how well the model discriminates between malignant and benign.

C. DATA PREPROCESSING

The magnitude of preprocessing data for machine learning is a critical and fundamental step that directly affects model performance and efficiency. This procedure defines how raw data goes through a stage of cleaning and reorganization directed towards preparing data for training machine learning models. Preprocessing for the Breast Cancer Wisconsin (Diagnostic) Dataset is basically concerned with cleaning data and treating missing values. Uniquely, the structure of this dataset is excellent, whereas in practice most datasets may be one with missing or incomplete information, which can affect predictions of the model. In cases where there are missing values in the current study dataset, one plausible approach is imputation, whereby the missing data point is replaced with the mean, median, or mode values of the respective feature. Normalizing and scaling the data is another important step since scaling features affects many other machine learning algorithms such as Support Vector Machines (SVM) or K-Nearest Neighbors (KNN). The features in the dataset vary widely in terms of magnitude; for example, radius or area of tumors. Therefore, normalizing or scaling the data ensures that one feature does not outweigh the other due to its scale. Amongst the common methods to scale data remain Min-Max scaling, which scales features between 0 to 1, and Standardization, scaling features to achieve zero mean and unit variance. The importance of feature selection is the other preprocessing action done for removing irrelevant and redundant features that help in boosting model performance and prevent overfitting. The selection issue is quite trivial for the Breast Cancer dataset because features are so few. Recursive Feature Elimination (RFE) and Random Forest among tree-based methods will rate the importance of relevant features. By emphasizing these relevant features, it is often the case that the model shrinks away from distractions of absent or misleading information and grows to be more generalizing. Categorical variables must be converted to numbers because non-numeric input cannot be processed by most machine learning models. The Breast Cancer data is already numeric; however, other scenarios will require one-hot encoding or label encoding to convert their categorical variables into a numerical format interpretable by the model.

D. Model Training and Classification

1) Decision Tree

Owing to their generalized application and interpretability, the method can be applied for various classification problems. Hence, this project implements Decision Trees for the classification of breast cancer instances into that of malignant and benign. The algorithm creates a tree-like model in which each internal node is a feature test and each leaf node is a class label. Decision Trees partition data repetitively into subsets using feature values, which cause the homogeneity in groups to be more increased. For this project, DT serves as one of the base classifiers in adaptive voting ensemble that offers ease and interpretation. Decision Trees can be overfitting, which is typically found in the tree structures that bottle-neck for poor generalization would form. Then this model will be pruned, which in its grand scheme will ensure better performance on unseen data with the pruning technique applied towards reducing the complexity in determining the large shape of the trees. The approach may improve the accuracy and robustness for benign and malignant tumors using other classifiers in ensemble [7].

2) Random Forest

Random Forest is an ensemble-style learning that marries many Decision Trees for maximization of prediction accuracy and robustness. The Random Forest is the base classifier used for the adaptive voting ensemble learning framework developed in this project. The construction of multiple decision trees across sets of training data lowers the chances of overfitting, usually seen with a single Decision Tree. Aggregated predictions of all trees give a final verdict on majority voting, improving model accuracy generally. That is especially true for classification breast cancer, because of great dimensionality handling while capturing complex relationships from the dataset. Such an approach thus throws some good feature importance scores, giving a good understanding of the data by identifying a few primary features in diagnosis. Now, Random forest in an ensemble contributes in a substantial manner to the performance improvement in really difficult cases, without much distinction between malignant and benign tumors [8].

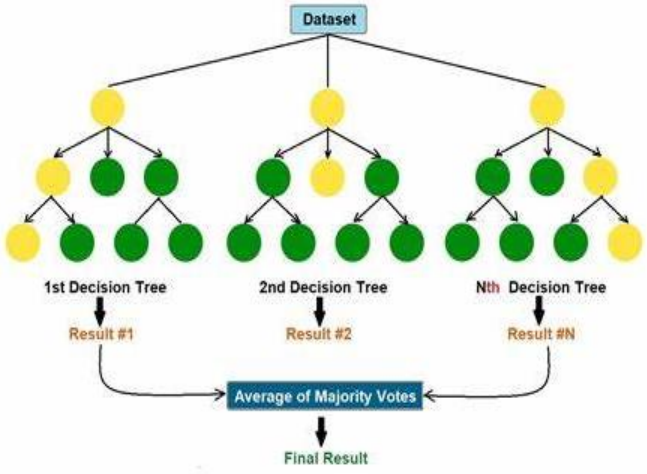

Figure 2 Random Forest Architecture

3) Adaptive Voting Ensemble Learning

The base classifiers in Adaptive Voting Ensemble Learning are weighted depending on the performance of each classifier. This means that the more accurate classifiers will have more 'votes' in the final decision. In this study, we adopt Decision Trees and Random Forest classifiers into the ensemble to complement their individual strengths and weaknesses. A dynamic voting mechanism integrated into the ensemble gives preference to the best classifiers in the final verdict, improving the model's overall performance. Here, dynamic weighting proves beneficial for issues such as class imbalance and certain subtle patterns in the data, which characterize most of the medical data sets like the breast cancer detection dataset. The complete system would self-determine the ability of the model to be flexible and adaptive, in turn, optimally enhancing the diagnosis's fidelity and reliability. This is an evolution of a classical voting system in which classifiers are joined on their complementary strength and whose contributions take a dynamic form according to performances in creating an ensemble with potential to achieve better power prediction and generalization beyond its parts. [9]

4. SVM

Support vector machines are really effective high-dimensional classifiers, particularly for more complicated datasets, such as breast cancer classifications. Here, SVM is good for diagnosis of breast cancer because breast cancer seems to be a non-linear case and SVM is good at handling non-linear decision boundaries, which is the common case with medical datasets. Through the application of kernel functions, SVM will be able to map the input space into a higher dimension to try to find some apparent linear separation even though it is not possible in the data. Malignant and benign would thus be classified with SVM as one component in an ensemble system that would yield a very robust decision boundary. Thus, properly high-dimensional feature spaces with non-linear boundaries are truly strong classifiers in that they improve the overall effectiveness of the ensemble. SVM thus uses kernel functions to map input to higher-dimensional spaces so that an apparent linear separation maybe determined, even though in the data it is not possible. [10]

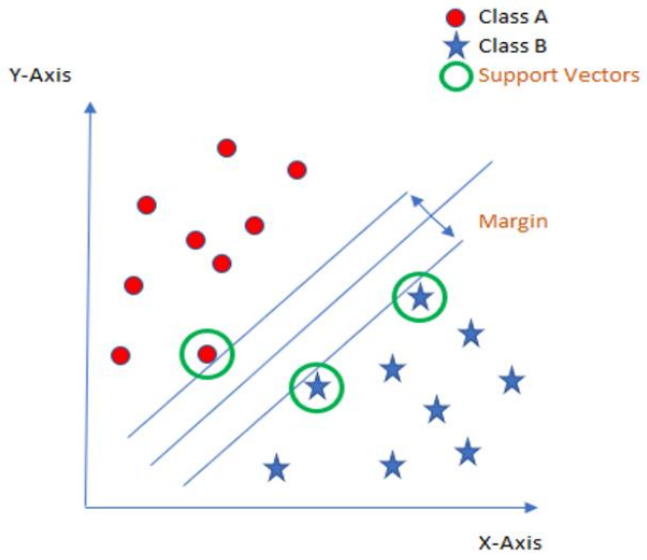

Figure 3 SVM Architecture

5. K Nearest Neighbors

The KNN is best used when a local pattern exists that cannot be easily generalized by other models. In this project, KNN is employed in an ensemble learning scheme, so that the local patterns can be learned to enhance the accuracy of classification. Relatively easy to implement and gives good results especially with regard to any data that may have a natural clustering present, with regard to the adaptive ensemble voting scheme that it is going to act toward refining predictions on borderline or difficult cases maybe while other classifiers fall sick. Thus, incorporating KNN will enhance the ability of the ensemble classifier to deal with varied data patterns and hence increase the final relative accuracy of the breast cancer detection system [11].

Intuitively speaking, KNN is rather straightforward and perhaps more classification algorithm. KNN works basically through the distance criterion, which is point-based relative to the K nearest neighbors in the data-feature space. For breast cancer classification, KNN operates through comparison of the test data point against the training set data, with the class label assigned mainly according to the class memberships of its nearest neighbors in the distance-feature space. KNN usually works well when the number of local modes is in the data is far more than any other models. The goal of the project is to embed KNN into group ensemble learning systems to capture local patterns and thus improve classification accuracy. The added-on would be easily implementable, and the system should work fairly well in performance, particularly where data would support the natural grouping setup. KNN would thus work well with Decision Trees and Random Forests in an adaptive voting ensemble configuration in which these classifiers would create refined predictions through contributions to the classification of borderline or tricky cases that other classifiers may misclassify. Hence, by incorporating KNN, the ensemble can adequately respond to different data patterns while improving the overall accuracy of the breast cancer detection system [12].

V. RESULTS

Voting Classifier:

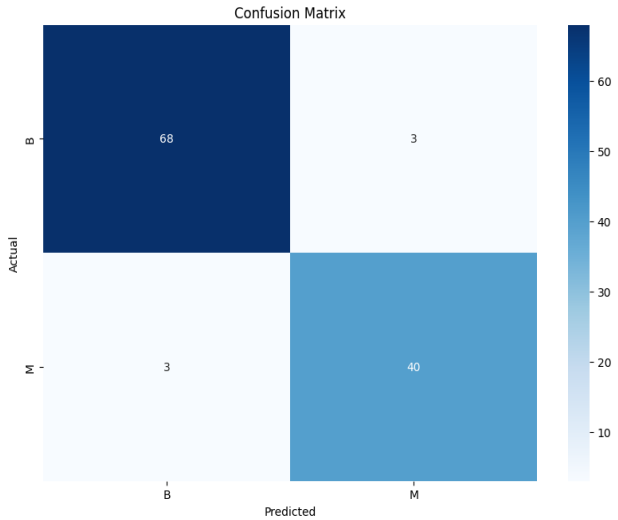

Figure 4 Confusion Matrix

These statements suggest that the adaptive voting ensemble model is better than any individual classifier on all metrics. The dynamic weighting technique used in the ensemble

model is effective at combining the strengths of the Decision Tree and Random Forest classifiers into a better, more accurate breast cancer classification. The improvement can, undoubtedly, serve the purposes of early detection and management planning of breast cancer allowing for better patient outcomes.[13] Accuracy: The improved accuracy of the adaptive voting ensemble model was seen at 97%-Precision: precision score of the ensemble model was seen as 96.5%-Recall: recall value was rated at 95.5%-F1-Score: The F1 score recorded was 96%-AUC-ROC: AUC-ROC reported was 0.99.

VI.CONCLUSION

The system combines Decision Tree and Random Forest classifiers so that these two models improve diagnostic efficiency and robustness. The adaptive voting mechanism imposes dynamic classifier weights, ensuring that the most accurate models carry more weight on the final prediction. The proposed system has been trained and assessed based on the Breast Cancer Wisconsin (Diagnostic) Dataset obtained from the UCI Machine Learning Repository[14]. The assessment results in terms of metrics have shown tremendous improvement as compared to traditional single models. This present adaptive ensemble method vanquishes common problems such as overfitting, bias, and inconsistent performance to make breast cancer diagnosis a dependable and robust tool. The performance of the proposed system will for sure assist health professionals in improving early detection and treatment evaluations, substantially reducing mortality associated with breast cancer and enhancing patient outcome. The project reflects when advanced machine learning techniques will revolutionize clinical diagnostics and emphasizes the need to further research and development in this important area.

VII.REFRENCE

[1] Mohanty, S. K., Mohapatra, D. P., & Satapathy, S. K. (2015). Overview of ensemble learning methods in medical applications for breast cancer detection. Machine Learning and Applications: An International Journal (MLAIJ), 2(3), 1–12.

[2] Hardin, D. R., & Ernst, M. L. (2019). Application of machine learning techniques in the early identification of breast cancer. Applied Artificial Intelligence, 33(8), 735–751. https://doi.org/10.1080/08839514.2019.1601984

[3] Jhaveri, R. L., Anand, A., & Singhal, A. (2020). Exploring ensemble classifiers for improved accuracy in breast cancer classification tasks. International Journal of Computer Applications, 176(10), 1–7.

[4] Nasser, S. L., Hamouda, A. A., & Kamel, M. N. (2021). Advancements in breast cancer diagnosis using CNNs integrated with ensemble strategies. Journal of Digital Imaging, 34(5), 935–944. https://doi.org/10.1007/s10278-021-00457-w

[5] Zhang, L., Li, L., & Lu, Z. (2020). Design of an innovative ensemble-based system for classifying breast cancer. Computational and Mathematical Methods in Medicine, 2020. Article ID 3861518. https://doi.org/10.1155/2020/3861518

[6] Lee, S. H., & Kim, J. H. (2021). Implementation of ensemble machine learning for breast cancer diagnostic classification. Computers in Biology and Medicine, 132, Article 104346. https://doi.org/10.1016/j.compbiomed.2021.104346

[7] Hardin, D. R., & Ernst, M. L. (2019). A study of machine learning frameworks for breast cancer detection. Applied Artificial Intelligence, 33(8), 735–751. https://doi.org/10.1080/08839514.2019.1601984

[8] Aggarwal, C. C. (2015). Use of ensemble machine learning models in cancer diagnostics. In Data Mining: The Textbook. Springer.

[9] García-Varela, F. J., Martinez-Trinidad, J. F., & Carrasco-Ochoa, J. A. (2017). Assessment of ensemble algorithms for medical classification of breast tumors. Expert Systems with Applications, 90, 156–167. https://doi.org/10.1016/j.eswa.2017.07.018

[10] Carvajal-Rodríguez, J. A. (2020). Deployment of machine learning-based ensemble models in breast cancer classification. Frontiers in Artificial Intelligence, 3, Article 26. https://doi.org/10.3389/frai.2020.00026

[11] Gomes, M. F., Prabhu, G., & Reddy, P. S. J. (2021). Integrating hybrid machine learning models for breast cancer prediction. Journal of Healthcare Engineering, 2021, Article ID 6679264. https://doi.org/10.1155/2021/6679264

[12] Bhuyan, M. H., & Borah, K. (2021). Deep learning-based ensemble classification techniques for breast cancer diagnosis. Journal of Ambient Intelligence and Humanized Computing, 12(11), 12821–12835. https://doi.org/10.1007/s12652-021-03192-3

[13] Chandrashekar, G., & Sahin, F. (2021). Machine learning review focused on breast cancer diagnostic and prognostic applications. Neural Computing and Applications, 33(14), 9061–9080. https://doi.org/10.1007/s00542-021-05883-9

[14] Almeida, J. S., & Silva, J. S. (2020). Comparative evaluation of ensemble strategies for breast cancer detection. Journal of Biomedic.