Journal of Science and Technology
ISSN: 2362-1443 (Print) 3059-9490 (Online)
2024, Vol. 4, No. 2 (Special Issue), 1-6
DOI : https://doi.org/10.3126/jost.v4i2.78683

Performance Enhancement of Breast Cancer Detection using AdaBoost Ensemble based on SVM and Decision Tree

Simran Gurung
Nepal College of Information Technology
Pokhara University, Nepal
simran.me.gurung@gmail.com

Ashim Khadka*
Nepal College of Information Technology
Pokhara University, Nepal
ashim.khadka@ncit.edu.np

* Corresponding author

Article History:
Received: 1 July 2024
Revised: 8 August 2024
Accepted: 17 December 2024

Keywords— Ensemble Learning model, Support Vector Machine (SVM), Decision Tree, Breast Cancer, Machine learning .

Abstract—**Breast cancer is the largest cause of death in women. Breast cancer is amenable to treatment, and a favourable prognosis can be attained if the condition is detected and addressed during its initial phases. Early and accurate prediction is the foundation for effective breast cancer management and improved survival rates. Accurate diagnosis plays a crucial role in effective treatment and patient outcomes. Correct classification of malignant and benign tumours can ensure better clinical decision-making, ultimately contributing to improved patient outcomes. This research proposes a novel AdaBoost ensemble method consisting of heterogeneous support vector machine and decision tree as machine learning algorithms in base models. The proposed model can consistently deliver a high precision of 0.98, recall of 0.95, and F_2 score of 0.96 scores, which indicates that the model is highly effective at correctly identifying malignant tumours while minimizing false positives. It not only surpasses traditional models but also outperforms other ensemble techniques, making it a reliable and effective tool for medical diagnostics.**

I. INTRODUCTION

Breast cancer is a disease in which cells in the breast develop abnormally and without control, leading to the formation of a mass known as a tumour. As shown in fig. 1 in 2022, the estimated number of female deaths due to different types of cancer across continents is 4.3 million. Breast cancer is the leading cause of cancer-related deaths among females [1]. Breast cancer, characterized by the growth of malignant tumorous cells in mammary glands, is among the most commonly occurring types of cancer among the female population. Benign tumours are usually well-defined and round or oval. Malignant tumours are generally poorly defined and irregular with lobules as shown in fig 2 [2]. A benign tumour poses minimal risk to the human anatomy and has a low probability of resulting inmortalities among individuals. A malignant tumour poses a greater hazard and has the potential to result in mortal outcomes in individuals. This type of tumour exhibits an expedited proliferation rate due to the anomalous multiplication of cells. The treatment of breast cancer has been

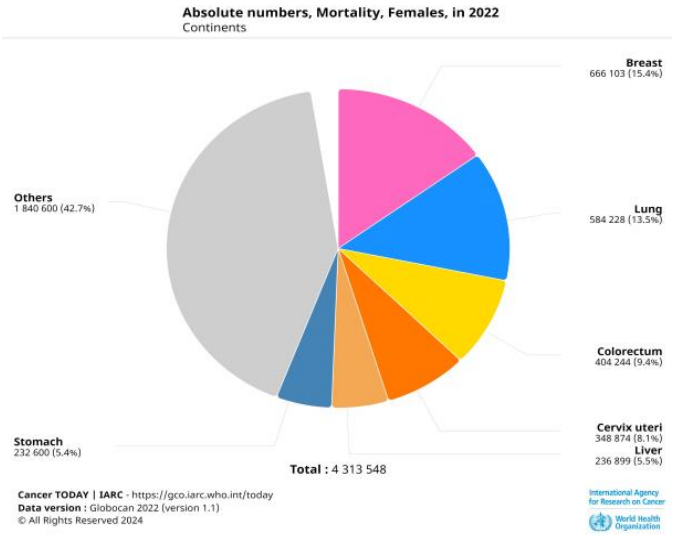

Fig. 1. Female mortality due to different types of cancer in 2022 across continents. [1]

observed to exhibit pronounced efficacy, particularly when accomplished at an early stage. Early and accurate prediction is the foundation for effective breast cancer management and improved survival rates. In this context, the development of robust prediction models is of paramount importance.

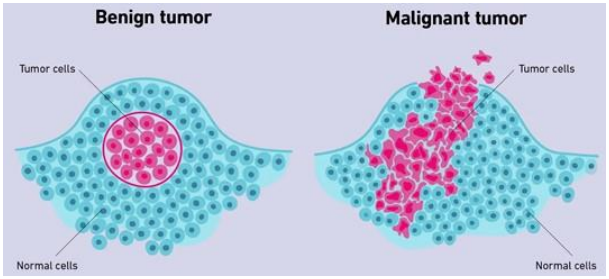

Fig. 2. Benign and malignant tumours cell structure [2]

The medical domain has recently witnessed the utilization of data mining methodologies for the effective analysis of medical data. Data mining methods should be employed for analysis and prediction, emphasising resolving classification challenges encountered in multiple applications [3]– [5]. It is plausible that machine learning systems could outperform standard models in terms of illness classification and prediction [6]–[9]. Machine learning plays a crucial role in the early diagnosis of Breast Cancer. Extensive research has been conducted on the diverse capabilities of Support Vector Machine (SVM), Artificial Neural Network (ANN), and Decision Tree (DT) based methodologies with the classification of medical data [10]–[16]. Many machine learning algorithms and neural network approaches have been used in the Breast Cancer Wisconsin diagnostic and prognostic dataset [17]. The Wisconsin diagnostic dataset consists of 63% benign tumours and 37% malignant tumours, i.e., a highly imbalanced dataset, which may lead to biased models that favour the majority class. Ensemble methods can effectively handle class imbalances by combining classifiers trained on balanced subsets of the data, thereby improving the overall predictive power for the minority class. Ensemble techniques incorporate techniques like resampling (e.g., boosting or bagging), which can mitigate the bias problem by giving more weight to malignant (minority) class instances or adjusting decision boundaries accordingly. Homogeneous ensembles typically use multiple instances of the same base classifier with variations in training data or parameters, which share similar biases and variances and are prone to overfitting, which may not offer sufficient diversity in model predictions. However, the heterogeneous ensembles combine classifiers that utilize different algorithms, which have different biases and variances and can help in reducing overfitting. In a heterogeneous ensemble, each model may handle variations and noise in the data differently. If one of the base models performs poorly on certain subsets of the data, maybe other types of base models in the ensemble might compensate for this, leading to a more balanced and robust performance. This is also crucial in cancer detection scenarios where missing malignant tumours (false negatives) can be much more costly than falsely accusing legitimate benign tumours as malignant (false positives) in the cancer detection system. The heterogeneous Adaboost ensemble method is proposed to detect breast cancer using SVM and DT, where each model weakness is overcome by other models of ensemble methods. The combination can generalize better to unseen data by leveraging the complementary strengths of the individual models. In ensemble methods, SVMs provide robustness in high-dimensional spaces, and DTs offer simplicity and interpretability.

The major contributions of this paper can be summarised as follows:

- This research proposes a novel AdaBoost ensemble method consisting of heterogeneous machine learning algorithms as base models.
- The performance of the ensemble classifier enhanced the prediction of breast cancer and is evaluated on the publicly available Wisconsin diagnostic and prognostic dataset.
- The proposed ensemble method outperforms various state-of-the-art methods for the detection of breast cancer.

The rest of the paper is organized as follows: Section II discusses the question of taxonomy and related works. Section III includes the details of the entire process and the model used. Section IV discusses the detailed results. Finally, a conclusion is in section V.

II. RELATED WORK

A study employed Decision Tree, Logistic Regression, and Naïve Bayes algorithms to determine the optimal feature subset that can achieve precise classification of the identified types of breast cancer. The review paper analysed comparatively the current advancement in the classification of medical data [18]. The two optimum automated breast cancer classification approaches are proposed based on the SVM radial basis function kernel with a hybridization of the Whale Optimization Algorithm (WOA) and Dragonfly Algorithm (DA). The proposed WOA-SVM method outperforms previous classification approaches on the Wisconsin Breast Cancer Database (WBCD) dataset. However, the proposed DA-SVM algorithm outperforms the previous classification algorithms on the Wisconsin Diagnosis Breast Cancer (WDBC) databases. Both the proposed methods use a 10-fold cross-validation dataset with the mean resultant accuracy of 97.89% and 99.27% respectively [10]. The SVM is hybridized with an extremely randomized trees classifier (extra trees) to classify breast cancer by removing unwanted irrelevant features. The proposed method achieved the highest accuracy of 80.23% and outperforms the current state-of-the-art. The proposed method also improved the prediction accuracy by 7.29% in contrast to without the feature selection method [11].

The particle swarm optimization (PSO) is used to optimize the efficiency of the Decision Tree algorithm on the WBCD where the skewed outcomes of classification are solved by a feature selection process. The performance of the proposed system is 92.26% accuracy and outperforms the state-of-the-art. The proposed method is used to create an early detection system for cancer diagnosis and helps health practitioners in decision-making [12]. The adaptive boosting is employed in the decision tree to handle the imbalance dataset of the WBCD. The AdaBoost enhances the performance of breast cancer detection by 4% than the decision tree [13].

The majority voting ensemble method is proposed to enhance the recall of breast cancer classification mechanism and outperforms the current state-of-the-art. The three classifiers simple logistic regression learning, support vector machine learning with stochastic gradient descent optimization and multilayer perceptron network, are used for ensemble classification using a hard voting mechanism and show better performance with 99.42%, as compared to the state-of-the-art algorithm for WBCD [19]. Early detection of breast cancer is important to reduce the mortality of patients. The optimized stacked ensemble learning (OSEL) is proposed where various machine learning methods (k−NN, LR, SVM, DT, RF) are stacked and genetic algorithm is used as optimizers. The proposed OSEL model achieved a maximum accuracy of up to 99.45% as compared to current state-of-theart boosting classifiers [20]. The heterogeneous ensemble learning techniques are used to improve the performance of the Bayesian network and Radial Basis Function classifiers on WBCD. The proposed method achieved optimal and remarkable accuracy of 97% to classify breast cancer and outperformed the existing homogenous ensemble methods and single classifiers [21].

III. METHODOLOGY

Fig. 3 shows the proposed block diagram of the AdaBoost Ensemble method where SVM and DT are used as classifiers to detect breast cancer on the WBCD dataset. Using heterogeneous base models i.e., SVM and DT in AdaBoost for breast cancer detection can leverage the strengths of each model, reduce overfitting, and improve robustness with generalization.

A. Data Collection

The research will utilize the Breast Cancer Wisconsin (WBCD) dataset, which contains a comprehensive set of features and corresponding breast cancer diagnoses. The datasets have been sourced from the machine learning repository at UCI. The aforementioned database comprises a total of 569 records, wherein each record encompasses 30 visual attributes out of 33 attributes pertaining to the characteristics of cancer cells [17]. As shown in fig 4, the

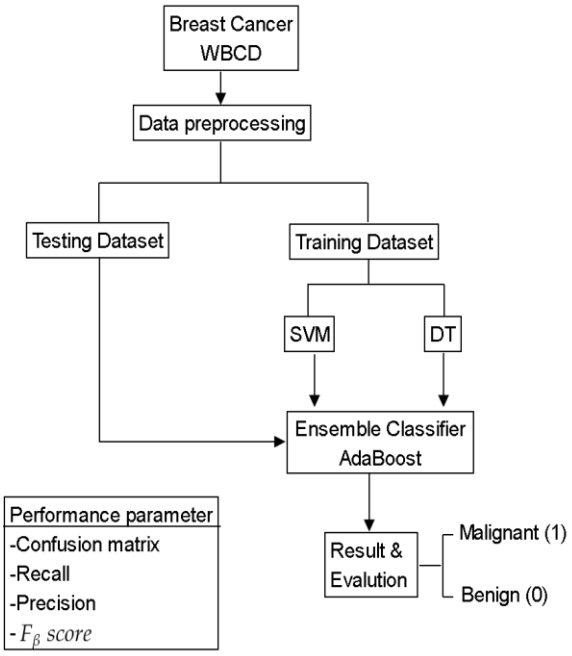

Fig. 3. Block diagram of proposed AdaBoost ensemble method based on SVM and DT to detect breast cancer

WBCD dataset is imbalanced i.e. one class has a larger number of instances of benign with 63% compared to another class of malignant with 37%. In the case of the imbalanced dataset, metrics such as accuracy tend to be misleading and to reduce the false negative recall is used as the main performance parameter.

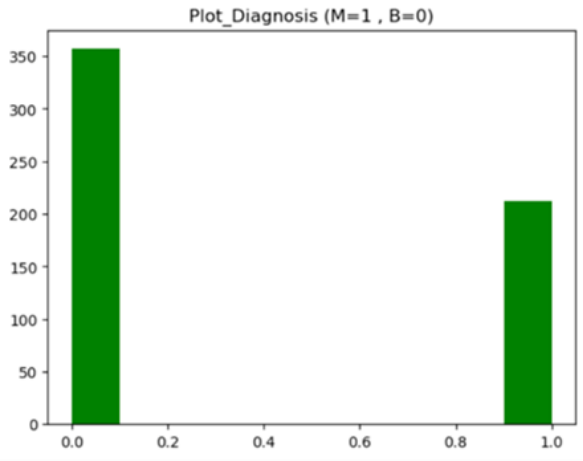

Fig. 4. Illustration of data distribution between malignant (M =1) and benign (B=0) classes of WBCD dataset

B. Preprocessing

In data preprocessing, the data is normalized using zscore normalization entails transforming information into a uniform and consistent structure that can be interpreted and comprehended by computer systems. It ensures the stability of algorithmic and avoids scale sensitivity of input attributes. Data Standardization helps the algorithm to improve optimization convergence, search efficiency, and overall improve its performance. It is a common preprocessing technique used to scale numerical features such that the mean is 0 and the standard deviation is 1.

$$z_i = \frac{x_i - \mu_i}{\sigma_i} \quad (1)$$

where the z_i is the normalized feature value of the $i-th$ of original feature x_i. The μ_i and σ_i are the mean and standard deviation of the $i-th$ feature values. The dataset is divided into two subsets i.e., training and test dataset. By dividing data, it is possible to assess the performance of a model on new data that it has not encountered before. This allows for the avoidance of overfitting, in which the model can remember the training data but does not generalize. It is important to allocate a significant amount of Data for Training and to ensure that there is sufficient Data for Testing. 80:20 ratio was used for splitting the dataset in this research.

C. Classification

The heterogeneous AdaBoost ensemble method is used to classify breast cancer into binary classes where SVM and DT are used as base models. The SVM provides strong performance in high-dimensional spaces and can effectively handle the boundary between classes where as the DT offers interpretability and handle categorical features. Each model has its strengths and weaknesses. AdaBoost focuses on the instances that previous models classified incorrectly. Errors made by heterogeneous models might not overlap significantly, which allows the ensemble to correct mistakes more effectively.

The diverse nature of heterogeneous models can enhance the generalization ability of the ensemble, making it more robust to new, unseen data.

D. Performance Evaluation

In the context of our ensemble learning prediction model for enhancing performance on the Breast Cancer dataset, we employed a range of performance evaluation metrics to comprehensively assess the model's capabilities. The model's performance was evaluated using the following metrics:

- Confusion matrix - It is especially useful when dealing with analyzing the performance in the sense of confusion in classifying two classes.
- Precision - It directly addresses the issue of false positives.
- Recall - If a model fails to identify a true positive case, can have severe consequences for patient safety. • F_β - It is particularly useful in situations where the class distribution is imbalanced, and there may be significant consequences for misclassifying certain classes. It is necessary to minimize the false negative, i.e., misclassification malignant as benign. Thus, the β =2 is used in the F_2 score for high recall.
- AUC-PR - It is a metric used to evaluate the performance of classification models, particularly when dealing with imbalanced datasets. It provides a way to measure the trade-off between precision and recall across different probability thresholds.

IV. RESULT AND DISCUSSION

In this section, the datasets of WBCD used in this research and the experimental evaluations show the usefulness of the proposed model. The cancer tumour is classified into two classes, i.e., malignant (M = 1) and benign (B = 0). In this research, the AUC-PR, recall and F_2 score parameters are used to evaluate the performance of the proposed model and compare it with the current state-of-the-art. The F_2 score places more emphasis on recall than precision. The performance of the proposed method is evaluated in terms of the confusion matrix as provided in fig. 5 where 114 test data are used. The proposed AdaBoost methods provides a very less number of false positive and false negative values indicating improved and stable performance with minimal misclassifications.

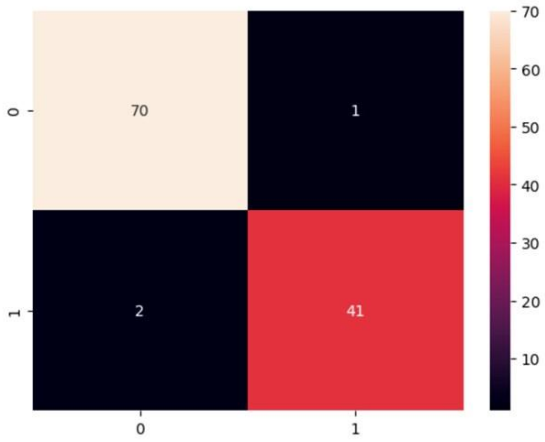

Fig. 5. Confusion matrix of binary classification using proposed methods where malignant (M =1) and benign (B=0)

The table I summarizes the performance metrics for the classification of tumours into malignant and benign classes using the proposed methods. It is observed that the proposed model can correctly identify 95% of all the malignant tumours.

TABLE I
Classification of tumour on malignant and benign classes using proposed methods

Diagnosis	Precision	Recall	F_2
Malignant (1)	0.98	0.95	0.96
Benign (0)	0.97	0.99	0.99

The model has a high precision of approximately 98%, indicating that when it predicts a malignant case. The F_2 score of approximately 96% demonstrates a good trade-off between precision and recall indicating that the proposed model can effectively distinguish tumours between malignant and benign cases with minimal misclassification. The proposed method also demonstrates outstanding performance in classifying benign tumours, with a precision of 97% and a recall of 99% which means very few false positives and almost no false negatives.

Table II compares the performance of the proposed methods with various state-of-the-art models in terms of precision, recall, F2 score, and AUC-PR (Area Under the Precision-Recall Curve). The overall performance of SVM and DT is increased by applying the AdaBoost ensemble learning algorithm on weak learners. The proposed method yields higher classification results in comparison to state-of-the-art methods.

TABLE II

Compare the proposed methods with state-of-the-art

Models	Precision	Recall	F_2	AUC-PR
SVM	0.95	0.95	0.95	0.96
ANN	0.95	0.91	0.92	0.95
DT	0.80	0.91	0.85	0.85
Soft Voting (ANN and SVM)	0.95	0.93	0.93	0.95
XGBoost (ANN, SVM and DT)	0.93	0.95	0.95	0.95
Adaboost (SVM and DT)	0.98	0.95	0.96	0.97

The decision tree model shows the lowest performance metrics among all, with significantly lower precision and AUC-PR. Similarly, the artificial neural networks model also shows the lowest recall value. The XGBoost which is a combination of ANN, SVM and DT, has a balanced performance with good precision and recall but falls slightly short of the proposed Adaboost method in precision and F score. The AUC-PR of 0.97 for the Adaboost method is the highest among all compared models, indicating that the proposed model always correctly predicts both classes even though the used WBCD dataset is highly imbalanced.

V. CONCLUSION AND FUTURE WORK

This paper aims to classify the types of tumour i.e. Benign tumour and Malignant tumour, using AdaBoost ensemble methods which is capable enough to classify the tumour with high recall and reduce false negatives. The heterogeneous base model, i.e., SVM and DT are used in AdaBoost methods by overcoming the weakness of one model by another model of ensemble methods. The classification results in this research affirmed the potency of ensemble learning for intricate breast cancer prediction. The proposed model can consistently deliver a high precision of 0.98, recall of 0.95, and F_2 score of 0.96 scores, which indicates that the model is highly effective at correctly identifying malignant tumours while minimizing false positives. It not only surpasses traditional models but also outperforms other ensemble techniques, making it a reliable and effective tool for medical diagnostics which ensures better clinical decision-making, ultimately contributing to improved patient outcomes. But when dealing with a very large dataset, training multiple SVMs as base classifiers in AdaBoost can significantly increase computational cost and complexity. So, for further work and research, a study on reduction on computational cost and complexity can be performed.

REFERENCES

[1] I. A. for Research on Cancer (IARC), "Globocan 2022," https://gco.iarc.fr/today/en/dataviz/pie?mode=cancer&group_populations=1&sexes=2&types=1.

[2] S. Whelan, "Benign vs malignant tumors," https://www.technologynetworks.com/cancer-research/articles/benign-vs-malignant-tumors-364765, 8 2022.

[3] A. T. Alhasani, H. Alkattan, A. A. Subhi, E.-S. M. El-Kenawy, and M. M. Eid, "A comparative analysis of methods for detecting and diagnosing breast cancer based on data mining," Methods, vol. 7, no. 9, 2023.

[4] S. F. Khorshid, A. M. Abdulazeez, and A. B. Sallow, "A comparative analysis and predicting for breast cancer detection based on data mining models," Asian Journal of Research in Computer Science, vol. 8, no. 4, pp. 45–59, 2021.

[5] V. Chaurasia, S. Pal, and B. Tiwari, "Prediction of benign and malignant breast cancer using data mining techniques," Journal of Algorithms & Computational Technology, vol. 12, no. 2, pp. 119–126, 2018.

[6] K. Bond and A. Sheta, "Medical data classification using machine learning techniques," International Journal of Computer Applications, vol. 183, no. 6, p. 1–8, Jun. 2021.

[7] S. Uddin, A. Khan, M. E. Hossain, and M. A. Moni, "Comparing different supervised machine learning algorithms for disease prediction," BMC Medical Informatics and Decision Making, vol. 19, no. 1, 12 2019.

[8] N. Sindhwani, A. Rana, A. Chaudhary et al., "Breast cancer detection using machine learning algorithms," in 2021 9th International conference on reliability, Infocom technologies and optimization (trends and future directions)(ICRITO). IEEE, 2021, pp. 1–5.

[9] S. A. Mohammed, S. Darrab, S. A. Noaman, and G. Saake, "Analysis of breast cancer detection using different machine learning techniques," in Data Mining and Big Data: 5th International Conference, DMBD 2020, Belgrade, Serbia, July 14–20, 2020, Proceedings 5. Springer, 2020, pp. 108–117.

[10] A. S. Elkorany, M. Marey, K. M. Almustafa, and Z. F. Elsharkawy, "Breast cancer diagnosis using support vector machines optimized by whale optimization and dragonfly algorithms," IEEE Access, vol. 10, pp. 69688–69699, 2022.

[11] G. Alfian, M. Syafrudin, I. Fahrurrozi, N. L. Fitriyani, F. T. D. Atmaji, T. Widodo, N. Bahiyah, F. Benes, and J. Rhee, "Predicting breast cancer from risk factors using svm and extratrees-based feature selection method," Computers, vol. 11, no. 9, p. 136, 2022.

[12] J. O. Afolayan, M. O. Adebiyi, M. O. Arowolo, C. Chakraborty, and A. A. Adebiyi, "Breast cancer detection using particle swarm optimization and decision tree machine learning technique," in Intelligent Healthcare: Infrastructure, Algorithms and Management. Springer, 2022, pp. 61–83.

[13] T. A. Assegie, R. L. Tulasi, and N. K. Kumar, "Breast cancer prediction model with decision tree and adaptive boosting," IAES International Journal of Artificial Intelligence, vol. 10, no. 1, pp. 184–190, 2021.

[14] M. H. Alshayeji, H. Ellethy, R. Gupta et al., "Computer-aided detection of breast cancer on the wisconsin dataset: An artificial neural networks approach," Biomedical signal processing and control, vol. 71, p. 103141, 2022.

[15] A. B. Nassif, M. A. Talib, Q. Nasir, Y. Afadar, and O. Elgendy, "Breast cancer detection using artificial intelligence techniques: A systematic literature review," Artificial intelligence in medicine, vol. 127, p. 102276, 2022.
[16] P. Paudel, R. Saud, S. K. Karna, and M. Bhandari, "Determining the major contributing features to predict breast cancer imposing ml algorithms with lime and shap," in 2023 International Conference on Electrical, Computer and Energy Technologies (ICECET), 2023.
[17] W. Wolberg, O. Mangasarian, N. Street, and W. Street, "Breast Cancer Wisconsin (Diagnostic)," UCI Machine Learning Repository, 1995, DOI: https://doi.org/10.24432/C5DW2B.
[18] S. A. Lashari, R. Ibrahim, N. Senan, and N. S. A. Taujuddin, "Application of data mining techniques for medical data classification: a review," in MATEC Web of conferences, vol. 150. EDP Sciences, 2018, p. 06003.
[19] A. S. Assiri, S. Nazir, and S. A. Velastin, "Breast tumor classification using an ensemble machine learning method," Journal of Imaging, vol. 6, no. 6, p. 39, 2020.
[20] M. Kumar, S. Singhal, S. Shekhar, B. Sharma, and G. Srivastava, "Optimized stacking ensemble learning model for breast cancer detection and classification using machine learning," Sustainability, vol. 14, no. 21, p. 13998, 2022.
[21] M. A. Jabbar, "Breast cancer data classification using ensemble machine learning." Engineering & Applied Science Research, vol. 48, no. 1, 2021.