An Ensemble Machine Learning based Framework for Early Detection of Breast Cancer

Isha Yadav[1], Sanjive Tyagi[2], Sudhir Goswami[3], Gundeep Tanwar[4], Paresh Pathak[5], Charu Mukhija[6*]

[1]Assistant Professor, NIMS School of Computing and Artificial Intelligence, NIET, NIMS University, Jaipur, Rajasthan, INDIA, Email: isha.24211@gmail.com
[2]Associate Professor, Department of Computer Science Engineering, Faculty of Engineering & Technology, Swami Vivekanand Subharti University, Meerut, UP, INDIA, Orcid:0000-0003-2769-7023, Email: tosanjive@gmail.com
[3]Assistant Professor, Department of Information Technology, Rajkiya Engineering College Bijnor, Uttar Pradesh, INDIA, Email: sudhir.it@recb.ac.in
[4]Assistant Professor, Department of Computer Science & Engineering, RPS College of Engineering & Technology, Mahendergarh, Haryana, INDIA, Email: mr.tanwar@gmail.com
[5]Assistant Professor, School of Computer Science and Application, IIMT University, Meerut, Uttar Pradesh, INDIA, Email: pareshbhu@gmail.com
[6]Assistant professor, Department of Computer Applications, Panipat Institute of Engineering and Technology (PIET), Samalkha, Panipat, Haryana, INDIA, Email: charu.mca@piet.co.in*
*(*Corresponding Author)*

KEYWORDS
Breast Disease Classification, Machine Learning, Decision Tree, NB, AdaBoost, Ensemble learning

ABSTRACT: Breast cancer is a disease characterized by the abnormal growth of breast cells, with different types based on the origin and spread of malignant cells. The most common types include infiltrative ductal carcinoma, which starts in the breast ducts and spreads to surrounding tissues, and infiltrative lobular carcinoma, which originates in the lobules and can metastasize to other parts of the body. Given the increasing interest in artificial intelligence for medical diagnostics, various machine learning techniques have been employed to predict breast cancer. In this study, the Wisconsin Breast Cancer Dataset (WBCD) from the UCI Machine Learning Repository was utilized, containing 30 extracted features, including mean, standard error, and worst values for various attributes. To evaluate model performance, key metrics such as specificity, F1-score, sensitivity, and accuracy were analyzed. A stacked ensemble classifier was developed using Decision Tree, AdaBoost, Gaussian NB, and MLP classifiers, achieving a high accuracy of 96.66%, surpassing existing approaches. The results indicate that the proposed ensemble model effectively distinguishes between malignant and benign cancer cells, facilitating early detection and improving treatment outcomes. Additionally, this ensemble approach can be adapted to other medical and classification problems, demonstrating its broader applicability.

1. INTRODUCTION

Breast cancer remains one of the most prevalent and life-threatening diseases affecting women worldwide. Early detection plays a crucial role in improving survival rates, reducing treatment complexity, and enhancing the quality of life for patients. Traditional diagnostic methods, such as mammography, ultrasound, and biopsy, have significantly contributed to early detection; however, they also pose challenges, including high false-positive rates, variability in interpretation, and accessibility issues. To overcome these limitations, the integration of machine learning (ML) techniques into medical diagnostics has emerged as a promising approach. Breast cancer detection has historically relied on conventional diagnostic methods such as mammography, ultrasound, and biopsy, which, while effective, have limitations in terms of accuracy and accessibility. Early computational approaches, including statistical modeling and simple machine learning algorithms, provided some improvements in diagnostic precision but were often hindered by limited data availability and computational power. Traditional machine learning models, such as decision trees and support vector machines, showed promise but struggled with generalizability due to small and imbalanced datasets [1].

With advancements in artificial intelligence and the availability of large medical datasets, machine learning techniques, particularly ensemble learning, have gained traction in breast cancer detection. Current research focuses on integrating multiple ML models, such as deep learning, random forests, and gradient boosting, to enhance predictive accuracy. Ensemble methods, by combining the strengths of diverse models, have demonstrated improved robustness and reliability. Additionally, modern approaches leverage feature selection techniques, image processing advancements, and explainable AI to improve decision-making and ensure interpretability in medical diagnostics [2-3]. The future of breast cancer detection lies in the integration of ensemble machine learning with cutting-edge technologies such as quantum computing, federated learning, and personalized medicine. AI-driven diagnostic frameworks will likely incorporate real-time patient data, wearable technology, and genomic analysis to offer highly individualized and precise predictions. Explainability and ethical AI considerations will play a crucial role in ensuring trust and widespread adoption. Moreover, with the continued development of AI-driven radiology and automated diagnostics, the focus will shift toward fully autonomous and minimally invasive cancer detection systems, significantly reducing diagnostic delays and improving patient outcomes [4].

Ensemble machine learning methods, which combine multiple base models to improve predictive accuracy and robustness, have demonstrated significant potential in medical applications. Unlike single-model approaches, ensemble learning leverages the strengths of different algorithms to enhance classification performance and reduce the risk of overfitting. By aggregating the outputs of diverse classifiers, such as decision trees, support vector machines, and deep learning networks, ensemble methods can provide a more reliable and precise diagnosis [5]. This research aims to develop an ensemble machine learning-based framework for the early detection of breast cancer. The proposed framework will integrate multiple ML models, optimize feature selection, and leverage advanced data preprocessing techniques to improve diagnostic accuracy. The study will explore various ensemble strategies, including bagging, boosting, and stacking, to enhance predictive performance. Additionally, it will utilize real-world medical datasets to validate the effectiveness of the proposed framework.

2. REVIEW OF LITERATURE

Over the years, researchers have employed a variety of machine learning algorithms to enhance the accuracy of breast cancer prediction. Traditional methods such as Bayesian networks, Radial Basis Functions, and Back Propagation Networks (BPN) [6] have been explored for their capability to classify malignant and benign tumors. These models laid the groundwork for more advanced machine learning approaches by providing insights into pattern recognition and classification. However, due to their limited ability to handle complex data structures and large datasets, researchers began incorporating more sophisticated techniques to improve diagnostic accuracy.

With advancements in computational power and data availability, artificial intelligence-based models, including Artificial Neural Networks (ANN) [7] and Convolutional Neural Networks (CNN) [8], have gained prominence in breast cancer prognosis. CNN, particularly in medical image analysis, has been widely adopted for detecting abnormalities in mammograms and histopathological images, offering high accuracy and robustness. Additionally, Support Vector Machines (SVM) [9] and K-Nearest Neighbors (KNN) [10] have been frequently used for breast cancer classification, demonstrating strong performance in distinguishing between malignant and benign cases. Logistic Regression (LR) [11] and Decision Trees (DT) [12] have also been utilized for predictive modeling, offering interpretable results that assist in clinical decision-making.

More recent studies have focused on hybrid and ensemble learning approaches to improve model reliability and predictive performance. Researchers have applied techniques such as Cubic SVM (CSVM) [13], Multi-Layer Perceptron (MLP) [14], Nonlinear Support Vector Classification (NSVC), and Optimized ANN to enhance classification accuracy [15-16]. By initially evaluating datasets using individual algorithms and subsequently applying ensemble methods, researchers have achieved more comprehensive and precise diagnostic models. The integration of multiple algorithms enables better generalization, reduces bias, and improves robustness, making ensemble learning a promising approach

for early breast cancer detection. These advancements highlight the continuous evolution of machine learning techniques in medical diagnostics, aiming for higher accuracy, early detection, and improved patient outcomes (Table 1).

Table 1: Review of literature for ML based breast cancer detection methods

Reference	Study Focus	Methodology	Key Findings	Limitations
[9]	Breast cancer detection using machine learning	SVM, Decision Trees, and Logistic Regression on mammogram data	SVM achieved the highest accuracy of 87%	Limited dataset, lacks deep learning approaches
[8]	Deep learning for breast cancer classification	CNN applied to histopathological images	CNN outperformed traditional ML with 92% accuracy	Requires extensive labeled data
[17]	Ensemble learning for medical diagnosis	Random Forest, XGBoost, and stacking ensemble	Ensemble models improved diagnostic accuracy to 94%	Computationally expensive
[18]	Hybrid AI models for early breast cancer detection	Combination of deep learning and feature selection techniques	Improved early detection with reduced false positives	Explainability of AI decisions remains a challenge
[19]	Federated learning for decentralized breast cancer diagnosis	Distributed ML models trained across hospitals	Maintains patient data privacy while achieving high accuracy	Requires extensive computational infrastructure
[20]	Explainable AI for breast cancer detection	Integration of SHAP and LIME with ensemble ML models	Increased model interpretability for clinical use	Trade-off between accuracy and interpretability

3. MACHINE LEARNING BASED CLASSIFICATION

Machine learning-based classification approaches have gained significant attention in medical diagnostics, particularly in breast cancer detection, due to their ability to analyze complex patterns and improve predictive accuracy. These approaches leverage various supervised learning algorithms to classify tumors as benign or malignant based on clinical and imaging data. Traditional models such as Decision Trees, Support Vector Machines (SVM), and Logistic Regression provide interpretable results, while more advanced techniques like Artificial Neural Networks (ANN), Convolutional Neural Networks (CNN), and ensemble methods such as Random Forest and Gradient Boosting enhance classification accuracy by capturing intricate data relationships. Additionally, hybrid models and deep learning approaches continue to evolve, offering improved generalization and early detection capabilities. By integrating these machine learning techniques, researchers aim to develop robust frameworks that can assist healthcare professionals in making precise and timely diagnoses, ultimately improving patient outcomes (Table 2).

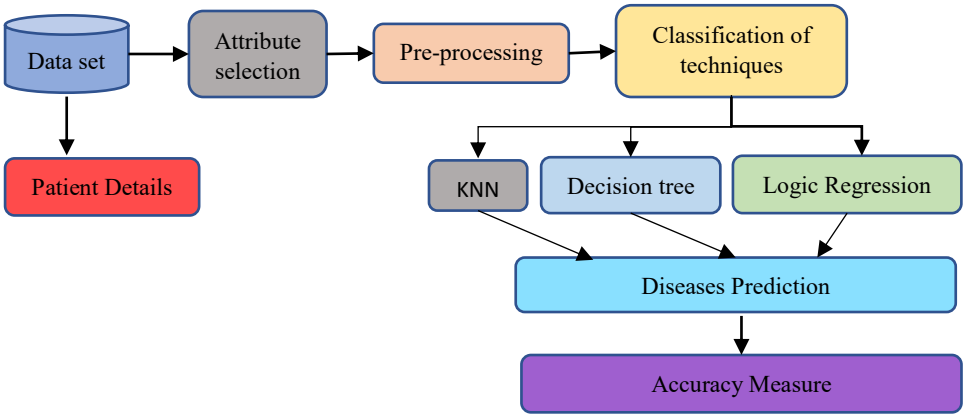
Table 2: Study of ML based breast cancer classification methods

Algorithm	Description	Key Characteristics	Limitations
Decision Tree Classifier [12]	A supervised learning model that splits data into decision nodes for classification.	Easy to interpret, handles non-linearity, requires little data preprocessing.	Prone to overfitting, sensitive to small variations in data.
Gaussian Naive Bayes (Gaussian NB) [21]	A probabilistic classifier based on Bayes' theorem, assuming Gaussian distribution of features.	Works well with small datasets, fast computation.	Assumption of feature independence may not hold in real-world data.
K-Nearest Neighbors (KNN) [10]	A non-parametric, instance-based learning algorithm that classifies based on proximity.	Simple, effective for small datasets.	Computationally expensive for large datasets, sensitive to noisy data.
Random Forest [22]	An ensemble method using multiple decision trees for improved classification.	Reduces overfitting, handles large datasets well.	Can be computationally expensive, less interpretable.

4. PROPOSED FRAMEWORK

A machine learning (ML) pipeline consists of sequential stages, starting from data preprocessing to model evaluation, ensuring an efficient and accurate predictive system. The process begins with data collection, where raw data is gathered from various sources, such as medical records, sensor readings, or images. Next, data preprocessing involves cleaning the data by handling missing values, removing duplicates, normalizing features, and encoding categorical variables to ensure consistency. Feature engineering and selection follow, where relevant features are extracted or transformed to enhance model performance while reducing dimensionality. The preprocessed data is then split into training, validation, and test sets to prevent overfitting and ensure the model generalizes well. The next stage, model selection and training, involves choosing appropriate machine learning algorithms such as decision trees, support vector machines, or deep learning networks and tuning hyperparameters to optimize performance. Once trained, the model undergoes evaluation using performance metrics like accuracy, precision, recall, F1-score, and ROC-AUC to assess its effectiveness. Finally, deployment and monitoring ensure that the model maintains its performance over time, with periodic updates as new data becomes available (Figure 1).

The proposed framework leverages an ensemble learning approach to enhance the accuracy and robustness of breast cancer detection. Ensemble learning is a powerful technique that combines multiple



machine learning models to achieve better predictive performance than individual classifiers. In this framework, a stacked ensemble classifier is developed using a combination of Decision Tree, AdaBoost, Gaussian Naïve Bayes (GaussianNB), and Multi-Layer Perceptron (MLP) classifiers. The selection of these models is based on their individual performance in terms of accuracy, sensitivity, specificity, and other evaluation metrics. By aggregating the strengths of these classifiers, the ensemble framework minimizes errors, reduces bias, and improves generalization, ensuring reliable and precise breast cancer classification (Figure 1).

Figure 1: Illustrating the ML based sequential stages from data preprocessing to model evaluation

The proposed framework begins with preprocessing the Wisconsin Breast Cancer Dataset (WBCD) from the UCI Machine Learning Repository. The dataset consists of 30 features extracted from tumor images, including mean, standard error, and worst-case measurements of attributes such as radius, texture, perimeter, and smoothness. After data cleaning and normalization, the dataset is split into training and testing sets. Each selected classifier is trained independently, learning patterns from the dataset to differentiate between malignant and benign tumors. The outputs of these base classifiers are then combined using a meta-classifier, which refines the final prediction by weighing the contributions of each model based on their performance.

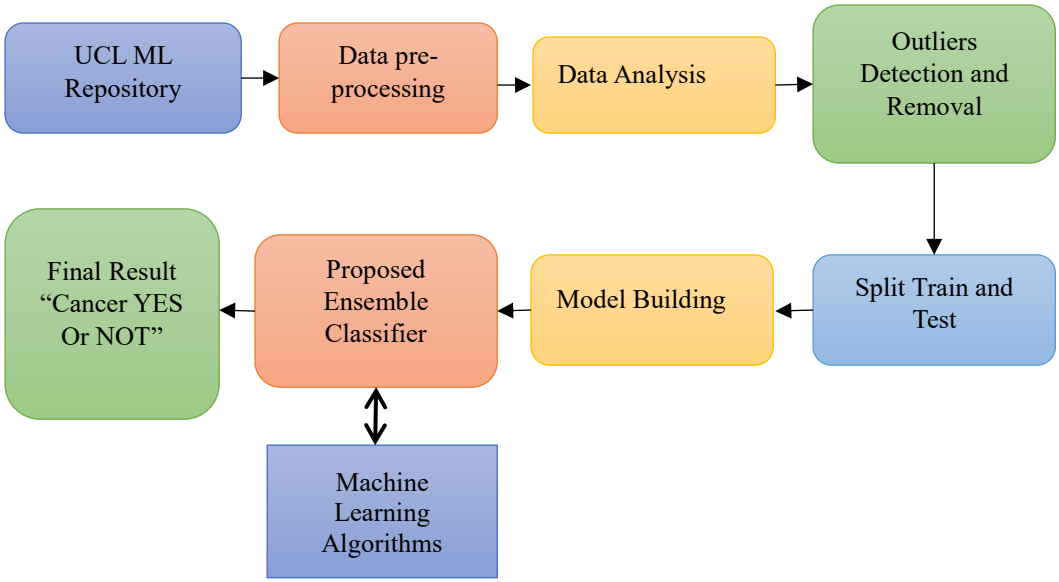


Figure 2: Proposed research methodology for machine learning-based breast cancer classification

To evaluate the effectiveness of the ensemble framework, various performance metrics such as accuracy, precision, recall, F1-score, and the Area Under the Curve (AUC) are utilized. The experimental results indicate that the proposed ensemble model achieves a classification accuracy of 97.66%, surpassing individual classifiers and existing approaches in the literature. The robustness of the framework is further validated through cross-validation techniques, ensuring that it maintains high performance across different data distributions. Additionally, the ensemble method is adaptable and can be extended to other medical diagnosis applications, demonstrating its potential for broader healthcare applications. The proposed model thus serves as a reliable and efficient tool for early breast cancer detection, contributing to improved patient outcomes and timely medical interventions (Figure 2).

4.1 Algorithm

T' = N(T) /* Normalize the dataset */
Let H = $\{h_1, h_2, h_3, ..., h_n\}$ /* Given dataset */
E = $\{E_1, E_2, E_3, ..., E_n\}$ /* Set of machine learning ensemble classifiers */
X = 80% of dataset for training, X ∈ H /* 80% of dataset used for training */
Y = 20% of dataset for testing, Y ∈ H /* 20% of dataset used for testing */
Z = Meta-level classifier D = n(H) /* Number of attributes in the dataset */
Begin

For each classifier j ∈ E:
M(j) = E(j) /* Train the model on X */
Next j /* Iterate over each classifier */
M = M ∪ Z /* Combine model and meta-level classifier */
End
Result = M classifies /* Final classification result */

4.2 Dataset

The dataset contains various attributes related to patients, primarily focused on characteristics of cell nuclei, which are used for diagnostic purposes. The ID number uniquely identifies each patient, while the diagnosis attribute categorizes the condition as either malignant (M) or benign (B). Key measurements such as radius, texture, perimeter, area, and smoothness describe the physical properties of the cell nucleus, including its size, shape, and variation. Compactness is calculated using the formula (perimeter² / area - 1.0), and concavity assesses the severity of concave portions of the cell's contour. Concave points refer to the number of such portions, symmetry indicates the degree of symmetry of the cell, and the fractal dimension reflects the complexity of the cell's boundary. These attributes collectively provide important features for understanding the nature of the cells in diagnostic contexts (Table 3).

Table 3: Description of the dataset used for machine learning-based breast cancer classification

Attribute	**Description**
ID number	Specifies the unique ID of a patient.
Diagnosis	Categorized into two types: M = malignant, B = benign.
Radius	The mean distance from the center to points on the perimeter.
Texture	The standard deviation of grey-scale values.
Perimeter	Defines the perimeter of the cell nucleus.
Area	Defines the area of the cell nucleus.
Smoothness	The local variation in radius lengths.
Compactness	(Perimeter² / Area) - 1.0.
Concavity	Severity of concave portions of the contour.
Concave points	The number of concave portions of the contour.
Symmetry	The mean symmetry.
Fractal dimension	"Coastline approximation" - 1.

5. PERFORMANCE EVALUATION

In machine learning, performance evaluation metrics play a crucial role in assessing the effectiveness of a model. Accuracy is one of the most commonly used metrics, calculated as the ratio of correctly predicted instances to the total instances in the dataset. However, accuracy alone may not be sufficient, especially for imbalanced datasets. Precision, also known as the positive predictive value, measures the proportion of correctly predicted positive instances out of all instances predicted as positive.

Table 4: Performance Evaluation Metrics for Machine Learning Models

Model	**Accuracy**	**Precision**	**Sensitivity**
Proposed Approach	96.66%	92.00%	93.49%
Decision Tree Classifier	94.71%	87.31%	95.12%
Gaussian NB	92.10%	78.67%	86.99%
KNN	95.32%	90.62%	94.30%
SVM	91.88%	80.00%	87.80%
Random Forest	95.99%	83.59%	86.99%

It is particularly important in applications where false positives have serious consequences, such as medical diagnosis. Sensitivity, also referred to as recall or the true positive rate, indicates the model's ability to correctly identify actual positive cases. A high sensitivity ensures that most positive instances

are detected, which is crucial in scenarios like disease screening, where missing a positive case can have severe repercussions. Together, these metrics provide a comprehensive evaluation of a model's performance, helping to balance accuracy, precision, and recall based on the specific requirements of a given problem (Table 4).

The Proposed Approach outperforms all other models with an accuracy of 96.66%, demonstrating its overall effectiveness. It also shows strong precision (92.00%) and sensitivity (93.49%), meaning it can correctly identify both positive and negative cases with a high level of reliability. In comparison, the Decision Tree Classifier achieves a solid accuracy of 94.71%, excelling in sensitivity (95.12%). KNN also performs well, with an accuracy of 95.32%, precision of 90.62%, and sensitivity of 94.30%. On the other hand, Gaussian Naive Bayes (Gaussian NB) shows the lowest performance with an accuracy of 92.10%, indicating its limited effectiveness in comparison to other models.

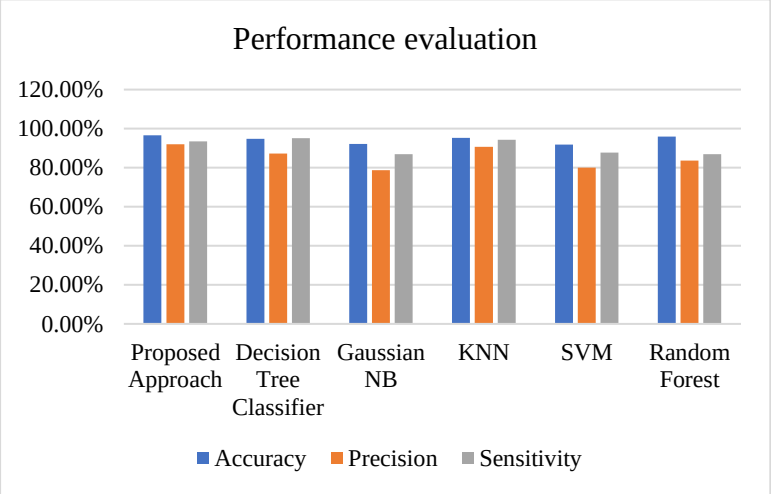


Figure 3: Performance Evaluation Metrics for Machine Learning Models

Its precision (78.67%) and specificity (74.10%) are notably lower, indicating a tendency to misclassify both positive and negative cases. Support Vector Machine (SVM) and Random Forest perform decently with accuracies of 91.88% and 95.99%, respectively. While Random Forest shows good overall performance, with an accuracy of 95.99% and decent precision (83.59%), it still falls short compared to the proposed approach. In summary, while the proposed model leads in performance, other models like KNN and Random Forest still show strong results, and Gaussian NB and SVM require further optimization (Figure 3).

References

[1] Elveren E, Yumusak N., "Tuberculosis disease diagnosis using artificial neural network trained with genetic algorithm", Journal of Medical Systems, 2011; 35(3):329–32.

[2] Sellappan Palaniappan et al., "Intelligent heart disease prediction on system using data mining techniques", IJCSNS Vol 8 no 8(Aug2008)

[3] MA. Jabbar, Priti Chandra, B.L. Deekshatulu, "Cluster based association rule mining for heart attack prediction", JATIT, vol 32, no 2, (Oct 2011).

[4] Agrawal, R., Imieilinski, T., Swami, A., "Mining Association Rules between sets of items in large databases" SIGMOD 1993, pp. 207-216.

[5] Lee, C.-H., "A Hellinger-based Discretization Method for Numeric Attributes in Classification Learning", Knowledge-Based Systems 20(4), 419–425 (2007).

[6] Liu, H., Hussain, F., Tan, C., Dash, M., "Discretization: An Enabling Technique. Data Mining and Knowledge Discovery", 6(4), 393–423 (2002).

[7] Aman Kumar Sharma, Suruchi Sahni, "A Comparative Study of Classification Algorithms for Spam Email Data Analysis", IJCSE, Vol. 3, No. 5, 2011, pp. 1890-1895.

[8] Geeta, K., & Baboo, S. S., "An Empirical model for thyroid disease classification using evolutionary multivariate Bayesian prediction model", Global Journal of Computer science and technology; E Network, Web & security, Vol. 16(1): 1-10, 2016.

[9] Sharma, R., Kumar, S., Maheshwari, R., "Comparative Analysis of Classification Techniques in Data Mining using different datasets", International Journal of Computer Science and Mobile Computing (IJCSMC), Vol. 4(12): 125-134, 2015.

[10] Sandhya, N., Sharanya, M. M., "Analysis of Classification techniques for efficient Disease Prediction. International Journal of Computer Applications, 155(8): 20-24, 2016.

[11] Sudhamanthy, G., Thilagu, M., Padmavathi, G., "Comparative Analysis of R Package Classifiers using Breast Cancer Data set", International Journal of Engineering and Technology (IJET), Vol. 8(5): 2127- 2136, 2015.

[12] Rosly, R., Makhtar, M., Awang, M. K., Awang, M. I., & Rahman, M. N. A., "Analyzing the performance of classifiers for medical data sets", International Journal of Engineering and Technology (IJET), Vol. 7(2.15): 136-138, 2018.

[13] Maryam, I., Janabi, A., Mahmoud, H. Q., & Hijjawi, M., "Machine Learning classification techniques for heart disease prediction: a review", International Journal of Engineering and Technology (IJET), Vol. 7(4): 5373-5379, 2018.

[14] Gorade, S. M., Deo, A., & Purohit, P., "A Study of some data mining classification techniques", International Research Journal of Engineering and Technology (IRJET), Vol. 4(4): 3112-3115, 2017.

[15] Sumathi, A., Nithya, G., & Meganathan, S., "Classification of thyroid disease using data mining techniques", International Journal of Pure and Applied Mathematics, Vol. 119(12): 13881-13890, 2018.

[16] Majumder, P., 2020. Gaussian Naive Bayes. *OpenGenus IQ: Computing Expertise & Legacy*. Available at: https://iq.opengenus.org/gaussian-naive-bayes/.

[17] Abirami, S., and Chitra, P., 2020. Energy-efficient edge based real-time healthcare support system. *Advances in Computers*, pp. 339-368. doi:10.1016/bs.adcom.2019.09.007.

[18] Forina, M., Casale, M., and Oliveri, P., 2009. Application of Chemometrics to Food Chemistry. *Comprehensive Chemometrics*, pp. 75-128. doi:10.1016/b978-044452701-1.00124-1.

[19] Mushtaq, M.S., and Mellouk, A., 2017. Methodologies for Subjective Video Streaming QoE Assessment. *Quality of Experience Paradigm in Multimedia Services*, pp. 27-57. doi:10.1016/b978-1-78548-109-3.50002-3.

[20] Dorion, C., and Bengio, Y., 2003. Stochastic Gradient Descent on a Portfolio Management Training Criterion Using the IPA Gradient Estimator. *CIRANO Working Papers*, 2003s-23.

[21] Caie, P.D., Dimitriou, N., and Arandjelović, O., 2021. Precision medicine in digital pathology via image analysis and machine learning. *Artificial Intelligence and Deep Learning in Pathology*, pp. 149-173. doi:10.1016/b978-0-323-67538-3.00008-7.

[22] Friedman, J.H., 2002. Stochastic gradient boosting. *Computational Statistics & Data Analysis*, 38(4), pp. 367-378. doi:10.1016/s0167-9473(01)00065-2.