

# A Robust Ensemble Learning Framework for Breast Tumor Classification and Early Detection

<sup>1</sup>Dr. Parbhakar Singh, <sup>2</sup>Dr. Ajay Sharma, <sup>3</sup>Dr. Ashish Jolly, <sup>4</sup>Dr. Shakti Kumar, <sup>5</sup>Gayatri Kumari, <sup>6</sup>Dr. Deepak Chandra Uprety, <sup>7</sup>Dr. Deepak Dagar, <sup>8</sup>Dr. Sanjeev Kumar

<sup>1</sup>Assistant Professor, Department of Computer Science, Shyam Lal College Evening (University of Delhi), Shahdara, New Delhi, INDIA, Email: [singhparbhakar87@gmail.com](mailto:singhparbhakar87@gmail.com)

<sup>2</sup>Associate Professor, Department of Computer Science, GNIOT Institute of Professional Studies, Greater Noida (UP), INDIA

<sup>3,4</sup>Assistant Professor, Department of Computer Science, Govt. PG College, Ambala Cantt. (Haryana), INDIA

<sup>5</sup>Assistant Professor, Department of Computer Science, Sanatan Dharma College, Ambala Cantt. (Haryana), INDIA

<sup>6</sup>Associate Professor and Research Coordinator, Department of Computer Science and Engineering, Noida Institute of Engineering and Technology, Knowledge Park 2, Greater Noida (UP), INDIA

<sup>7</sup>Associate Professor, Department of Business Administration, Maharaja Agrasen Institute of Management Studies, New Delhi, INDIA

<sup>8</sup>Sr. Assistant Professor, Department of Computer Science, Maharaja Agrasen Institute of Technology, New Delhi, INDIA

**Abstract:** Breast tumour, marked by the uncontrolled proliferation of breast cells, presents in various forms, most notably infiltrative ductal carcinoma and infiltrative lobular carcinoma both capable of metastasis. With the rise of artificial intelligence in healthcare, machine learning has become a vital tool in the early identification of such conditions. This study introduces an ensemble machine learning based system aimed at the early and accurate classification of breast tumors as malignant or benign. Utilizing the Wisconsin Breast Tumour Dataset (WBCD) from the UCI Machine Learning Repository, which comprises 30 diagnostic features including mean, standard error, and extreme values, the proposed framework integrates multiple classifiers, Decision Tree, AdaBoost, Gaussian Naive Bayes, and Multi-Layer Perceptron into a stacked ensemble model. Performance was evaluated using metrics such as accuracy, sensitivity, specificity, and F1 score, with the ensemble achieving a superior accuracy of 96.66%. The findings demonstrate that the ensemble model enhances predictive reliability and supports timely diagnosis, with potential for adaptation in broader clinical and classification tasks.

**Keywords:** Breast Disease Classification, Machine Learning, Decision Tree, NB, AdaBoost, Ensemble learning

## 1. INTRODUCTION

Breast cancer remains one of the most prevalent and life-threatening diseases affecting women globally. Early and accurate detection is critical for improving survival rates, minimizing treatment complexity, and enhancing patients' quality of life. While traditional diagnostic tools such as mammography, ultrasound, and biopsy—have long played a key role in early detection, they present notable challenges. These include high false positive rates, inter observer variability, and limited accessibility, particularly in under resourced regions.

To address these limitations, the integration of machine learning (ML) into diagnostic workflows has emerged as a powerful solution. Early computational approaches, including basic statistical methods and conventional ML algorithms like decision trees and support vector machines, offered incremental improvements in diagnostic accuracy. However, these models often struggled with small, imbalanced datasets and lacked generalizability across diverse patient populations [1]. The recent surge in artificial intelligence capabilities combined with access to large, high quality medical datasets—has opened new avenues for enhanced diagnostic systems.

Among these, ensemble learning has gained significant traction for its ability to improve classification performance and model robustness. By combining multiple base learners such as decision trees, deep

neural networks, and support vector machines ensemble methods mitigate the weaknesses of individual models and reduce overfitting. Techniques such as bagging, boosting, and stacking allow these systems to learn from diverse patterns in data, resulting in more reliable and precise diagnostic outcomes [2-3]. Additionally, modern frameworks often integrate advanced feature selection, image processing, and explainable AI to ensure both effectiveness and interpretability in clinical environments.

Looking forward, the future of breast tumor detection lies in the convergence of ensemble machine learning with next generation technologies. Innovations such as quantum computing, federated learning, and personalized medicine are expected to revolutionize predictive healthcare. These systems may eventually incorporate real time patient monitoring, wearable devices, and genomic data to enable highly personalized and context aware diagnostics. Moreover, ethical AI practices and explainable models will be essential for fostering trust and adoption in clinical settings [4].

This research proposes the development of an ensemble machine learning based framework for the early detection of breast tumors. The framework aims to combine multiple ML algorithms with optimized feature selection and robust data preprocessing techniques to maximize diagnostic performance. Different ensemble strategies including bagging, boosting, and stacking—will be evaluated to determine their effectiveness in improving prediction accuracy. The model will be validated using real world medical datasets to assess its clinical relevance and generalizability [5]. This work aspires to contribute to the development of intelligent, scalable, and interpretable diagnostic tools that can support timely and accurate breast cancer detection.

## 2. REVIEW OF LITERATURE

Over the years, researchers have employed a variety of machine learning algorithms to enhance the accuracy of breast tumour 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 tumour 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 tumour 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 tumour detection. These advancements highlight the continuous evolution of machine learning techniques in medical diagnostics, aiming for higher accuracy, early detection, and improved patient outcomes.

## 3. MACHINE LEARNING BASED CLASSIFICATION

Machine learning based classification approaches have gained significant attention in medical diagnostics, particularly in breast tumour 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 tumour 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 pre-processed 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 tumour 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 (Gaussian NB), 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 tumour classification (Figure 1).

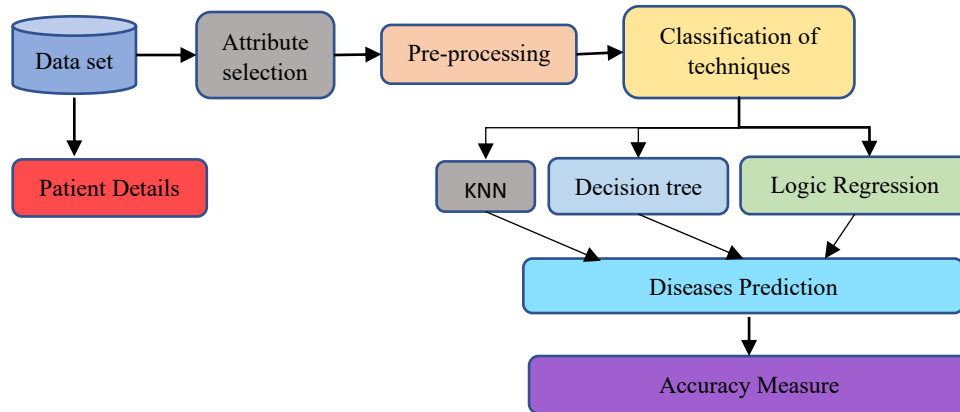


Figure 1: Illustrating the ML based sequential stages from data preprocessing to model evaluation

The proposed framework begins with preprocessing the Wisconsin Breast Tumour 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.

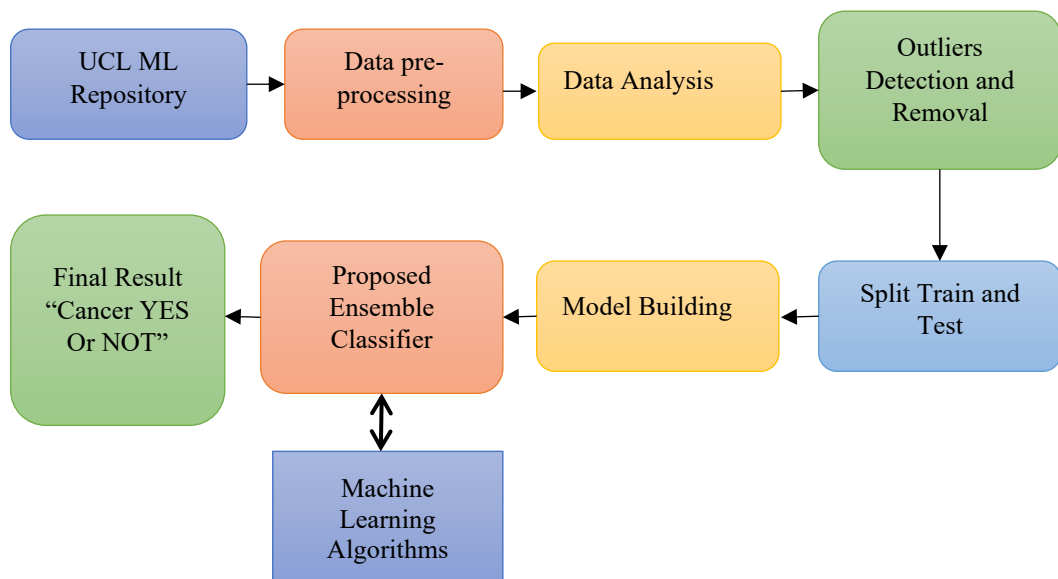


Figure 2: Proposed research methodology for machine learning based breast tumour classification

10.48047/jocaaa.2024.33.08.144

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 tumour detection, contributing to improved patient outcomes and timely medical interventions (Figure 2).

#### 4.1 Algorithm

```

T' = N(T) /* Normalize the dataset */
Let H = {h1, h2, h3, ..., hn} /* Given dataset */
E = {E1, E2, E3, ..., En} /* 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  $(\text{perimeter}^2 / \text{area} - 1.0)$ , 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 tumour 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	$(\text{Perimeter}^2 / \text{Area}) - 1.0$ .
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.

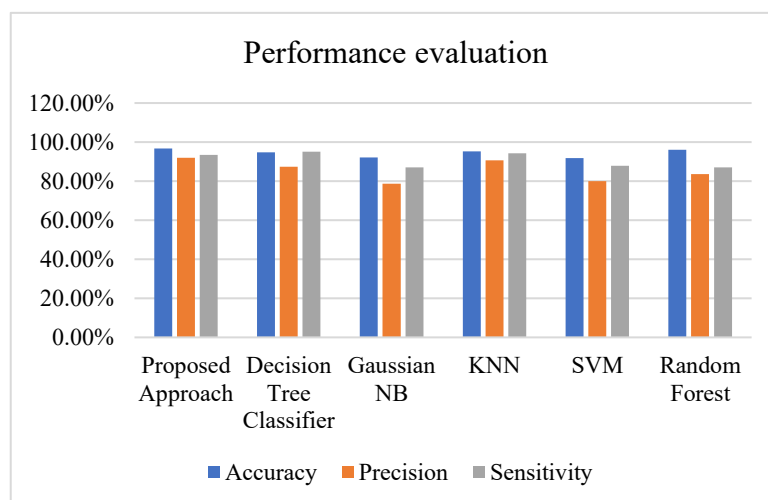


Figure 3: Performance Evaluation Metrics for Machine Learning Models

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).

## 6. CONCLUSION

In conclusion, the proposed ensemble machine learning-based framework demonstrates high accuracy and robustness in the early detection of breast tumors, effectively distinguishing between malignant and benign cases. By leveraging the strengths of diverse classifiers—Decision Tree, AdaBoost, Gaussian Naive Bayes, and Multi-Layer Perceptron—within a stacked ensemble model, the system achieved a notable accuracy of 96.66% on the Wisconsin Breast Cancer Dataset. This result highlights the potential of ensemble approaches in improving diagnostic performance and reducing false predictions, ultimately contributing to more timely and accurate clinical decisions. Looking ahead, future research can focus on expanding the framework to include deep learning models and image-based data from mammograms or histopathological slides to enhance diagnostic precision further. Additionally, the integration of real-time clinical data, genomic markers, and wearable device inputs could lead to more personalized and dynamic tumor detection systems. Exploring federated learning and privacy-preserving techniques will also be crucial to enable secure deployment in real-world healthcare settings. Finally, efforts toward developing interpretable AI models will be essential to support clinical trust and ensure ethical, transparent decision-making.

## 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 Tumour 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.

10.48047/jocaaa.2024.33.08.144

- [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/b978-0-444-52701-1.00124-1.