ISSN: 2633-4828 Vol. 5 No. S4, (July-August, 2023)

International Journal of Applied Engineering & Technology

ADVANCEMENT IN THYROID DETECTION USING XGBOOST ALGORITHM

Asish Paul Stains[1], Aman Singh[2], Rhea Roy[3], S. Vasanthadev Suryakala[4*]

[1, 2, 3, 4] Department of Electronics and Communication, College of Engineering and Technology, SRM Institute of Science and Technology, Kattankulathur, Chennai, India

[1] as4667@srmist.edu.in, [2] as9099@srmist.edu.in, [3] rr7855@srmist.edu.in, [4] suryakas@srmist.edu.in

[*]Corresponding Author: S. Vasanthadev Suryakala, suryakas@srmist.edu.in

Abstract

Thyroid disorders represent a prevalent health concern, necessitating accurate and efficient diagnostic methods for timely intervention. This research explores the comparative analysis of Logistic Regression, Bagging SVM, Decision Tree, Random Forest and XGBoost, and the application of the XGBoost algorithm for thyroid detection, aiming to improve diagnostic accuracy and efficiency. The study involves the utilization of a diverse dataset comprising various thyroid-related parameters and employs XGBoost, a powerful machine-learning algorithm known for its effectiveness in classification tasks. Through comprehensive experimentation and evaluation, our findings indicate promising results in enhanced sensitivity and specificity in thyroid detection. This research contributes to the growing body of literature on integrating advanced machine learning techniques in medical diagnostics, offering a potential avenue for more reliable and timely identification of thyroid conditions. The implications of these findings extend to the broader field of healthcare, highlighting the potential of XGBoost in optimizing thyroid disorder detection methodologies. The collective efforts of numerous researchers yielded suboptimal results. Consequently, our study undertook a comparative analysis involving XGBoost, Bagging SVM, Decision Tree, Random Forest, and Logistic Regression models. We determined that Logistic Regression obtained an accuracy of 62.02%, Decision Tree obtained an accuracy of 98.54%, Random Forest obtained an accuracy of 98.11%, SVM obtained an accuracy of 61.19% and XGBoost obtained an accuracy of 98.66%. Through rigorous analysis, XGBoost outperformed the others, achieving a remarkable accuracy of 98.66%.

Keywords: *XGBoost, Bagging SVM, Decision Tree, Random Forest, Logistic Regression.*

INTRODUCTION

The thyroid gland plays a crucial role in regulating metabolism and overall bodily functions. Undetected and untreated thyroid disorders may lead to serious health implications, such as cardiovascular problems and mental health challenges. Supervised learning, a machine learning technique, is preferred for thyroid detection because of its ability to effectively analyze intricate medical data[1]. Through the utilization of labeled data, supervised learning algorithms can discern subtle patterns suggestive of thyroid dysfunction, facilitating early identification and tailored treatment strategies to enhance patient well-being. Numerous studies have delved into machine learning techniques for thyroid disease classification and detection, each offering distinct insights [4]. Alyas et al. (2022) proposed a pragmatic framework involving preprocessing steps like image normalization and feature selection, while Singh et al. (2022) showcased the remarkable performance of the Gradient Boosting Classifier after fine-tuning. Prathibha et al. (2022) demonstrated notable accuracy in classifying thyroid diseases from X-ray images using a modified ResNet architecture, and Chandan et al. (2021) developed a practical web application integrating logistic regression for thyroid disease prediction based on blood test data, suggesting enhancements involving ultrasound scans for improved detection. These studies underscore the potential of machine learning methodologies in diagnosing and classifying thyroid diseases, offering avenues for further research. However, challenges such as data quality, model interpretability, generalization, feature selection, and clinical validation must be addressed to ensure the effectiveness and applicability of these models. Collaboration between machine learning researchers, clinicians, and domain experts is essential to overcome these challenges and advance towards more accurate and clinically relevant solutions. Our approach addresses these challenges comprehensively by utilizing real-time data obtained from blood reports of actual patients, incorporating over 3000 samples and employing five different algorithms. We have meticulously assessed performance metrics such as accuracy, precision, recall, and F1 score, ultimately determining XGBoost as the superior algorithm for thyroid disease detection.

METHODOLOGY

Thyroid detection initiates with meticulous data collection, gathering essential information for the study. Subsequently, the collected data undergoes pre-processing to ensure cleanliness and analysis readiness, including steps like data cleaning and normalization. The dataset is then partitioned into two subsets: training data and test data. Fig.1 shows a clear depiction of the framework.

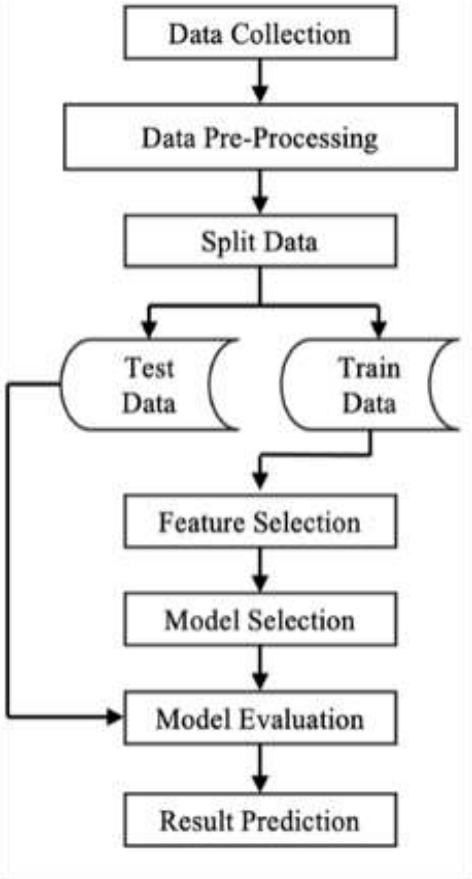

Fig.1 Methodology Framework

The training data undergoes rigorous feature selection and model selection processes to enhance its effectiveness in capturing relevant patterns and relationships. Simultaneously, the test data is fed into the chosen model for evaluation, assessing its performance and generalizability to unseen data. Post-evaluation, the system generates predictions based on the model's performance, offering valuable insights into the accuracy and efficiency of the selected approach. This iterative process of data collection, pre-processing, model training, evaluation, and prediction ensures the development of robust machine learning models for accurate thyroid disorder detection. These models facilitate timely interventions and personalized treatment strategies, ultimately improving patient outcomes and quality of care.

ALGORITHMS

Five algorithms have been applied - Bagging SVM, Decision Tree, Logistic Regression, Random Forest, and XGBoost - to the methodology outlined above. The necessity for employing different algorithms in thyroid detection arises from the complexity of medical data, diverse feature representation capabilities, and varying performance metrics. Each algorithm brings unique strengths, like interpretability or computational efficiency, enhancing the overall accuracy and effectiveness of thyroid detection systems.

1. Bagging SVM

Bagging SVM is an ensemble technique for improving SVM performance and stability. It creates multiple subsets of training data via bootstrap sampling, trains individual SVM models independently, and aggregates predictions through majority voting for classification or averaging for regression tasks[1]. It mitigates overfitting and enhances generalization, particularly beneficial when the base SVM model exhibits high variance. Fig.2 illustrates how Bagging SVM works. The effectiveness depends on dataset characteristics and learning problem challenges.

Let D be the original dataset with N samples. A bootstrap sample D_i is created by randomly sampling N samples with replacement from D [8].

$$D_i = \{(x_1, y_1), (x_2,y_2),...,(x_N, y_N)\} \quad (1)$$

where

x = input features or independent variables of the dataset

y = corresponding output or dependent variable of the dataset

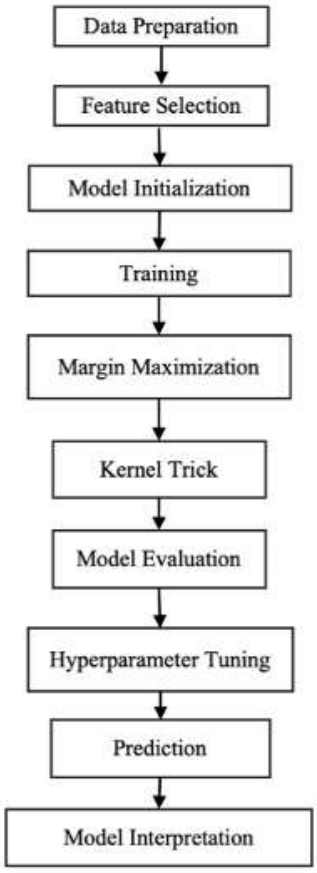

Fig.2 Bagging SVM Block Diagram

2. Decision Tree

Decision trees in supervised machine learning partition continuous data based on specified parameters. Attributes form internal nodes, conveying decision rules through branches to interpret final outcomes at terminal nodes. Categorization follows a Top-Down approach, sorting from root to leaf nodes[8]. Fig.3 illustrates how Decision Tree works.

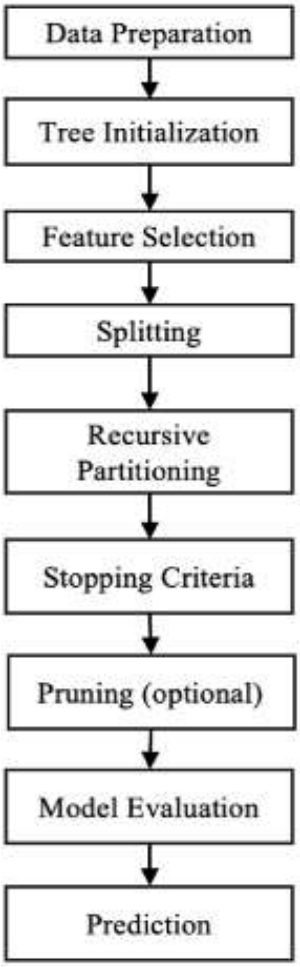

Fig.3 Decision Tree Block Diagram

Training data subdivides into smaller subsets using metrics like Gini Index, Information Gain, Entropy, Gain Ratio, and Chi-Square. The Gini Index is a statistical measure used to quantify inequality in a distribution, ranging from 0 (perfect equality) to 1 (maximal inequality). The Information Index and Gini Index, expressed through Eqs. (1) and (2) respectively, are commonly employed in assessing the majority of datasets. This process is iterated for each child tuple until all tuples belong to the same class, and no additional attributes are required. The utilization of the Gini Index contributes to heightened precision and diagnostic accuracy[8].

Information Gain = Entropy(S) − [(weighted avg) ∗ Entropy (each feature)]. (1)

$$\text{Gini Index} = 1 - \sum_{i=1}^{c} P^2{}_i \quad (2)$$

In Equation (2),$P_i$represents the fraction of samples associated with class c within a specific node.

$$\text{To calculate Entropy(S)} = \sum_{i=1}^{c} - p_i \log_2 p_i \quad (3)$$

3. Logistic Regression

Logistic regression is a statistical technique used to model the relationship between a dependent variable (target) and one or more independent variables (features) [2]. Its objective is to identify the optimal logistic relationship that explains how changes in predictors influence changes in the target variable. Fig.4 shows the working of Logistic Regression.

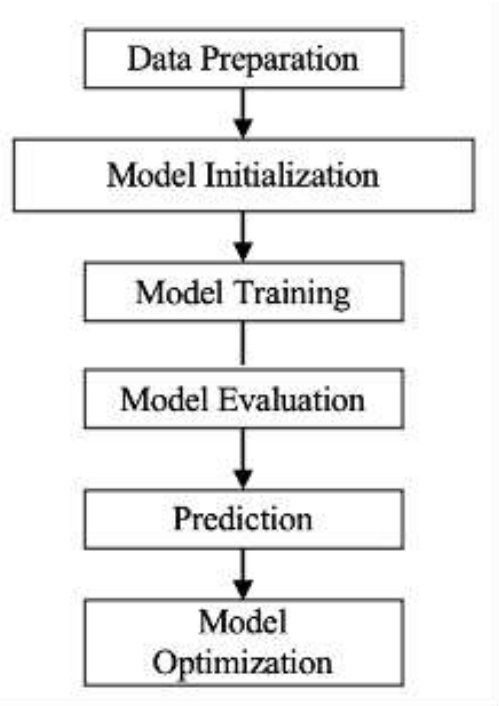

Fig.4 Logistic Regression Block Diagram

During training, coefficients are calculated to minimize the disparity between predicted and actual values, often employing optimization methods like gradient descent [8]. In logistic regression, there is only one independent variable, and the relationship is represented by a straight-line equation:

$$y=mx+b \quad (4)$$

where:

y represents the dependent variable,

x denotes the independent variable,

m signifies the slope of the line, and

b is the y-intercept [7].

For multiple logistic regression, with more than one independent variable, the equation becomes:

$$y=b_0+b_1x_1+b_2x_2+\ldots+b_nx_n \quad (5)$$

where:

b_0 = y-intercept,

b_1, b_2,…,b_n = coefficients for each independent variable $x_{1,2}$,…,x_n [7].

4. Random Forest

Random Forest, a popular ensemble learning algorithm, combines outputs from multiple decision trees to produce a single result. It incorporates bagging and feature randomness, selecting a random subset of features to minimize correlation among trees[3]. This enhances performance and identifies significant elements, crucial for classification accuracy.

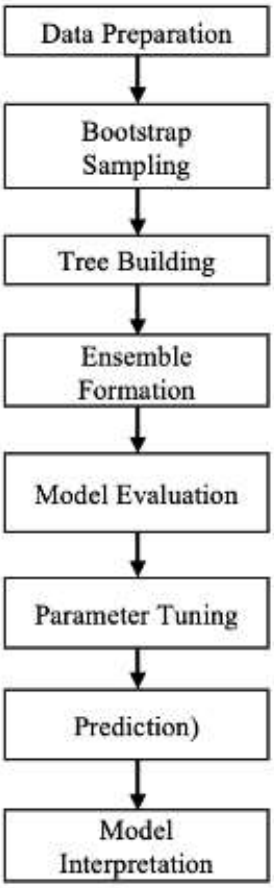

Fig.5 Random Forest Block Diagram

Feature selection methods like filtering and encapsulation improve accuracy by focusing on relevant features independently of the algorithm. Random forests offer more reliable forecasts compared to single decision trees, making them superior in ensemble learning.

Procedure for feature selection:

$$\text{norm } fi_i = \frac{fi_i}{\Sigma_{ij \in \text{all features}} f_i} \quad (6)$$

where

norm fi_i - the normalized importance of feature i

fi - the importance of feature i

$$\text{RF } fi_i = \frac{\Sigma_{j=\text{all trees}} normfi_i}{T} \quad (7)$$

where

RF fi_i - the importance of feature i calculated from all trees in random forest model

norm fi_i - The normalised feature importance for i in tree j

T - the total number of trees.

5. XGBoost

XGBoost, an abbreviation for eXtreme Gradient Boosting, stands out as a highly acclaimed and effective machine learning algorithm suitable for both classification and regression tasks. Rooted in the gradient boosting framework, XGBoost systematically builds an ensemble of weak learners, often in the form of decision trees. Each subsequent learner refines errors from its predecessors, contributing significantly to the algorithm's efficacy[8]. Fig.6 shows the working of XGBoost.

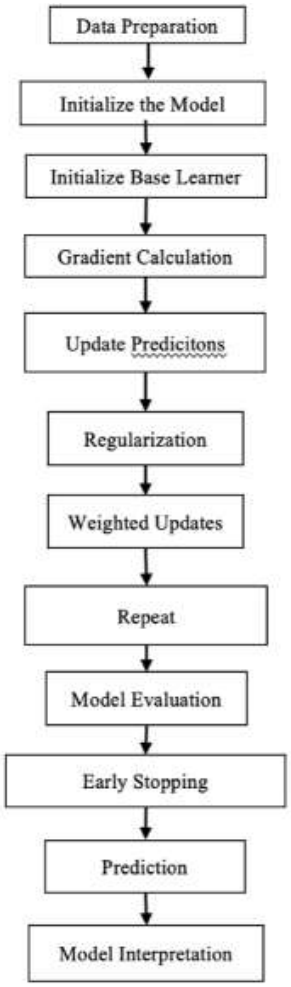

Fig.6 XGBoost Block Diagram

Known for its rapid computation, outstanding performance, and scalability, XGBoost has risen to prominence in the realm of machine learning. Its standout features include a strong emphasis on combating overfitting through regularization, the use of parallel computing techniques for faster processing, and innovative methods like tree pruning. This adaptability also extends to effectively handling missing data values and accommodating various data formats, solidifying XGBoost as a robust and versatile tool for predictive modeling across diverse domains. As a result, it has become a cornerstone in contemporary machine learning applications [5].

$$\min[obj] = \min\{L(F_m(X), Y+\Omega(f) + C\} \quad (8)$$

where:

$L(F_m(X), Y)$ = Loss Function.

It independently handles missing data and oversees structured databases within classification and regression predictive modeling duties. XGBoost anticipates residuals or errors from previous models, merging them to generate the ultimate prediction. Additionally, it improves effectiveness by addressing overfitting concerns in datasets.

RESULTS

The dataset, acquired from Kaggle, consists of 3772 patient samples with 30 attributes, focusing on thyroid disease records from the Garavan Institute and J. Ross Quinlan of the New South Wales Institute, Sydney, Australia, dating back to 1987. Its primary aim is binary classification, distinguishing thyroid disease presence or absence in each patient. With ample data for analysis, it enables exploration of attribute relationships and the binary target variable, potentially improving diagnostic accuracy and patient care strategies.

Attributes describe patient characteristics, medical measurements, demographics such as age and sex, medical conditions, treatments, symptoms, and measurements of thyroid-related hormones and substances in the blood like TSH, T3, TT4, T4U, FTI, and TBG levels, with indicators of whether these measurements were taken. Additionally, attributes indicate psychological concerns, surgeries, and referral sources. Each attribute has a corresponding data type, aiding comprehensive analysis of thyroid-related data.

The dataset contains 30 attributes grouped into labels, binary categorical features, and continuous numerical features. Labels like age and sex provide demographic data. Binary categorical features such as 'on_thyroxine' and 'sick' reveal medical history, while continuous numerical features like 'T3' and 'TSH' aid in diagnosing thyroid disorders. Biochemical profiles distinguish thyroid disorders, such as hyperthyroidism with elevated T3 and/or T4 and decreased TSH, and primary hypothyroidism with reduced T3 and/or T4 and elevated TSH. These profiles inform tailored treatment strategies for effective management.

Data Pre-Processing

i. Handling Missing Values:

The dataset undergoes thorough scrutiny to identify the presence of '?' symbols, which typically signify missing or unknown values. Once detected, these '?' values are systematically replaced with numpy.nan, ensuring a consistent approach to managing missing data throughout the dataset.

This process is crucial for maintaining data integrity and reliability during subsequent analysis and modeling stages. Fig.7 shows the graph of missing values of the dataset.

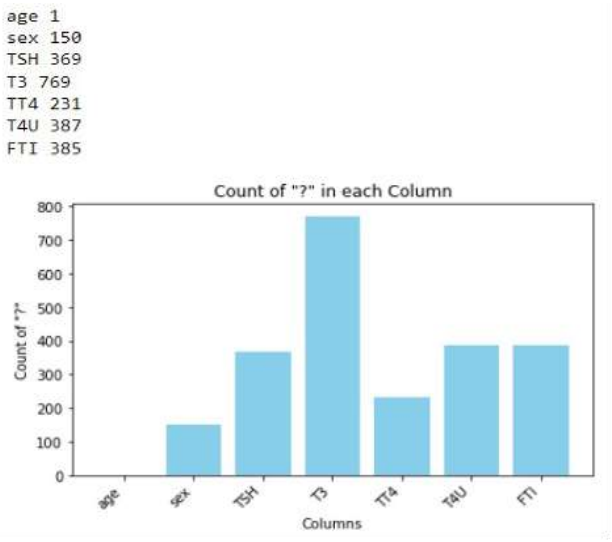

Fig.7 Missing Values of the dataset

Fig.8 shows the handling of missing values after pre-processing.

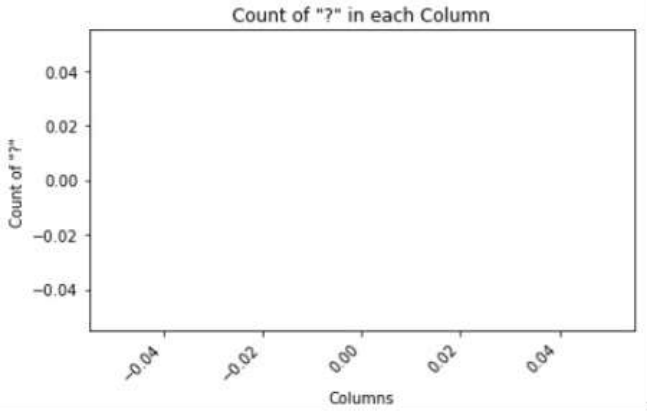

Fig.8 Missing Values handled

ii. Handling Categorical Data:

Categorical data, like gender, undergoes Label Encoding for numerical representation. Binary categorical columns are mapped to 0 and 1 for consistency, streamlining interpretation and analysis in machine learning algorithms. Fig.9 shows the graph for Categorical Data of the dataset.

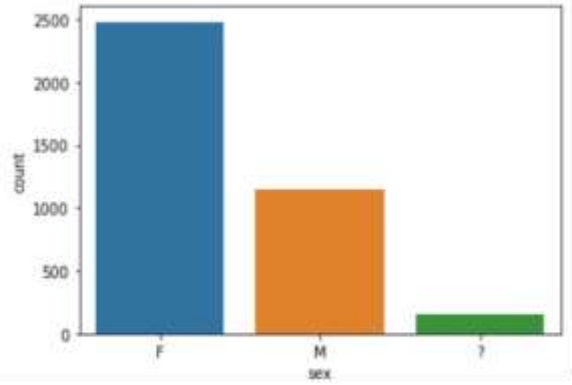

Fig.9 Categorical Data

Fig.10 shows the handling of categorical data after pre-processing.

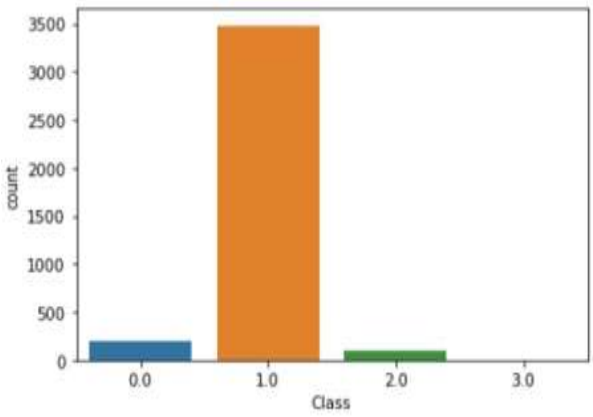

Fig.10 Categorical Data handled

iii. Handling Imbalanced Data:
The imbalanced-learn library is used to oversample the minority class using RandomOverSampler.

Fig.11 shows the graph of imbalanced data of the dataset.

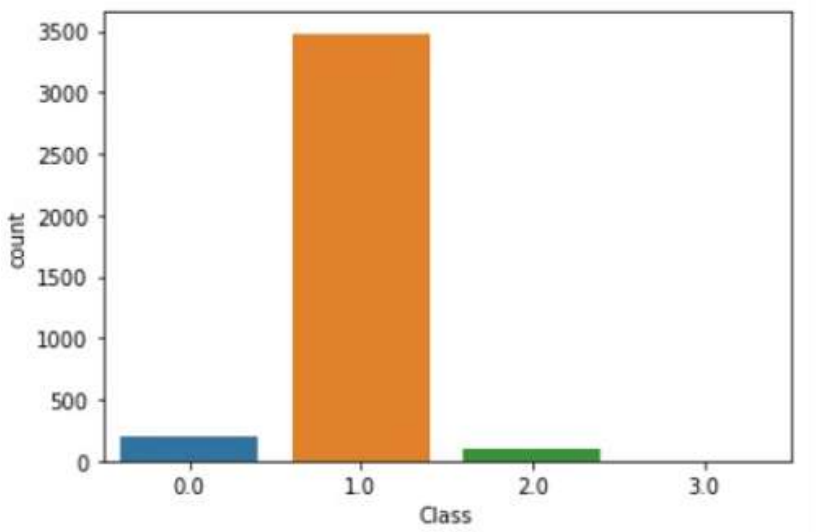

Fig.11 Imbalanced Data

Fig.12 shows how imbalanced data was handled after pre-processing.

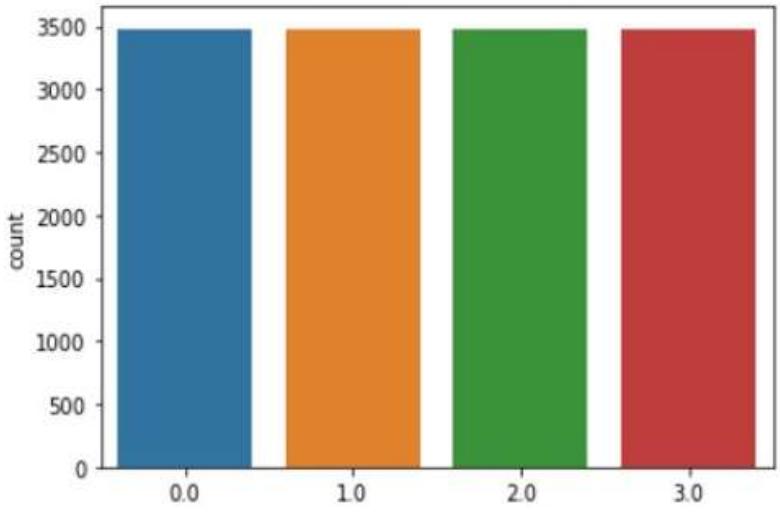

Fig.12 Balanced Data

ROC Curve

The Receiver Operating Characteristic (ROC) curve serves as a valuable graphical tool employed in the evaluation of binary classification models. Its primary function is to illustrate the trade-off between the True Positive Rate (sensitivity) and the False Positive Rate (specificity subtracted from one) as the discrimination threshold of the classifier undergoes variation. By depicting this relationship, the ROC curve aids in the assessment of a model's ability to differentiate between positive and negative instances across a spectrum of threshold settings.

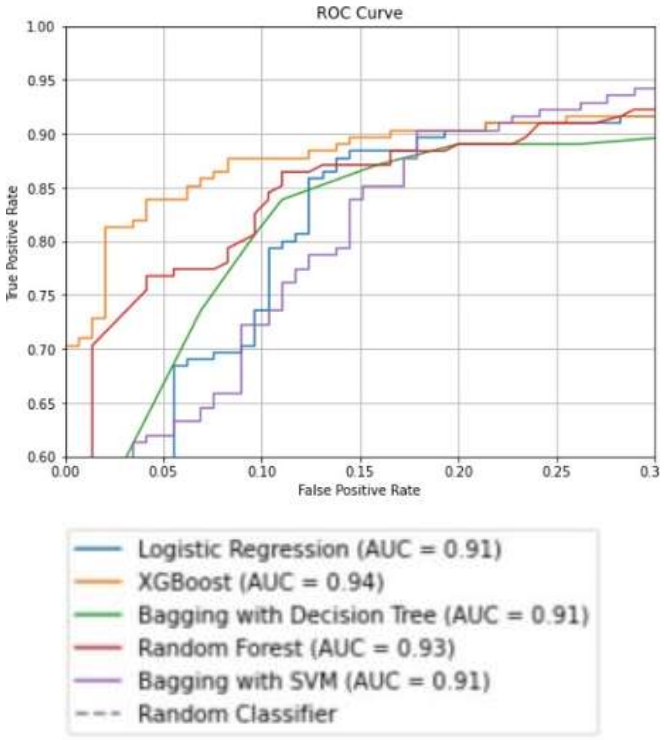

Fig.13 ROC Curve of all algorithms

A curve closer to the top-left corner of the plot indicates better performance, with the ideal point representing perfect classification. Overall performance is quantified by the Area Under the Curve (AUC), where a higher AUC value suggests superior discriminative ability of the model. Fig.13 indicates that XGBoost has achieved the highest AUC value of 0.94, making it the superior performing algorithm in comparison.

Validation

To effectively evaluate a deep learning model, it's essential to incorporate validation using key classification metrics like Precision, Accuracy, F1-score, and Recall. These metrics are computed based on the counts of True Positive, False Positive, True Negative, and False Negative. Below are the equations used to calculate these metrics[8].

Accuracy: In classification tasks, accuracy measures the proportion of correct predictions made by the model out of all potential predictions. It's computed by dividing the number of correct predictions by the total predictions, then multiplying by 100 [8].

$$\text{Accuracy} = \frac{T_P+T_N}{T_P+F_P+F_N+T_N}$$

In this context, TP (True Positive) refers to instances where actual euthyroid sick syndrome (1) is correctly classified as sick-euthyroid (1). TN (True Negative) indicates cases where the absence of euthyroid sick syndrome (0) is accurately classified as Negative (0). FP (False Positive) occurs when the absence of euthyroid sick syndrome (0) is erroneously classified as sick-euthyroid (1). FN (False Negative) denotes instances where euthyroid sick syndrome (1) is incorrectly classified as Negative (0).

Precision: Precision represents the ratio of true positives and total positives predicted [8].

$$\text{Precision} = \frac{T_P}{T_P+F_P}$$

Precision signifies the ratio of individuals diagnosed with euthyroid sick syndrome who actually have it, derived from true positives (TP) and false positives (FP).

Recall: Recall measures the ratio of true positives to all positive instances in the ground truth [4].

$$\text{Recall} = \frac{T_P}{T_P + F_N}$$

Recall indicates the percentage of individuals with euthyroid sick syndrome accurately identified as sick-euthyroid by the algorithm. True positives (TP) encompass both genuine positives (euthyroid sick syndrome) and false negatives (FN), representing individuals with the condition despite the model's erroneous prediction

F-1 Score: The F1-score combines precision and recall metrics, representing their harmonic mean. A high F1-score indicates robust precision and recall, rendering it valuable for handling imbalanced classification tasks [4].

$$\text{F-1 Score} = \frac{2}{\frac{1}{\text{Precision}} + \frac{1}{\text{Recall}}}$$

Based on the metrics and formulas provided above, Table 1 presents a comparative analysis of all algorithms, utilizing the same evaluation criteria.

Table I: Comparative Analysis of all algorithms

Algorithm	Accuracy $\frac{T_P+T_N}{T_P+F_P+T_N+F_N}$	Precision $\frac{T_P}{T_P+F_P}$	Recall $\frac{T_P}{T_P+F_N}$	F1 Score $\frac{Precsion*Recall}{2}$
Logistic Regression	62.02%	59.20%	62.03%	59.40%
Decision tree	98.54%	98.60%	98.53%	98.53%
Random Forest	98.11%	98.15%	98.10%	98.11%
SVM	61.19%	58.82%	61.11%	56.56%
XGBoost	98.66%	98.72%	98.65%	98.66%

The given table above clearly demonstrates that Bagging SVM resulted in an accuracy of 61.19%. Logistic Regression resulted in an accuracy of 62.02%. Random Forest resulted in an accuracy of 98.11%. Decision Tree resulted in an accuracy of 98.54%. XGBoost resulted in an accuracy of 98.66%.

As a result, it can be determined that XGBoost has been prevailed as the most accurate algorithm with an accuracy of 98.66%.

CONCLUSION

In conclusion, the application of ensemble learning algorithms for thyroid detection represents a significant advancement in medical diagnostics. Utilizing ensemble learning algorithms for thyroid detection offers a precise and efficient diagnostic approach. These algorithms, such as support vector machines and decision trees, analyze patient data to identify thyroid disorders early. With the increasing prevalence of thyroid abnormalities globally, early detection is essential for effective treatment and management. Machine learning enables non-invasive and cost-effective screening, particularly beneficial in regions with limited healthcare access. By leveraging comprehensive datasets, these models enhance diagnostic accuracy and streamline the diagnostic process. Incorporating ensemble learning in thyroid detection not only improves patient outcomes but also reduces

healthcare costs and burdens on healthcare systems. As technology advances, the integration of ensemble learning shows potential for revolutionizing thyroid disorder management and enhancing overall healthcare delivery.

REFERENCES

[1] Alyas, T., Hamid, M., Alissa, K., Faiz, T., Tabassum, N., & Ahmad, A, "Empirical Method for Thyroid Disease Classification Using a Machine Learning Approach," BioMed Research International, 2022.

[2] Singh, V. K., Yadav, N. D., Singh, R. K., & Sahu, M, "Detection of Thyroid Using Machine Learning Approach", ISSN:-2394-0298, 3, 65-80, NIU International Journal of Human Rights, 2022.

[3] Anari, S., Tataei Sarshar, N., Mahjoori, N., Dorosti, S., & Rezaie, A, "Review of Deep Learning Approaches for Thyroid Cancer Diagnosis.", Mathematical Problems in Engineering, 2022.

[4] Soma Prathibha1,*, Deepak Dahiya2, C. R. Rene Robin3, Cherukuru Venkata Nishkala4 and S. Swedha, "A Novel Technique for Detecting Various Thyroid Diseases Using Deep Learning",2022.

[5] Pothamsetty Pavani1, Dr.P.P.Sadu Naik M.Tech, PhD2. 1PG Student, "Prediction of Thyroid Disease Using Machine Learning Techniques.", Department of CSE, Dr. Samuel George Institute of Engineering and Technology, Markapuram - 523316, A.P, India2 Professor, Department of CSE, Dr. Samuel George Institute of Engineering and Technology, Markapuram - 523316, A.P, India

[6] Rajasekhar Chaganti 1, Furqan Rustam 2,† , Isabel De La Torre Díez 3, Juan Luis Vidal Mazón 4,5,6 , Carmen Lili Rodríguez 4,7 and Imran Ashraf 8, "Thyroid Disease Prediction Using Selective Features and Machine Learning Techniques" , 2022.

[7] Chandan R Department of ECE,DSU, Bangalore, Karnataka, India. Chethan MS Department of ECE. DSU, Bangalore, Karnataka, India, Chetan Vasan Department of ECE DSU,Bangalore, Karnataka, India. Devikarani H S Department of ECE. DSU, Bangalore, Karnataka, India, "Thyroid detection using machine learning.", 2022.

[8] L. Ozyılmaz and T. Yıldırım, "Diagnosis of thyroid disease using artificial neural network methods," in: Proceedings of ICONIP'02 9th international conference on neural information processing (Singapore: Orchid Country Club, 2002) pp. 2033–2036.

[9] K. Polat, S. Sahan and S. Gunes, "A novel hybrid method based on artificial immune recognition system (AIRS) with fuzzy weighted pre- processing for thyroid disease diagnosis," Expert Systems with Applications, vol. 32, 2007, pp. 1141-1147.

[10] F. Saiti, A. A. Naini, M. A. Shoorehdeli, and M. Teshnehlab, "Thyroid Disease Diagnosis Based on Genetic Algorithms Using PNN and SVM," in 3rd International Conference on Bioinformatics and Biomedical Engineering, 2009. ICBBE 2009.

[11] G. Zhang, L.V. Berardi, "An investigation of neural networks in thyroid function diagnosis," Health Care Management Science, 1998, pp. 29- 37. Available: http://www.endocrineweb.com/ thyroid.html, (Accessed: 7 August 2007).

[12] Hosseinzadeh, M.; Ahmed, O.H.; Ghafour, M.Y.; Safara, F.; Hama, H.; Ali, S.; Vo, B.; Chiang, H.S., "A multiple multilayer perceptron neural network with an adaptive learning algorithm for thyroid disease diagnosis in the internet of medical things.", J. Supercomput. **2021**, 77, 3616–3637.

[13] Abbad Ur Rehman, H.; Lin, C.Y.; Mushtaq, Z, "Effective K-Nearest Neighbor Algorithms Performance Analysis of Thyroid Disease.", J. Chin. Inst. Eng. **2021**, 44, 77–87.

[14] Mishra, S.; Tadesse, Y.; Dash, A.; Jena, L.; Ranjan, P., "Thyroid disorder analysis using random forest classifier.", In Intelligent and Cloud Computing; Springer: Berlin/Heidelberg, Germany, 2021; pp. 385–390.

[15] Alyas, T.; Hamid, M.; Alissa, K.; Faiz, T.; Tabassum, N.; Ahmad, A., "Empirical Method for Thyroid Disease Classification Using a Machine Learning Approach.", BioMed Res. Int. **2022**, 2022, 9809932.

[16] Jha, R.; Bhattacharjee, V.; Mustafi, A., "Increasing the Prediction Accuracy for Thyroid Disease: A Step towards Better Health for Society.", Wirel. Pers. Commun. **2022**, 122, 1921–1938.

[17] Sankar, S.; Potti, A.; Chandrika, G.N.; Ramasubbareddy, S., "Thyroid Disease Prediction Using XGBoost Algorithms.", J. Mob. Multimed. **2022**, 18, 1–18.

[18] Wajner, S.M.; Maia, A.L., "New insights toward the acute non-thyroidal illness syndrome.", Front. Endocrinol. **2012**, 3, 8.

[19] Cai, J.; Luo, J.; Wang, S.; Yang, S., "Feature selection in machine learning: A new perspective.", Neurocomputing **2018**, 300, 70–79.

[20] UCI. UCI Machine Learning Repository: Thyroid Disease Data Set. Available online: https://archive.ics.uci.edu/ml/datasets/th yroid+disease.