ThyroStack: A Stacking Model for Thyroid Disease Prediction

Kazi Hasebul Islam[1], Dilu Biswas[1], Md. Mohibbullah Akash[1], Tanpia Tasnim[1], and Al Zadid Sultan Bin Habib[2]
[1]Department of Computer Science and Engineering, Green University of Bangladesh, Narayanganj-1461, Dhaka, Bangladesh
[2]Lane Department of Computer Science and Electrical Engineering, West Virginia University, Morgantown, WV 26505, USA
Email: hasebulislam47@gmail.com, dilubiswas750@gmail.com, mohibbullahakash17@gmail.com, tanpia@cse.green.edu.bd, ah00069@mix.wvu.edu

***Abstract*—Thyroid disorders, which impact the thyroid gland's functionality, present substantial health challenges globally. Current research in thyroid disease emphasizes prediction using machine learning (ML) and deep learning (DL) algorithms, with a particular interest in ensemble models for prediction. However, existing thyroid disease prediction models often encounter challenges with imbalanced datasets, limited data sizes, and inadequate feature selection. In this paper, a stacking model was introduced for predicting thyroid disease. The model utilized Extreme Gradient Boosting (XGBoost) and Multilayer Perceptron (MLP) as base models for thyroid disease prediction. These models were integrated into a stacking framework where Logistic Regression (LR) is the meta-model, aggregating the predictions from the base models to generate final predictions. Also, the feature selection technique was applied to identify significant features. Additionally, a SMOTE-ENN a hybrid sampling technique was incorporated to address the class balancing issue in the thyroid dataset, enhancing the model's ability to learn effectively from both classes. Furthermore, a SHapley Additive exPlanations (SHAP) technique was used to enrich the interpretability and transparency of the proposed model. The proposed stacking model demonstrated superior performance with an impressive accuracy of 99.78%. Finally, the stacking technique effectively combined the strengths of XGBoost and MLP, using their complementary features to boost predictive accuracy through LR. The proposed model enhanced prediction accuracy and offered valuable insights into the intricate patterns and relationships among features associated with thyroid disease.**

***Index Terms*—Thyroid Disease, Stacking Model, MLP, XGBoost, Logistic Regression, SHapley Additive exPlanations (SHAP).**

I. Introduction

Thyroid diseases are prevalent worldwide and can significantly impact an individual's health and well-being. More than 30 million Americans have hypothyroidism, which is the most common endocrine disease of the thyroid gland [1]. About 5 out of every 100 Americans ages 12 and up have hypothyroidism. Women are much more likely to be harmed than men; the rate of occurrence is up to 9 times higher in people who were assigned female at birth [1]. In Bangladesh, a few occupational groups said that nearly 7% of their members had hypothyroidism [2]. In the same way, almost 20% of people in India, a neighboring Southeast Asian country, had a thyroid problem. Hypothyroidism was the most common type of thyroid disease [2]. Hypothyroidism, also known as underactive thyroid, is a condition characterized by a deficiency in the production of certain hormones by the thyroid gland [3]. Hypothyroidism manifests with elevated serum TSH levels alongside normal serum-free T4 and T3 levels, progressing to a decrease in serum-free T4 levels over time [2].

Nowadays, thyroid disorders are a significant and rapidly spreading concern, demanding extensive time and effort for accurate investigation and diagnosis. Precisely predicting these diseases has emerged as a critical challenge within the healthcare system. Researchers primarily emphasize using ultrasound imaging to enhance the accuracy of predicting thyroid disease. The main drawback lies in the collection of high-quality ultrasound image datasets, which hampers the achievement of better-performing predictive models for thyroid disease. Additionally, ultrasound images suffer from inherent limitations due to susceptibility to noise and artifacts. Furthermore, extensive research is being conducted in the realm of thyroid disease, concentrating on aspects like thyroid disease prediction and predicting treatment using ML and DL algorithms. Certain researchers have employed ensemble models to predict thyroid disease. Existing thyroid disease prediction models frequently face challenges with imbalanced datasets, small data sizes, and limited feature selection [5] [6] [7]. To address the issues of the existing research, the following objectives are considered in achieving the goal:

- To introduce a stacking model to predict thyroid disease.
- To identify the most significant features contributing to the prediction.
- To apply hybrid sampling model for addressing class imbalance issue.
- To evaluate the performance of these models using appropriate metrics and compare them with traditional models.

The remaining portion of the paper is described as follows:
In Section II, the paper conducts a literature review focusing on related works relevant to the proposed idea. Section III outlines the methodology, including the workflow diagram,

979-8-3315-3197-3/24/$31.00

selection techniques, stacking model implementation, and methods for evaluating performance. Section IV presents the results, featuring a comparison table, confusion matrix, and ROC curve analysis. Section V concludes the paper, discussing its findings, and limitations, and suggesting directions for future research.

II. Literature Review

Numerous researches focus on predicting thyroid disease. Currently, ML and DL algorithms are pivotal in the prediction of thyroid disease. This section summarizes key findings from previous studies regarding the prediction of thyroid disease.

M.L. M. Prasad and R. Santhosh [5] introduced a novel multi-stage weight-based ensemble learning process to create proper decisions. Through the integration of multiple classifiers, the system aimed to enhance prediction accuracy and achieve superior results with high-quality predictions. The study utilized two distinct datasets, but it was constrained by a small dataset comprising only 247 instances and focused on predicting hypothyroidism disease.

In [6], N. Afshan et al. introduced a novel idea for diagnosing thyroid cancer that integrated an adaptive synthetic sampling technique with a weighted average voting ensemble of two different super learners (SLs). Although the ensemble model demonstrated enhanced performance, its computational complexity may increase due to the utilization of multiple base estimators and iterations of the super learner technique.

T Akhtar et al. [7] introduced an effective uniform ensemble of ensembles combined with diverse feature selection techniques to improve the accuracy of detecting thyroid disorders. The paper applied Select From Model (SFM), Select K-Best (SKB), and RFE strategies. The authors employed DT, Gradient Boosting (GB), LR, and RF classifiers as feature estimators. The approach included homogeneous ensembles based on bagging and boosting classifiers, integrated using a voting ensemble technique with soft and hard voting mechanisms. Despite achieving 100% accuracy with a compact feature set, the study was limited by a small dataset and did not address issues related to dataset imbalance.

In [8], L. Aversano et al. explored the prediction of thyroid LT4 treatment using ten different established machine-learning algorithms. The study encompassed multiple datasets, with findings indicating the Extra-Tree Classifier performed strongly, achieving an 84% accuracy rate. However, a notable drawback of the study was the dataset's quality. The absence of any feature selection technique was another drawback of the paper.

D. C. Yadav and S. Pal [9] proposed a decision tree-based bagging ensemble method for thyroid disease prediction. They utilized a dataset containing 3,710 cases and 29 features, employing decision tree (DT), RF, and extra tree (ET) classifiers for predictions. Despite achieving 100% accuracy with the bagging ensemble model, it is noted for its potential increase in computational complexity due to multiple iterations. Furthermore, the study did not address feature selection techniques or discuss the issue of data imbalance.

A. Selwal and I. Raoof [10] proposed a model for thyroid disease prediction using a MLP machine-learning approach. The study utilized a dataset comprising 120 individuals from SKIMS, Jammu, and Kashmir. The classification process involved using 7 to 11 features to classify individuals into normal, hyperthyroid, and hypothyroid categories. However, the authors did not extensively discuss their feature selection methodology. Despite working with a very small dataset, the model achieved an accuracy of 99.8%.

D. C. Yadav and S. Pal [11] proposed a voting-based ensemble method using decision trees for predicting thyroid disease. Their study utilized a dataset consisting of 499 thyroid patients and applied J48, Random Tree, and Hoeffding Tree classifiers to predict hyperthyroidism, hypothyroidism, and euthyroidism based on experimental values of T3, T4, and TSH. The ensemble model achieved an accuracy of 99.2% with a small dataset. However, the research did not address the issue of data imbalance and did not discuss appropriate feature selection techniques.

In contrast to existing studies, this paper introduces a stacking ensemble model aimed at enhancing the accuracy of thyroid prediction. Feature selection technique was utilized to extract pertinent features from a large set of attributes. To tackle the issue of imbalanced data, the paper applied the SMOTE-ENN hybrid sampling technique. Finally, evaluate and compare the result with the different traditional techniques to introduce the most effective model for thyroid prediction.

III. Methodology

In this section, Fig. 1 illustrates the overall workflow of the proposed model. The model proceeds through several stages: data collection, data preprocessing, feature selection, balancing data, design stacking model, and performance evaluation for predicting thyroid disease. For transparency of the model was used the SHAP technique. A detailed explanation of the methodology will follow in the subsequent section.

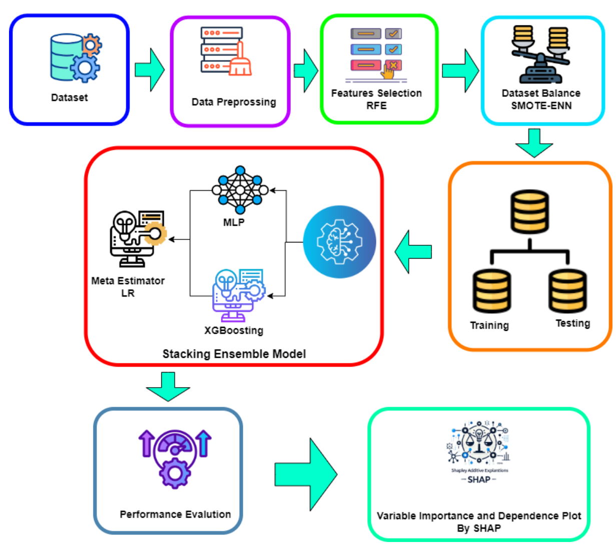


Fig. 1: Workflow Diagram of the Proposed Stacking Model

A. Data Collection

The dataset is the most important component of the ML and DL approaches. The most challenging task was to collect the correct dataset. So, it was more difficult to manage this dataset for our work. Our dataset was collected from Kaggle provided by the UCI Machine Learning Repository and uploaded by Emmanuel F.Werr. The dataset has a total number of 9172 data with 31 attributes of thyroid disease [12].

B. Data Pre-processing

Raw datasets have enormous problems such as missing data, null values, unimportant features, imbalance datasets, and others. To address this issues, the paper used various techniques like, data cleaning and handling, duplicate data removal, redundant feature cleaning, out layer removal, handling missing values, and filling missing values using KNN importer.

C. Feature Selection Technique

Feature selection involves automatically identifying the most effective features that contribute remarkably to predicting the target attributes. Some irrelevant features in the dataset effect the accuracy of the prediction, create overfitting issues, and increasing the processing time [13]. So, in this paper, the Recursive Feature Elimination technique has been used to overcome the issues.

Recursive Feature Elimination (REF): The REF technique is a method for selecting essential features for a model. It begins by training the model using all available features. After training, REF ranks these features based on their impact on model performance, often using coefficients or feature importance. It then removes the least significant features from the current set. REF operates recursively, iteratively refining the set of attributes by eliminating those deemed less crucial and building subsequent models with the remaining attributes [14]. In our model, REF has been used to overcome the overfitting problem and increase the accuracy by identifying the core features and removing the redundant features. Fig. 2 depicts the most significant features for this model such as, TSH, FTI, TT4, T3, 4U, on thyroxine, age, tumor, sex, psych, and query hyperthyroid.

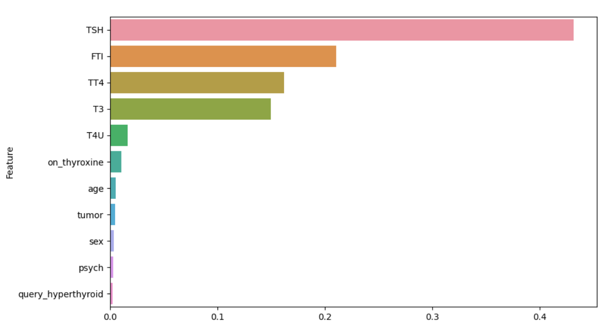
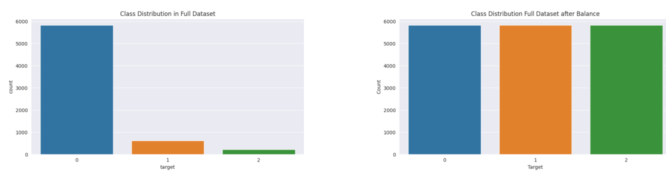


Fig. 2: Feature selection by REF

D. Imbalance Dataset

In datasets, unequal distribution of attributes often leads to data imbalance. Predicting accurately with imbalanced datasets poses a primary challenge. Effective data balancing is crucial for optimal model performance. Oversampling and undersampling are methods used to address this issue. In undersampling, removing samples from the majority class may eliminate valuable information that could potentially diminish classifier performance [15]. Conversely, the oversampling technique would introduce challenges such as overlapping, boundary issues, or noisy samples when generating new data. These issues have been tackled by the SMOTE-ENN hybrid technique. This approach aggregates SMOTE, which synthesizes minority class samples to balance the dataset, with ENN (Edited Nearest Neighbors), which removes noisy or borderline samples. By incorporating these methods, SMOTE-ENN aims to enhance dataset quality by reducing noise and improving class balance, thereby bolstering model performance on imbalanced datasets [15]. The imbalance and balanced dataset for the model is depicted in Fig.3 (a-b).



(a) Imbalance Dataset (b) Balanced Dataset

Fig. 3: Data Balancing with SMOTE-ENN hybrid technique

E. Classification Models

1) Multi-Layer Perceptron (MLP): An MLP is an artificial neural network defined by its layered architecture, comprising multiple interconnected layers of neurons. Neurons within each layer are connected to neurons in subsequent layers, and each connection is characterized by a weight that adjusts during the training phase [10]. The technique is extensively utilized for tasks, such as classification, regression, and function approximation because of their capacity to represent intricate correlations between input and output variables.

2) Extreme gradient boosting (XGBoost): XGBoost is a widely used ML algorithm that extends the principles of gradient boosting. Initially, XGBoost trains the model using a weak classifier. Subsequently, it fits a new weak classifier to the residuals of the current model, aiming to enhance accuracy without disrupting the existing model. This iterative process continues, refining the model to achieve better performance with each step [16]. The technique is mainly utilized for its speed and efficiency. Also, it handles large datasets and complex models effectively.

3) Logistic Regression (LR): LR is a statistical technique designed for binary classification and multi-class classification [17]. It is trained using optimization techniques to determine coefficients that most accurately fit the data by minimizing the disparity between predicted probabilities and observed outcomes. The main benefit of using LR is that it provides a probabilistic framework to make predictions.

F. Proposed Stacking Ensemble Model

A stacking model combines multiple base models using a meta-model. Base models are trained on the original data, and a meta-model is trained on the predictions made by the base models [17]. Stacking often combines multiple models resulting in better performance than any single base model. It also reduces the risk of overfitting because it exploits the strengths of various models.

Hence, MLP and XGBoost as base models, with LR as the meta-model have been selected in our proposed stacking model. MLP and XGBoost were selected as base models for the stacking ensemble because of their complementary strengths. MLP is particularly effective at modeling complex, non-linear relationships between input features, while XGBoost excels in preventing overfitting through regularization and is highly scalable for large datasets. Combining these models offers a balanced combination of flexibility and robustness, making them more suitable for enhancing predictive performance in the ensemble. The overall procedure of the proposed stacking model is presented in Algorithm 1.

Algorithm 1 Proposed Stacking Ensemble Model

Requirement: Training dataset (T) consists of input features and target features. **Base models:** $f_{\text{MLP}}(T)$ predicts using MLP model for input data T. and $f_{\text{XGB}}(T)$ predicts using XGBoost model for input data T. **Meta-model:** $f_{\text{LR}(T)}$ combines the predictions of base models.

Outcome: Thyroid disease prediction.

(1) Splitting the training dataset using k-fold cross-validation

(2) Train the base models.

- Train the MLP model $f_{\text{MLP}}(T)$ and the XGBoost model $f_{\text{XGB}}(T)$ on the training set.
- Generate the prediction of the base models from the validation set.

(3) Combine these predictions into a feature set for the meta-learner.

$$X_{\text{Meta}} = [f_{\text{MLP}}(T), f_{\text{XGB}}(T)]$$

(4) Train the meta-model $f_{\text{LR}(T)}$ using X_{Meta} as input features and target features from the validation set.

(5) Predict the target variable for the testing set using the trained meta-model to obtain the final ensemble prediction Y_{Stack}. Finally, evaluate the accuracy of the proposed model Y_{Stack}.

G. SHapley Additive exPlanations (SHAP)

SHAP is an effective technique for improving model interpretability by offering a detailed view of feature contributions. Each feature in the model is assigned a SHAP value that indicates its effect on the model's output. This process clarifies how each predictor influences specific predictions, enhancing transparency in the model [18]. Hence, the SHAP technique is applied in this work to identify the core features that significantly contribute to the model performance.

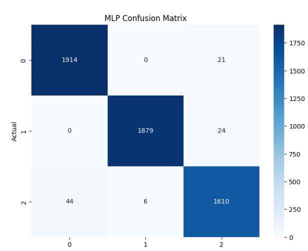
IV. Result and Discussion

The performance of each model to identify the most effective predictor is assessed in the section. The study evaluates model performance using metrics such as accuracy, recall, precision, and F1-score [19]. The results of the techniques are presented in Table I, and Fig. [4-5]. A comparison with existing research papers has been shown in Table II.

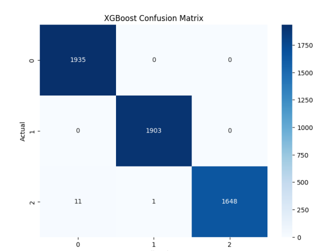
TABLE I: Performance Evaluation compared with our proposed model and other traditional models

Model	Accuracy	F1-score	Precision	Recall
Random Forest	98.18	98.18	98.19	98.18
ExtraTree	97.68	97.66	97.65	97.68
SVM	93.40	92.49	93.42	93.40
AdaBoost	94.81	95.54	96.28	94.81
DecisionTree	98.00	97.99	98.00	98.00
Our Proposed Model	**99.78**	**99.80**	**99.79**	**99.78**

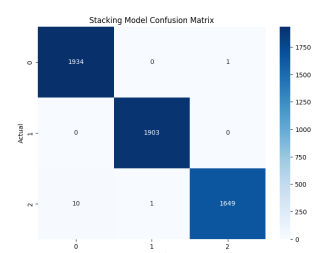
Table I presents a comparison of our proposed stacking model against several traditional models evaluated using four different metrics. The result showed that the Proposed Stacking Model performed the best, with 99.78% of accuracy and 99.80% of F1-score. In contrast, the accuracy of the Random Forest, Decision Tree Classifier, and Extra Trees Classifier, is 98.18%, 98.00%, and 97.68% respectively. SVM and AdaBoost Classifier have poor performance compared to others.



(a) MLP confusion Matrix



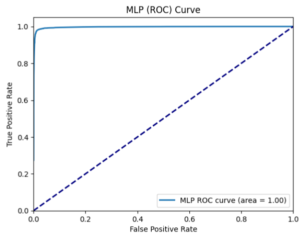
(b) XGBoost confusion Matrix



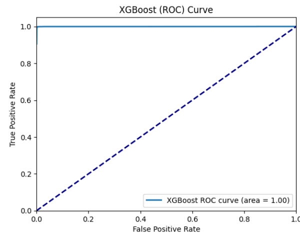
(c) Stacking Ensemble confusion Matrix

Fig. 4: Confusion Matrices of MLP, XGBoost, Stacking Ensemble Model

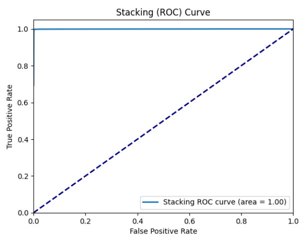
The confusion matrices for MLP, XGBoost, and the Ensemble Model indicate high accuracy (Fig. 4).



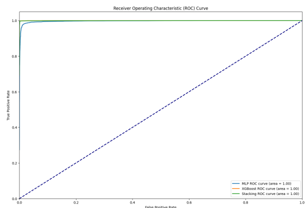
(a) MLP ROC Curve



(b) XGBoost ROC Curve

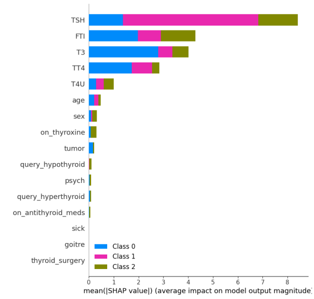


(c) Stacking ROC Curve

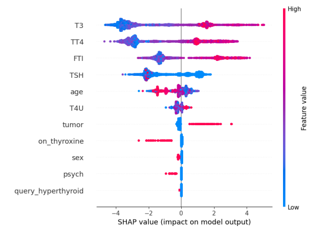


(d) All ROC Curve Combination

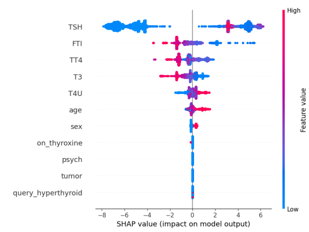
Fig. 5: ROC Curve of MLP, XGBoost, Stacking Ensemble Model



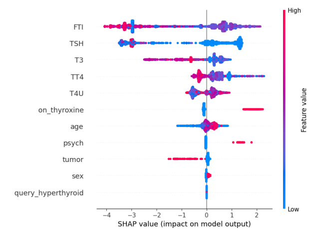
(a) Summary view of the contribution of each feature



(b) Summary plot of mean SHAP values for Class 0



(c) Summary plot of mean SHAP values for Class 1



(d) Summary plot of mean SHAP values for Class 2

Fig. 6: Interpreting the model with the SHAP values

The receiver operating characteristic curves (ROC Curves) show how well a classification model works across all levels. Fig. 5 depicts that the performance of the proposed Model is the best. The ROC curve indicates that the MLP and XGBoost models complemented each other effectively.

Table II represents an overview comparison of our proposed model with several recent research papers to show the effectiveness of our model. In [6], the authors used a weighted-based voting ensemble model in two different datasets and achieved the accuracy of 99.6% and 99.7% respectively. However, the computational complexity of the model is a limitation of this paper. In [9], the authors used the Bagging ensemble model in the UCI dataset and achieved 100% accuracy for 35 seed values. However, it is noted that the complexity of the model is potentially increased due to multiple iterations. Furthermore, the study did not mention the issue of data imbalance. In [20], the author used a self-stack ensemble with random forest and achieved 99.0% accuracy. One drawback of this study is that each target class was reduced to 230 data instances through the process of down-sampling. On the other hand, our proposed model used a stacking model with the SMOTE-ENN hybrid technique in the UCI dataset and achieved 99.78% accuracy. In summary, the stacking model adeptly integrated the strengths of XGBoost and MLP, utilizing their complementary features to enhance predictive accuracy through Logistic Regression. To verify the role of the features in thyroid prediction, the model generated SHAP values. Fig. 6 (a-d) illustrates the SHAP summary feature plots for the three classes and provides the feature importance and their effects on the model prediction.

For Class 0, the features T3, TT4, FTI, and TSH are the most impactful. Higher values of T3 and TT4. Also, lower values of FTI and TSH also influence this class. Other features like age, T4U, tumor, and on_thyroxine have small impacts with varying influence based on their values.
Class 1 is mostly influenced by TSH, where high values strongly contribute to predictions. FTI, TT4, T3, and T4U also play vital roles. Age and sex have small impacts, age showing a higher impact in older patients. On_thyroxine and query_hyperthyroid have smaller influences on predictions.
For Class 2, FTI, TSH, T3, and TT4 are the main influencing features. High values of FTI and T3 influence the model towards Class 2, while low values of TSH and TT4 are indicative of this class. On_thyroxine shows an impact, with higher values pushing towards Class 2. Age, psych, and tumor also have varying impacts in class 2.
Overall classes, TSH, FTI, T3, TT4, and T4U consistently appear as core features, although their directions and magnitudes of impact vary. Age shows variable importance, indicating its influence is context-dependent. On_thyroxine is particularly notable for Class 2. Other features tumor, sex, psych, and query_hyperthyroid have more specific and fewer impacts across the classes. Shap provides a clear understanding of the underlying patterns in the data. Finally, this approach increased prediction transparency and delivered meaningful insights into the complex patterns and relationships among features.

V. CONCLUSION

Our paper presents the development and evaluation of a stacking ensemble learning model for predicting thyroid disease. our proposed method combined the strengths of MLP and XGBoost, for the improvement of the accuracy. Additionally, our model effectively addresses the challenge of imbalanced datasets, which is a prevalent issue in medical diagnosis. The SMOTE-ENN technique ensures that our model achieves high

TABLE II: Comparison of our proposed model with different existing research papers

Ref. No	Dataset	Dataset Size	Models	Accuracy
[6]	Dataset 1: KEEL thyroid dataset Dataset 2: Hypothyroid UCI repository	Dataset 1: 5760 samples Dataset 2: 3017 samples	Weighted Voting Ensemble, LR, Decision Trees, SVM, RF, AdaBoost, Bagging Classifier	Dataset 1: 99.6% Dataset 2: 99.7%
[9]	UCI machine learning repository	3710 instances	Bagging Ensemble Model: RF, Decision Tree, Extra Tree	100% in case of seed value 35
[20]	UCI Machine Learning repository	9172 samples	Self-Stack Ensemble Model: Random Forest, down-sampling	99.0%
Our Proposed Model	**UCI Machine Learning repository**	**9172 samples**	**Stacking Ensemble: MLP, XGBoost, Logistic Regression**	**99.78%**

accuracy across all classes, including less common thyroid disorders. Moreover, the SHAP technique has been demonstrated to be an effective method for understanding feature relationships and their influence on thyroid predictions. This validation has offered valuable insights into the consistency and reliability of our model. Finally, the performance of our model has been outstanding, with an accuracy of 99.78%. In addition, Our proposed model significantly enhanced prediction accuracy and provided valuable insights into key features. This aligns seamlessly with the principles of Industry 5.0 by offering practical, high-impact solutions that advance healthcare technology and improve patient care outcomes. Nevertheless, our paper's limitation is that our model was used with a single dataset. In the future, we will focus on incorporating real-time data collected from various healthcare sources. Additionally, integrating advanced machine learning techniques, such as deep learning and reinforcement learning could enhance predictive capabilities. Moreover, developing personalized medical approaches will enable precise diagnosis and treatment recommendations.

VI. ACKNOWLEDGMENT

We sincerely appreciate to the Center for Research Innovation and Transformation (CRIT) at Green University of Bangladesh for their financial support in conducting this research.

REFERENCES

[1] D. Childs, "Thyroid Disorder," Available online: https://www.restartmed.com/hypothyroidism-statistics/ (Accessed on June 11, 2024).

[2] S. Selim, M. Mustari, T. A. Khan, "Approach to Management of Hypo and Hyperthyroidism in Bangladesh: A Nationwide Physicians' Perspective Survey", Frontiers in Endocrinology, Vol. 14, pp. 1322335, 2023.

[3] "Hypothyroidism", National Institute of Diabetes and Digestive and Kidney Diseases (NIDDK), Available online: https://www.niddk.nih.gov/health-information/endocrine-diseases/hypothyroidism (Accessed on June 24, 2024).

[4] Mayo Clinic. Hyperthyroidism: Symptoms and causes. Retrieved June 24, 2024, Available online:https://www.mayoclinic.org/diseases-conditions/hyperthyroidism/symptoms-causes/syc-20373659.

[5] M. L. M. Prasad, and R. Santhosh, "Design of a Multi-Stage Ensemble Model for Thyroid Prediction Using Learning Approaches", Intelligent Automation & Soft Computing, vol. 39, 2024.

[6] N. Afshan, Z. Mushtaq, F. S. Alamri, M. F. Qureshi, N. A. Khan, and I. Siddique, "Efficient Thyroid Disorder Identification with Weighted Voting Ensemble of Super Learners by Using Adaptive Synthetic Sampling Technique", AIMS Mathematics, vol. 8, pp.24274-24309, 2023.

[7] T. Akhtar, S.O. Gilani, Z. Mushtaq, S. Arif, M. Jamil, Y. Ayaz, S.I. Butt, A. Waris," Effective Voting Ensemble of Homogenous Ensembling with Multiple Attribute-Selection Approaches for Improved Identification of Thyroid Disorder", Electronics, vol. 10, PP.3026, 2021.

[8] L. Aversanoa, M. L. Bernardia, M. Cimitileb, M. Iammarinoa, P. E. Macchiac, I. C. Nettorec, C.Verdone, "Thyroid Disease Treatment Prediction with Machine Learning Approaches", Procedia Computer Science, Elsevier, vol.192, pp. 1031—1040, 2021

[9] D. C. Yadav, and S. Pal,"Prediction of Thyroid Disease Using Decision Tree Ensemble Method", Human-Intelligent Systems Integration, vol. 2, pp. 89-95, 2020.

[10] A. Selwal, and I. Raoof, "A Multi-layer Perceptron Based Intelligent Thyroid Disease Prediction System", Indonesian Journal of Electrical Engineering and Computer Science, vol. 17, pp. 524-532, 2020.

[11] D. C. Yadav, and S. Pal, "Decision Tree Ensemble Techniques to Predict Thyroid Disease", International Journal of Recent Technology and Engineering, vol. 8, pp. 8242-8246, 2019.

[12] N. E. Fwerr. "Thyroid Disease Data." Kaggle. Accessed July 7, 2024. https://www.kaggle.com/datasets/emmanuelfwerr/thyroid-disease-data/data.

[13] M. Riajuliislam, K. Z. Rahim, and A. Mahmud, "Prediction of Thyroid Disease (Hypothyroid) in Early Stage Using Feature Selection and Classification Techniques", International conference on information and communication technology for sustainable development (ICICT4SD), IEEE, pp. 60-64, 2021.

[14] A. Sultana, and R. Islam, "Machine Learning Framework with Feature Selection Approaches for Thyroid Disease Classification and Associated Risk Factors Identification", Journal of Electrical Systems and Information Technology, vol. 10. p.32, 2023.

[15] Z. Xu, D. Shen, T. Nie, and Y. Kou, "A Hybrid Sampling Algorithm Combining M-SMOTE and ENN Based on Random Forest for Medical Imbalanced Data", Journal of Biomedical Informatics, vol. 107, p.103465, 2020.

[16] S. Sankar, A. Potti, G.N. Chandrika, and S. Ramasubbareddy, "Thyroid Disease Prediction Using XGBoost Algorithms", Journal of Mobile Multimedia, vol. 18, pp.1-18, 2022.

[17] G. Obaido et al., "An Improved Framework for Detecting Thyroid Disease using Filter-based Feature Selection and Stacking Ensemble", IEEE Access, 2024.

[18] S. Sankar, and S. Sathyalakshmi, "A Study on the Explainability of Thyroid Cancer Prediction: SHAP Values and Association-Rule Based Feature Integration Framework", Computers, Materials & Continua, vol. 79, 2024.

[19] M.A. Latif, Z. Mushtaq, S. Arif, S. Rehman, M.F. Qureshi, N.A. Samee, M. Alabdulhafith, Y.H. Gu, and M.A. Al-masni, "Improving Thyroid Disorder Diagnosis via Ensemble Stacking and Bidirectional Feature Selection", Computers, Materials & Continua, vol. 78, 2024.

[20] S. Ji, "SSC: The Novel Self-Stack Ensemble Model for Thyroid Disease Prediction", Plos one, vol. 19, p.e0295501, 2024.