Volume 3, Issue 9, 2025

Frontier in Medical & Health Research

ISSN: (e) 3007-1607 (p) 3007-1593

# ENSEMBLE LEARNING AND UNCERTAINTY ESTIMATION FOR PREDICTIVE MODELING OF HYPOTHYROIDISM

**Zill e Huma[1], Muazzam Ali[*2], M U Hashmi[3], Affan Ahmad[4], Abdul Manan[5]**

[1]*Department of Basic Sciences, Superior University, Lahore, Pakistan*
[*2,3,4,5]*Department of Computer Science, Superior University, Lahore, Pakistan*

[*2]muazzamali@superior.edu.pk

DOI:https://doi.org/10.5281/zenodo.17777113

**Keywords**
*Hypothyroidism Detection, Ensemble Machine Learning, SMOTE, Principal Component Analysis (PCA), Uncertainty Estimation*

**Article History**
*Received: 20 September 2025*
*Accepted: 15 November 2025*
*Published: 29 November 2025*



**Corresponding Author: ***
**Muazzam Ali**

## Abstract

*This study highlights the development of prediction of predictive model for predicting the hypothyroidism by leveraging the ensemble machine learning techniques that are combined with ensemble techniques The issues in associated with imbalanced and noisy datasets the study evaluated the use of SMOTE for balancing the class and PCA to reduce the dimensions. The models evaluated included AdaBoost, Bagging (Random Forest) and Blending (Stacking). With the focus on abilities to handle the complex medical data and increase the accuracy of prediction. The research highlights effectiveness of ensemble techniques methods in enhancing the reliability and accuracy of hypothyroidism prediction. Key findings shows that Blending (Stacking) shows best balance between precision, recall that provides superior performance as compared to others even when PCA and SMOTE are used. Moreover, the importance of uncertainty quantification in refining the predictive abilities is focused that evaluated its potential in high – stakes medical applications. The applications of PCA and SMOTE while beneficial in increasing the model performance also introduced some complexity and potential for over fitting.*

## 1. Introduction

When thyroid gland doesn't produce or release enough hormones into your bloodstream, you get hypothyroidism [1]. Metabolism consequently slows down. This may result in inadvertent weight gain and persistent fatigue. Doctor may check for hypothyroidism with a simple blood test, even if symptoms like weariness and weight gain aren't unique to the illness. Hypothyroidism is generally relatively curable [2]. With medicine and frequent follow-up appointments with an endocrinologist, the majority of people can manage the disease. Long-term hypothyroidism can prove fatal if treatment is not received [3]. It's crucial to get evaluated if patient start experiencing new symptoms. There are four main types of hypothyroidism. When primary hypothyroidism is present, the thyroid generates insufficient levels of thyroid hormones [4]. TSH, or thyroid stimulating hormone, is then produced in greater amounts by the pituitary gland. It can be caused by autoimmune disorders including Hashimoto's disease, radiation therapy, or thyroid gland surgery. Primary hypothyroidism is the most common type. Secondary hypothyroidism is caused by an underactive pituitary gland, a pea-sized structure at the base of the brain [5]. This rare type of hypothyroidism occurs when the pituitary gland is unable to send TSH to the thyroid gland. Tertiary hypothyroidism is caused by the hypothalamus, a region of the brain that regulates homeostasis, producing too little thyrotrophic-releasing hormone (TRH) [6]. As a result, the pituitary

Frontier in Medical & Health Research

ISSN: (e) 3007-1607 (p) 3007-1593

gland cannot generate enough TSH. When TSH levels are slightly elevated but all other thyroid hormone levels are within the normal range, this is known as subclinical hypothyroidism, or mild thyroid failure. Subclinical hypothyroidism typically goes away on its own in three months. Hypothyroidism symptoms can develop gradually over time. In certain cases, it might take years. Brain fog, which is the inability to focus or remember things, is one of the possible symptoms, symptoms such as sadness and anxiety, dry, coarse skin and hair, high blood cholesterol, and fatigue [7]. Menstrual periods that are heavy or frequent, a hoarse voice, an inability to withstand cold temperatures, tingling or numb hands, facial changes such puffiness around the eyes and drooping eyelids, tired or weak muscles, and unexplained weight gain [8].

Several investigations have been carried out in the past that identify specific symptoms of thyroid problems. These symptoms are indicative of a specific thyroid disease [9]. For example, thyroiditis, Graves' disease, Hashimoto's disease, thyroid goiter, thyroid cancer, nodules, etc., can all be detected and predicted using these tests and symptoms. Thyroid disease may be categorized using a variety of indicators and traits. Current research provides several features relevant to thyroid problems. For example, thyroid detection may be done utilizing lithium, goiter, hypopituitary, Psych, TSH, T3, T4, T4U, FTI, and TBG. Numerous studies have already examined thyroid conditions and their symptoms [10-12]. Some approaches focus on data analytics, while others document tests and symptoms. Depending on the disease detection technique statistical analysis, machine learning, or deep learning the precision and robustness of these approaches vary. Most existing approaches involve overfitting models and publicly available datasets. There are very few samples for the positive (disease) class in most of the datasets that are currently available due to an imbalanced class problem. Such a dataset causes a machine learning or deep learning model to overtake the majority class and produce erroneous predictions for the minority class. Another limitation of existing research is that only a limited number of thyroid illnesses are used for classification; much of it focuses on the binary-class problem, making those approaches unsuitable for real-world sickness detection. This work aims to solve this issue by increasing the number of minority class samples through the use of synthetic data samples.

The aim of this research is to develop a predictive model for the detection of hypothyroidism by incorporating the ensemble techniques and uncertainty estimation. The key objective is to increase the accuracy and reliability of prediction in the context of imbalanced and uncertain dataset. This study use the ensemble techniques like AdaBoost, bagging and stacking with uncertainty estimation techniques to enhance the decision process. The research aims to address the common issues in medical prediction of tasks and imbalanced and noise by using the innovative machine learning methods like SMOTE for class balance and principal component analysis for dimensionally reduction. Furthermore, the study explains the substantial development in model performance with accuracy and focuses on the importance of uncertainty estimating in refining the predictive capabilities. The evaluation metrics provide the evaluation of ensemble techniques in handling complex medical data. This study brings forth new insights combining several approaches that justify the prolix model to hypothyroidism, and which are of great importance in medicine.

**Literature Review**

This study aimed at determining the most efficient AI model to predict the states of the thyroid gland while considering the region's liver enzymes and hormones as independent factors. The multi-layer perception MLP model recorded gains of 3.77% and 12.54% when set against the support vector SVM and Hammerstein-Weiner HW models, respectively. In addition to the previously mentioned individual AI models, three more ensemble methods were tested neural network ensemble, weighted average ensemble, and simple average ensemble. While performing the NNE technique, testing all three AI models, Usman and coworkers reported increases in

Frontier in Medical & Health Research

ISSN: (e) 3007-1607 (p) 3007-1593

performance accuracy of 7.44%, 11.212%, and 19.98% respectively [13]. In [14], they made a prediction of thyroid cancer using the modified xgboost model with clinical and molecular features. This algorithm enhances cancer prediction of thyroid nodules through integrating gene expression and fine needle aspiration biopsy. This model outperforms other machine learning models such as C5.0, CART, CHAID, QUEST, LSVM, random tree, and some other models. It is very surprising to see that this model works well with an astonishing accuracy of 94.6 percent! Its effectiveness is demonstrated in real time. The research also highlights the use of machine learning in healthcare predictions, along with the associated security challenges.

The model should be enhanced with accurate information on patients, their lifestyle, environmental elements, genetic data, and bias management. Thyroid disease prevention and prediction can be done internally by clinicians with the assistance of XGBoost. This research proposes an algorithm that uses bagging-based ensemble linear discriminant analysis with SMOTE for diagnosing thyroid disorders. The algorithm proposed incorporates four components: Developing a LDA classifier, applying SMOTE to each instance, "with replacement" random sampling, and classification voting of all trained classifiers. The subject of the investigation is hyperthyroidism and hypothyroidism. It was verified that the model is suitable for a core dataset of 1092 records and a secondary set of 7200 records. It was determined that for primary data, the accuracy was 85.45%, and for secondary data, it was 82.71%. The model was also tested against conventional machine learning classifiers for design accuracy assessment. This research concludes that the proposed approach successfully predicts thyroid conditions [15].

An Iraqi study focused on patients extracted data from Iraqi patients and applied an assortment of algorithms: SVM, RF, DT, NB, LR, KNN, MLP, and LDA to classify thyroid disease. The objective of the studies was to find out the most accurate and efficient algorithm which showed promise for the application of machine learning methods in the diagnosis of thyroid disorders [16]. To increased characterization accuracy, super learners were created in a way that they had multiple base and meta learners and resampling methods for changing the dataset. Evaluations on the UCI hypothyroid and thyroid disease datasets showed significant improvements in F1-score, recall, accuracy, and precision when compared to individual models. However, issues with the computational complexity of the WAV collection and the optimal selection of base assessors inside SLs still need to be investigated [17].

In [18], researchers aim was to improve the pycaret classification model's estimator list in order to more accurately forecast hypothyroidism. An intelligent feature classification model with blunge calibration was put out, supporting the very accurate diagnosis of hypothyroidism. Python was used to implement the model in the Anaconda Navigator IDE's Spyder editor. The model maintained an accuracy of 99.5% for the BCRM and BCCM models and 89.5% for regressor feature-selection techniques, according to the data [19]. By combining many basic models, the framework enhances prediction performance while cutting expenses and screening time. With ROC-AUC scores of 99.9%, tests on a dataset of thyroid diseases revealed that the ensemble learning method and filter-based feature selection strategy had superior discriminative performance. This work demonstrates how ensemble techniques might enhance machine learning-based thyroid illness identification. Despite its potential, the research is biased due to the use of secondary data and thus needs improvement. Using varied datasets could better predict the gravity of thyroid disorders [20]. After evaluating four machine learning ensemble algorithms, the Bagging Classifier stood out as the most effective method for improving the accuracy of identifying complex thyroid disorders. With 99.87 accuracy, it surpassed AdaBoost that had 98.87. The results have the potential to facilitate the development of automated tools for the early detection and monitoring of thyroid disorders which will ultimately improve patient care and accelerate medical processes [21]. Whenever an estimate of

Frontier in Medical & Health Research

ISSN: (e) 3007-1607 (p) 3007-1593

the possibility of a condition is to be given, the traditional expert systems TES in medicine often tend to resort to a certainty factor CF approach. This method, however, is not always capable of estimating vague issues such as the likelihood of the disease. A supervised learning approach was taken to solve this problem using linear regression. The research placed emphasis on so-called analytic methods of a machine learning expert system MLES, and, as it turned out, the models which were strongest in forecast Thyroid abnormalities illness risk were multiple linear regression MLR and multiple polynomial regression MPR [22].

## 3. Research Methodology

The text above means that there were several pre-processing techniques used during the research to ensure that the data was clean and ready to be used by machine learning algorithms. The preprocessing steps are broken down and explained below:

### (a) Missing Values

Not having complete datasets is one of the most common issues which leads to biased or incomplete models. Different strategies were employed to overcome the issue of missing values depending on the type of feature present. For missing categorical columns, the missing values were filled with the modes of that feature. For categorical items, that value is assumed to be its mode. The median imputation method is used for categorical columns to fill missing values, ensuring data skewness. However, it may not be ideal for extremely skewed data or poor estimation models.

### (b) Class Imbalance

Though class imbalance is rarely handled while building models, it is a challenge when dealing with classification problems. In metrics with class imbalance, F1-score, precision, recall, and AUC-ROC are more useful than accuracy [23]. By considering all the metrics in the evaluation matrices, it can be hypothesized that there is class imbalance in the dataset but that would require further investigation to confirm the hypothesis.

### (c) Feature Selection

The Principal Component Analysis (PCA) was utilized for feature selection, ensuring high variance containment and managing data overfitting issues by capturing essential features.

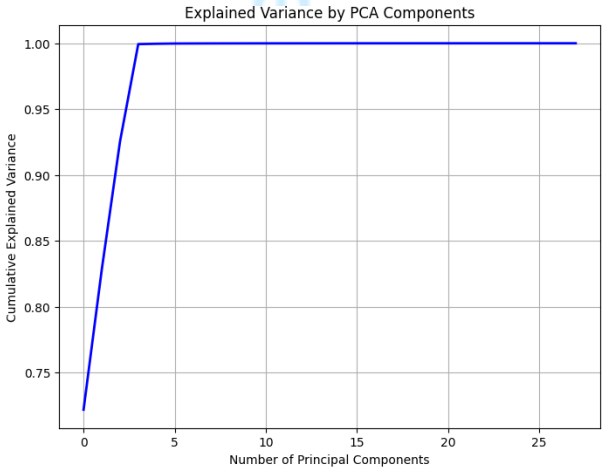



**Figure 1. Explained Variance by PCA Components**

### (d) Data Transformation

For training non-numerical data on a Machine Learning model, it was necessary for the data to go through transformation encoding [24]. This processing step was divided into the following parts: Categorical types of variables like 'sex', 'queryonthyroxine', 'referralsource' were numerically encoded through Label Encoding to transform the features into numbers. The approach assigns categories to unique integers for

Frontier in Medical & Health Research

ISSN: (e) 3007-1607 (p) 3007-1593

easy analysis but is insufficient for nominal features like 'referral source', requiring tremble one-hot encoding. For the target variable 'binaryClass', Label Encoding was used to change binary-like categorical values (f' and 't') into numeric values (0 and 1). This would make the target variable suitable for binary classification.

**(e) Splitting Data into Training and Testing**

The data frame was split into testing and training portions according to a ratio of 70%-30%. Checking model accuracy is done using the test data which comprises 30% of the total records while the remaining 70% is used to train. The split was executed with the train_test_split method from sklearn to random shuffle the data for improved model generalization and performance evaluation.

**(f) Data Visualization**

**Histogram**

The histogram visualizes the distribution of different numerical features in dataset like Age, TSH, T3, T4, T$U and FTI, the analysis provides insights. Features like TSH and T3 provides left skewed distribution where most values are concentrated at the lower end with few extreme outliers. This shows that mostly of individuals have lower levels of these features with small group shown extreme values Features like Age and T4 also shows skewness and potential outlier which may affect model performance and necessitate further data preprocessing or investigation. FTI and T4U show bimodal distribution that shows the presence of two distinct groups within in the data. This information is valuable for understanding features behavior which could influence predictive modeling and approach to handling the features [25].

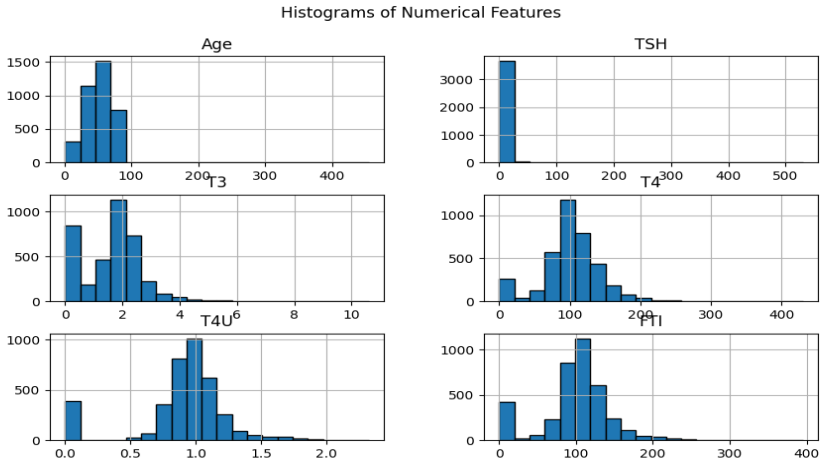



**Figure 2. Histogram of Dataset**

**Box Plot**

The boxplot depicts the numerical data distribution visually using the minimum, Q1, median, Q3, and maximum. Also, it points out some outliers. For instance, in TSH, T4, and T3 there are several outliers which are indicated by points outside the whiskers of the boxplot. These outliers could either be valid significant values or mistakes which need to be further analyzed and could be treated specifically in the model (for example, capped or removed). The range on T4 and FTI shows these features have large amounts of variance which is important to assess how these features can impact the model. It suggests transforming features to improve model log transformation fit, as TSH's low-range values and disproportionate high values indicate skewness.

Volume 3, Issue 9, 2025

Frontier in Medical & Health Research

ISSN: (e) 3007-1607 (p) 3007-1593

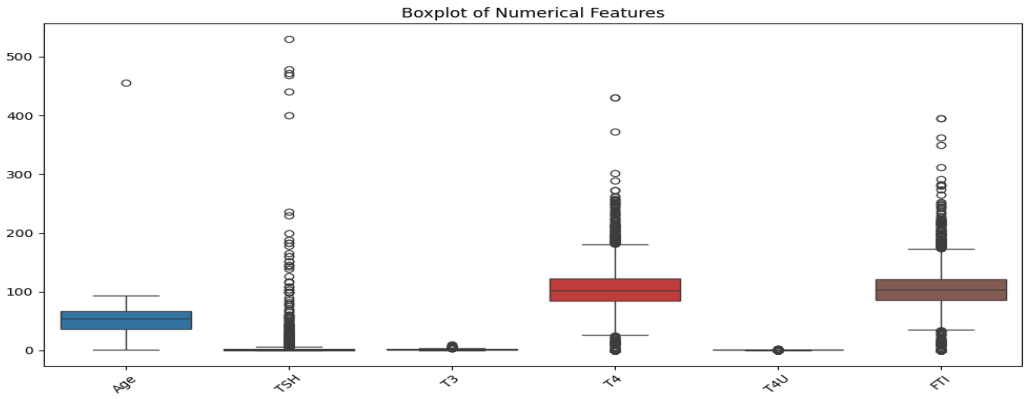



Figure 3: Boxplot of Dataset

**Correlation Heat Map**

A correlation heat map depicts the parallel association between numerical values. With color gradients, it shows the strength and direction of the correlation, red shows strong positive correlations and blue shows negative ones [26]. The 0.78 correlation between T4 and FTI indicates a rather strong association meaning that these two features most probably have similar underlying factors, or one might be influencing the other. One needs to take this into account in the model to mitigate multi-collinearity. Features like Age, TSH and T3 display weak or negative associations with most of the remaining features which indicate that they are peripheral from the rest of the thyroid-dominant features within the other parts of the dataset. Selecting features with high correlation and dimensionality reduction is crucial for model performance. Combining T4 and FTI into informative components through PCA may improve model performance.

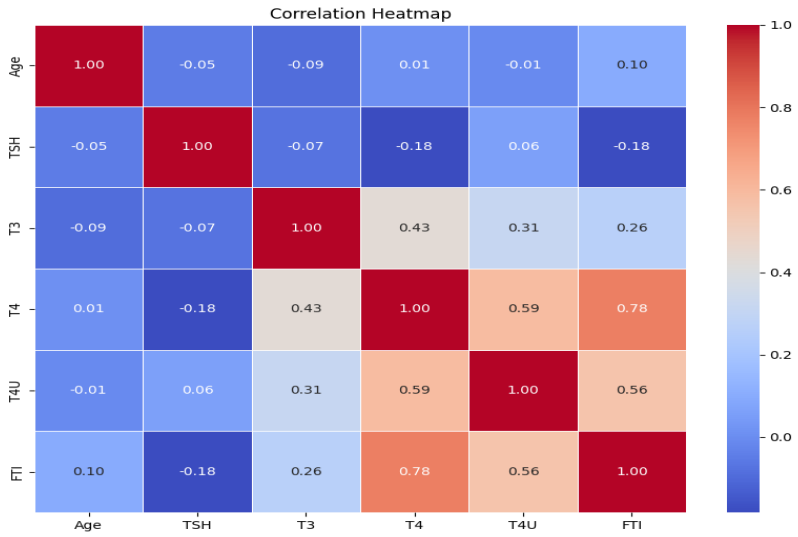



Figure 4: Correlation Heatmap of Dataset

Volume 3, Issue 9, 2025

Frontier in Medical & Health Research

ISSN: (e) 3007-1607 (p) 3007-1593

**CountPlot**

Count plots show a class imbalance in categorical attributes like gender, antithyroid medication use, and pregnancy, indicating a need for more accurate data analysis [27]. For instance, suppose we look at the patients who are not pregnant, there are more cases than those who are pregnant. This might bias the predictions that can be made by the model if some measures like oversampling, under sampling, or using SMOTE synthetic sample generation techniques are not employed. Some features like lithium, goiter, and tumor are not common in some categories and may inhibit the model's generalization capabilities. These low frequency categories may be handled differently (for instance, merging small sized classes or specific targeting algorithms designed to work with imbalanced data). The binary Class variable represents the target variable, assuming positive or negative hypothyroidism cases, crucial for understanding class composition and imbalance in the dataset.

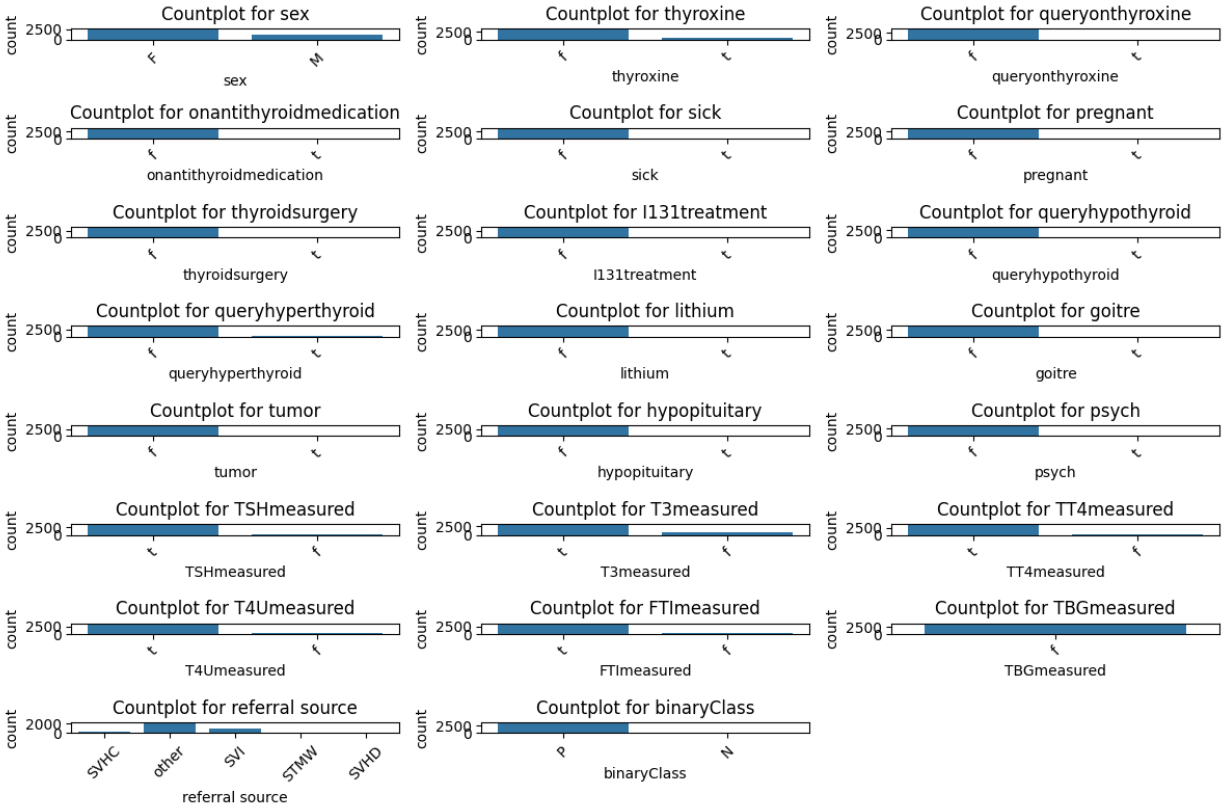



**Figure 5: Countplot for Dataset**

**Overall Critique of Visualizations**

Such visualizations provide important information regarding the presence of distributions, interrelations between various features, and other potential issues such as outliers and class imbalance. Grasping the data's structure allows for decisions to be made concerning data preprocessing, which includes outlier removal, skewed data transformations, feature selection, and handling imbalanced class distributions. Without such preprocessing work, machine learning models can be trained on imprecise data and result in highly variable predictions, thus skewing the outcome.

**(g) Ensemble Techniques**

Multiple base models' strengths were utilized in various ensemble techniques to improve performance. For this purpose, the following ensemble techniques were used:

Frontier in Medical & Health Research

ISSN: (e) 3007-1607 (p) 3007-1593

**AdaBoost**

An iterative process of weak classification improvement was executed using AdaBoost or Better put, boosting. Each weak 'learner' is trained in a sequential manner. In simpler terms, multiple decision trees are used as learners in this example. Each next model in this sequence attempts to fix mistakes done by the previous model. This method is very useful because it assists in bias reduction when working with noisy datasets.

**Bagging (Bootstrap Aggregating)**

Variance was reduced to enhance stability in a model using multiple base learners together, and thus bootstrapping, Bagging or Bootstrap Aggregating was employed. The model first generates distinct bootstrapped subsets of the training data to train various base learners like decision trees. When the model attempts to make a final prediction, it combines the outcome of all other models through voting or averaging. For tree models such decision trees, which are high variance models, Bagging greatly assists in reducing overfitting.

**Stacking**

Blending is an ensemble approach that works by combining different types of models into a single stronger model. For this project, Support Vector Machine (SVM), Random Forest, and Decision Trees are all used as base models in the blending process. A Logistic Regression model, which relies on the predictions of the base models to make the final prediction, combines these base models. Furthermore, blending excels at combining small, focused models that are trained on the different features of the data.

**(h) Evaluation Metrics**

The classification performance of the models was evaluated by using a set of metrics that were crafted to thoroughly capture their classification performance.

**Accuracy**

In classification tasks, accuracy is the most popular metric, which in this context denotes the ratio of correctly predicted instances to the total number of instances. It can be deceptive in cases of lopsidedness in datasets, in which a model can be accurate by simply predicting the majority class.

**Precision**

Precision and recall are vital with datasets that exhibit imbalance. While precision focuses on determining the true positive rate from all positive predictions done precision, recall is concerned with how well a model does in predicting the actual positives and focuses on true positive claims only. These metrics are very crucial in situations where, misclassifying a positive as a negative and vice versa is very critical (i.e. medical diagnosis).

**F1score**

The F1 score combines both the precision and the recall (true positive rate) using geometric distribution. It is a balanced metric normally adopted in situations where uneven class distribution is prevalent because unlike other metrics such as accuracy or precision, the single measure F1 score does not strongly rely on one aspect of the distribution of data. The F1 score helps ensure that both precision and recall are considered together.

**Cohen's Kappa**

Cohen's kappa captures the match between the anticipated selection and the actual selection made while considering random turning points. A Kappa score that is considerably high suggests that the model predictions concords with the true labels and is not by chance alone.

**Confusion Matrix**

The confusion matrix shows model performance on a finer scale by providing counts of true positives, true negatives, false positives, and false negatives. This is especially important in understanding what kinds of mistakes the model is making and the balance between various classes.

Frontier in Medical & Health Research

ISSN: (e) 3007-1607 (p) 3007-1593

**ROC Curve and AUC**

The Receiver Operating Characteristic (ROC) curve shows the ratio of true positives to false positives at different thresholds. The AUC gives us a measure of how well the model can discriminate between the two classes pos and neg. The larger the AUC, the better the model does, and the ROC curve provides an straightforward way to see how well the model performs at different levels of classification.

**Aleatoric Uncertainty**

Aleatoric uncertainty lies within the model as the degree of noise in the dataset sources out-of-sample models. This could be for example due to chance variation, alteration, or error during the measurement. While considering selection bias, one can observe aleatoric uncertainty in essences related to unresolved factors by assessing the precision of the model's multiple predictions. Moreover, concentration level reveals predicted noise and uncertainty factors beneath reliable estimation boundaries.

**Epistemic Uncertainty**

Declare the epistemic model that best suits your environment and execute relevant information in its meta. Epistemic uncertainty is caused when a model has insufficient data due to unlearned knowledge or incompletely trained model. It shows how confident a model can be in assuming it has corrected prediction made for a given task. The area selected for deletion highlights unnecessary borders that contain the wires sent on the base meta object. This fusion mostly gives more composite Metas that would perfectly work with the relevant data, if they had more advanced relevant data.

**Predictive Uncertainty**

Predictive uncertainty incorporates the two types of uncertainties- epistemic i.e. associated with a lack of knowledge and aleatoric which is randomness. Predictive uncertainty serves to indicate how confident one can be about the overall outcome of a model's prediction. As stated, it can indicate predictive reliability for decision making for various fields including medicine and finance. By calculating predictive uncertainty, it is possible to manage areas of low confidence, such as by spending more additional effort, more models, etc.

**Results and Discussions**

Following table 1 shows the results of the evaluation metrics which we computed using the ensemble techniques.

**Table 1. Results of Evaluation Metrics**

| Model | Accuracy | Precision | Recall | F1 Score | Cohen's Kappa |
|---|---|---|---|---|---|
| Boosting (AdaBoost) | 0.994695 | 1 | 0.99424 | 0.9971 | 0.964152 |
| Bagging (Random Forest) | 0.99646 | 0.999039 | 0.99712 | 0.9980 | 0.975609 |
| Blending (Stacking) | 0.997347 | 0.998084 | 0.99904 | 0.9985 | 0.98142 |
| Results with PCA | | | | | |
| Boosting (AdaBoost) | 0.954907 | 0.95841 | 0.994247 | 0.976 | 0.605797 |
| Bagging (Random Forest) | 0.965517 | 0.973585 | 0.989454 | 0.981455 | 0.7364 |
| Blending (Stacking) | 0.967286 | 0.972744 | 0.99233 | 0.982439 | 0.744075 |
| Results with SMOTE | | | | | |
| Boosting (AdaBoost) | 0.95137 | 0.985265 | 0.961649 | 0.973314 | 0.700134 |
| Bagging (Random Forest) | 0.966401 | 0.989289 | 0.974113 | 0.981643 | 0.783865 |
| Blending (Stacking) | 0.971706 | 0.988406 | 0.980825 | 0.984601 | 0.810719 |

Frontier in Medical & Health Research

ISSN: (e) 3007-1607 (p) 3007-1593

**(i) Accuracy**

The accuracy metric measures the proportion of correct prediction by the models. The highest accuracy of ensemble technique is observed of the Blending (Stacking) which is 0.9973 without PCA and SMOTE. The bagging (random forest) model perform well with accuracy 0.9965 but boosting (AdaBoost) slightly low in accuracy. While blending appears to be the best in term of accuracy that shows its importance that accuracy is not reliable metric for the context of imbalance class. This paper focuses on the F1score, precision and recall in addition to accuracy which is more reliable informative metrics for the imbalanced dataset. The little variation in accuracy shows that model may be influenced by the complexity of ensemble techniques. Blending (stacking) benefits of combining many models that capture more diverse patterns in data that may provides its superior performance. The accuracy drops in AMOTE and PCA for all models that show that transformation could change distribution of the data that impact the ability of model to produce as effectively as in baseline configuration.

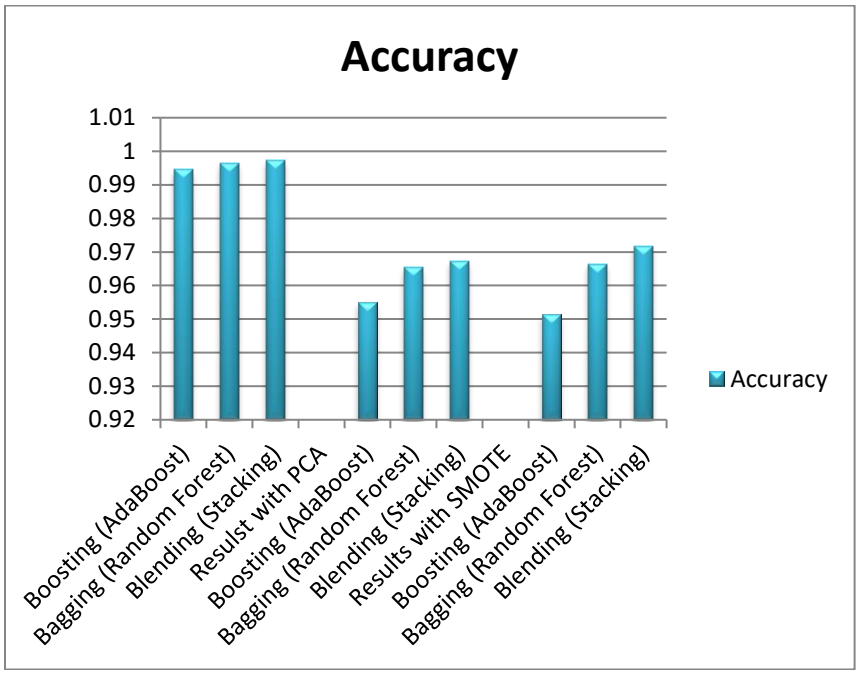



**Figure 6. Accuracy Comparison of the Ensemble Techniques**

**(ii) Precision**

The boosting (AdaBoost) model reaches the perfect precision score of 1. That shows that all positive prediction made by this model was accurate. While Boosting (AdaBoost) perform well in term of precision, Bagging (Random Forest) and Blending (Stacking) also shows high precision with 0.9990 and 0.9980 respectively. This shows that ensemble techniques decrease the false positive rate which is very important in medical imaging especially when predicting a serious condition such as hypothyroidism. The drop in precision with PCA and SMOTE is less pronounced that shows that dimensionally reduction and data balancing techniques might slightly affect the ability of model to avoid false positive rate. The Blending (Stacking) models still maintain the high accuracy.

Volume 3, Issue 9, 2025

Frontier in Medical & Health Research

ISSN: (e) 3007-1607 (p) 3007-1593

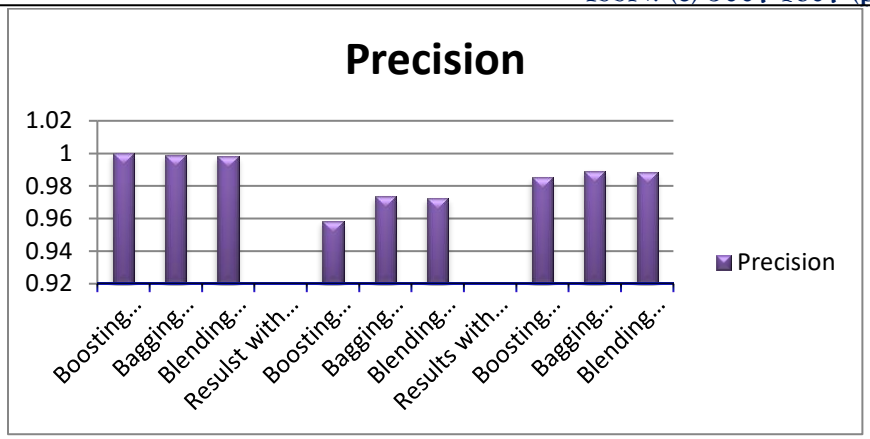



**Figure 7. Precision Comparison of the Ensemble Techniques**

**(iii) Recall**

Recall measures the proportion of true positive identities out of all actual positive cases. For blending (stacking), the value of recall 0.9990 which is highest among all and next value is 0.9971 of Bagging (random forest). These models show that they are adept at predicting nearly all the positive cases in the dataset. Boosting (AdaBoost) shows lower recall value that is 0.9942 as compare to others which shows that it might miss few positive cases. Its precision score is best that measure all correctly identified all positive prediction as true that suggest a trade-off between recall and precision in certain models. For all models the PCA and SMOTE decreases the recall. This decline may be due to alteration of features space during PCA which could obscure some important patterns that are required for detecting positive cases. Blending (Stacking) remains quite strong with 0.9923 with PCA or SMOTE that shows its superior ability to balance the sensitivity within the transformation.

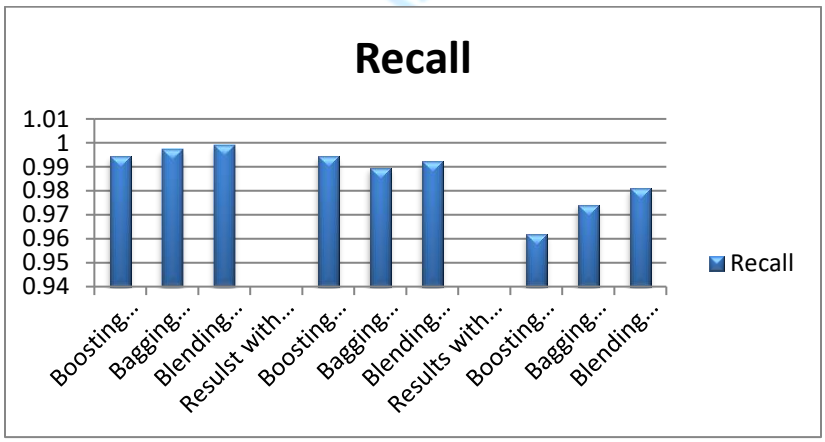



**Figure 8. Recall Comparison of the Ensemble Techniques**

**(IV) F1-Score**

Blending (Stacking) again excel with F1Score of 0.9986 without PCA or SMOTE which shows that it is most balanced model in term of precision and recall. Both Bagging (Random Forest) and Boosting (AdaBoost) perform excellently with F1-score of 0.9981 and 0.9971 respectively. F=The F1-score dos not show deterioration when using PCA and SMOTE with Blending (Stacking) continue to achieve 0.9824 value which is still high. This is important result as its shows that while the features engineering

Frontier in Medical & Health Research

ISSN: (e) 3007-1607 (p) 3007-1593

and class balancing techniques may affect some aspects of the model they do not harm the model overall performance.

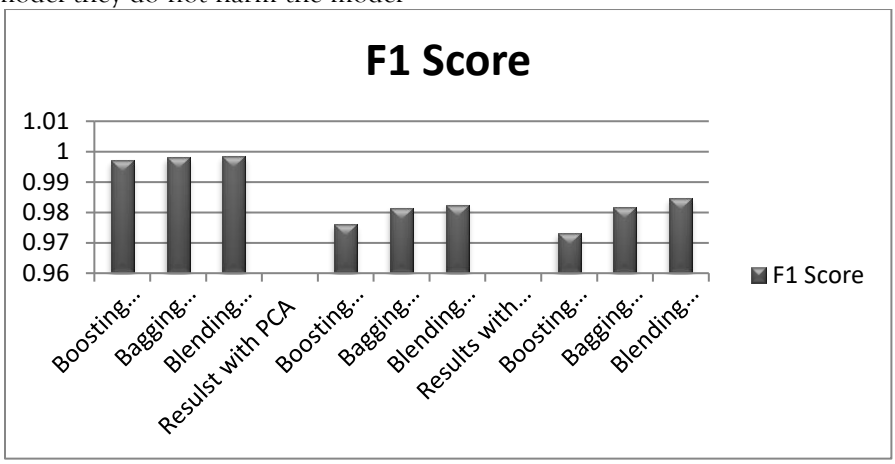



**Figure 9. F1-Score Comparison of the Ensemble Techniques**

**(v) Cohen's Kappa**

It is measure of agreement between predicted and actual classification adjusting for chance. Blending (Stacking) scores highest that is 0.9814 that shows strong agreement between predicted and actual labels. This metric is very important for understanding the model consistently when dealing with imbalanced datasets. Bagging also shows high performance with 0.99756 while Boosting (AdaBoost) achieve 0.9642 the lowest level. The drop in Kappa value with PCA and SMOTE is less significant as compare to accuracy that shows core agreement between predicted and true value remain strong even when dimensionally reduction or class imbalance techniques are used.

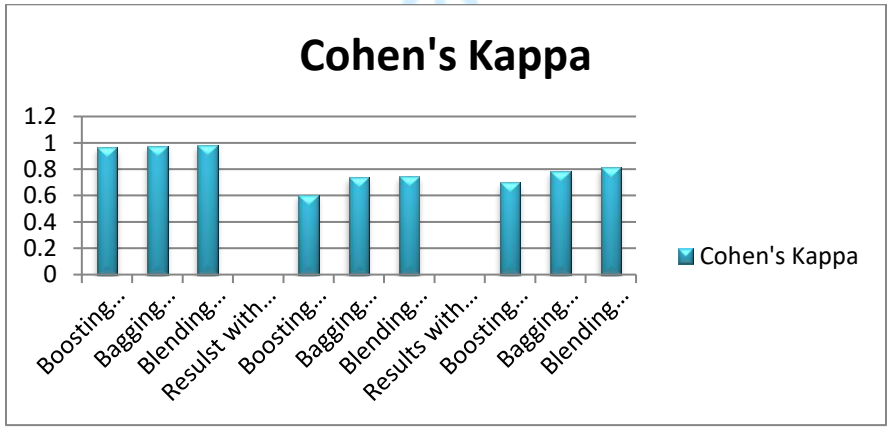



**Figure 10. Cohen's Kappa Comparison of the Ensemble Techniques**

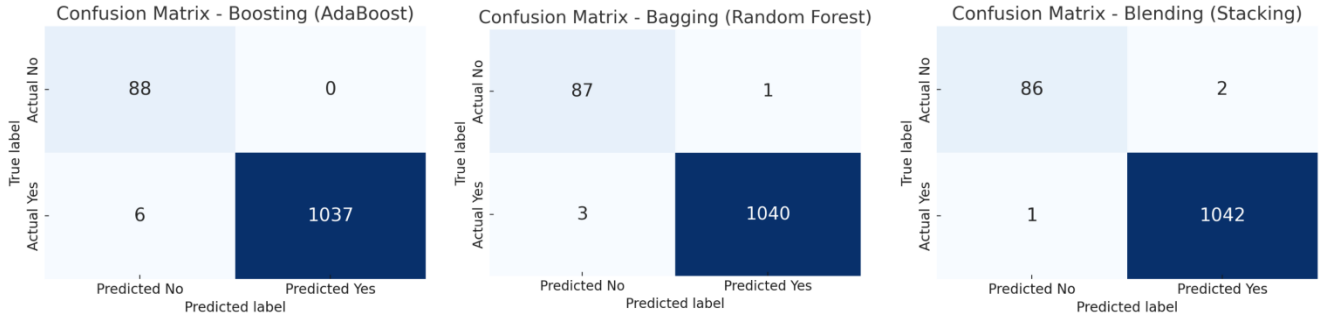
**(vi) Confusion Matrix**

Blending (Stacking) comes out on top for its ability to handle classes with relative precision. Bagging (Random Forest) performed well with recall, but AdaBoost may be slightly more cautious with its predictions, which might be beneficial in situations where avoiding False Positives is more critical than detecting all positives.

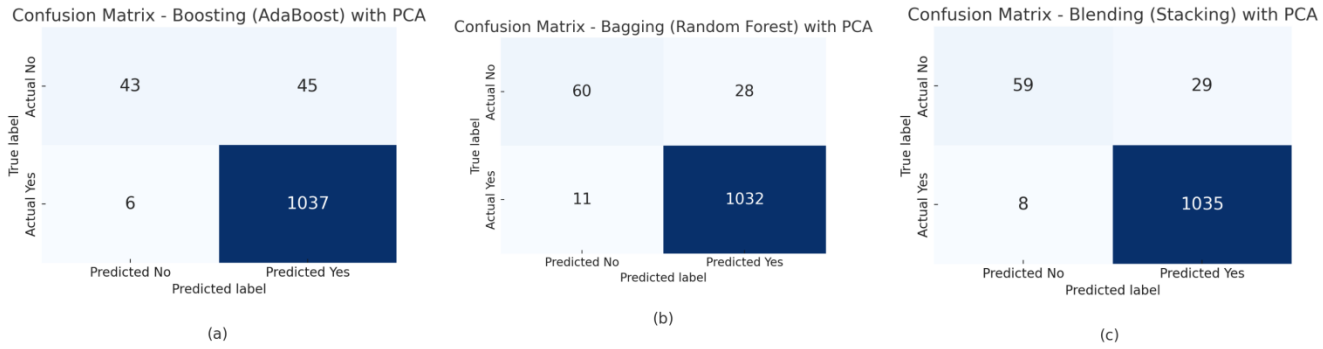
Volume 3, Issue 9, 2025

Frontier in Medical & Health Research

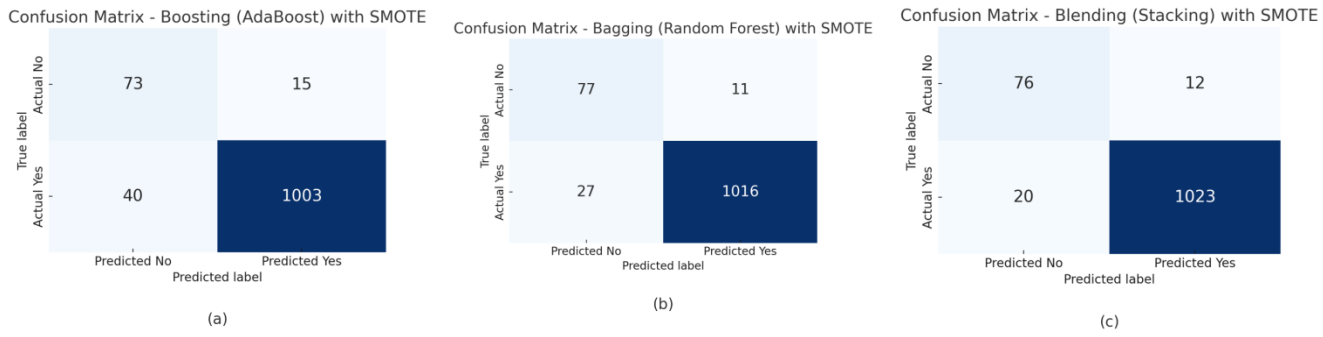
ISSN: (e) 3007-1607 (p) 3007-1593




Blending (Stacking) shows a balance between precision and recall that make it most versatile model. Bagging (Random Forest) is great for identifying "Yes" cases but at the cost of increased False Positives, while AdaBoost shows potential but struggles with its handling of both classes.




Each model brings its strengths and trade-offs. AdaBoost with SMOTE is strong in predicting the "Yes" class but struggles with the "No" class, which could be a limitation in certain applications. Bagging (Random Forest) with SMOTE excels at identifying "Yes" and decrease the False Negatives, that make it a better option for reducing missed positives, some precision with higher False Positives. Blending (Stacking) with SMOTE shows the best balance between precision and recall, provide a best performance across both classes with highest number of True Positives. For applications requiring consistent, well-rounded results, Blending (Stacking) stands out as most reliable choice.



Volume 3, Issue 9, 2025

Frontier in Medical & Health Research

ISSN: (e) 3007-1607 (p) 3007-1593

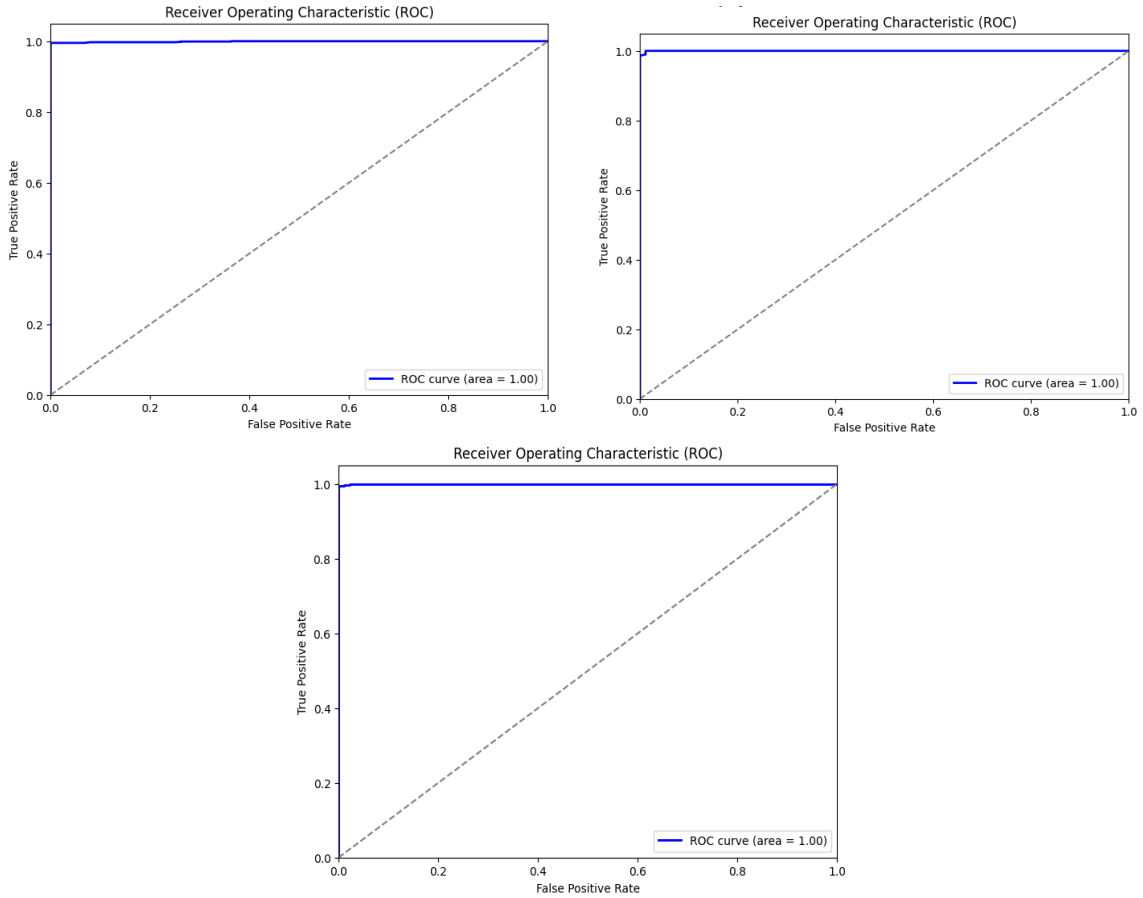
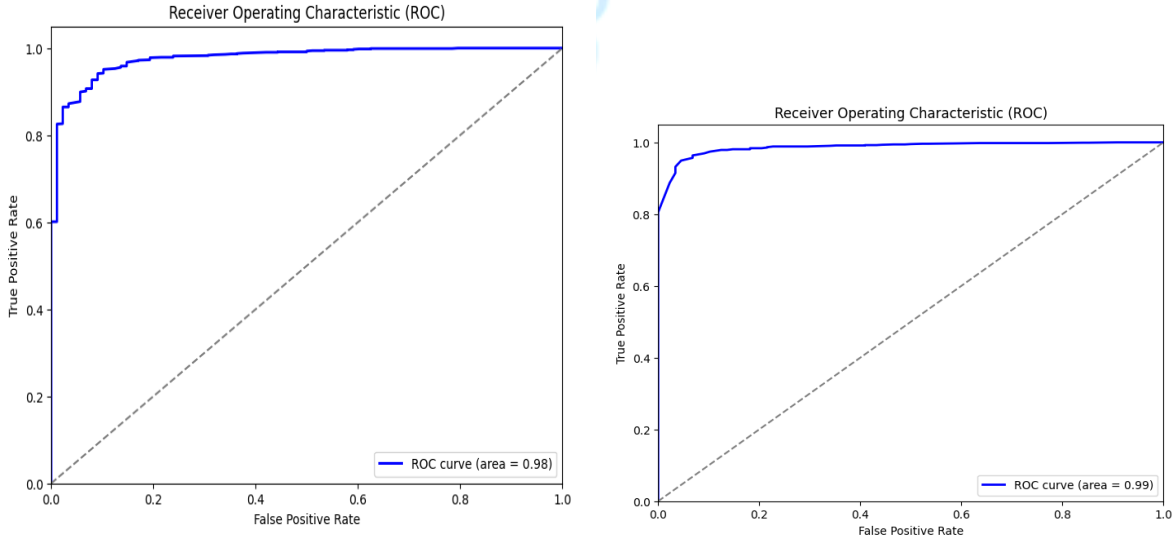



Figure 11. ROC Curves of Ensemble Techniques



Volume 3, Issue 9, 2025

Frontier in Medical & Health Research

ISSN: (e) 3007-1607 (p) 3007-1593

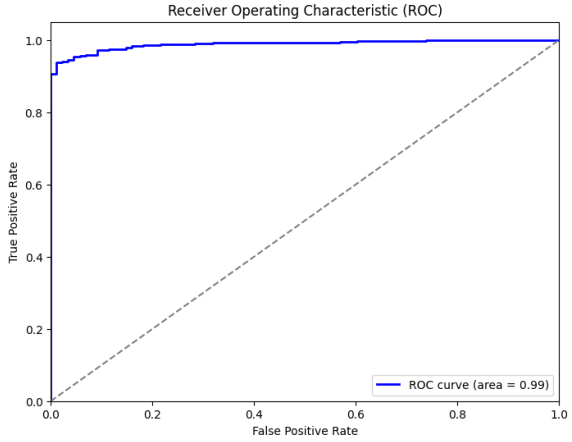



**Figure 12. ROC Curves of Ensemble Techniques with PCA**

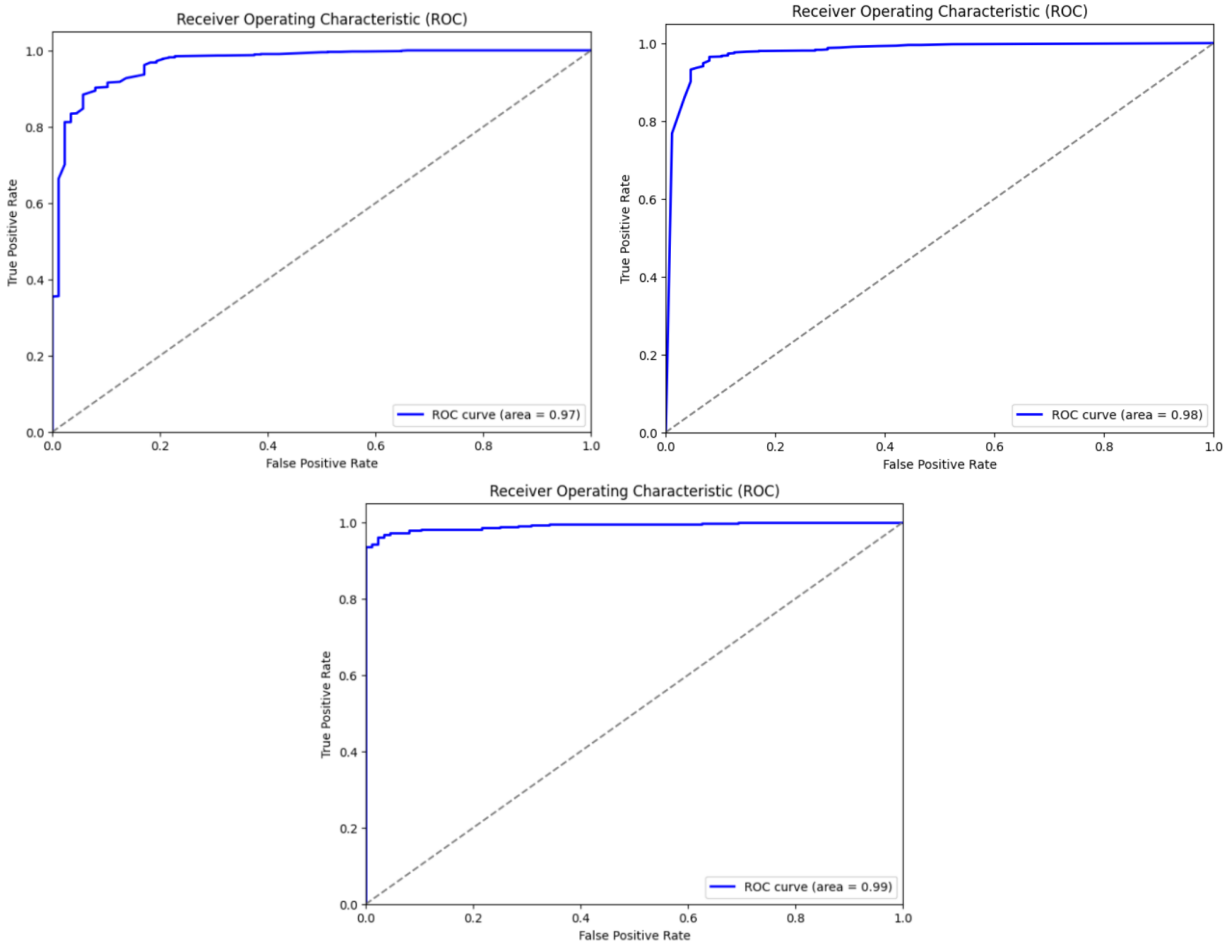



**Figure 14. ROC Curves of Ensemble Techniques with SMOTE**

**(vii) Aleatoric Uncertainty**

Boosting (AdaBoost) has low aleatoric uncertainty of 0.0085 that shows minimal random noise in its predictions. Blending (Stacking) and Bagging (Random Forest) show higher aleatoric uncertainties, with Blending (Stacking) at 0.0622, that it is more susceptible to variability possibly due to the diversity of models involved in the ensemble. The PCA and SMOTE transformations tend to increase aleatoric

Frontier in Medical & Health Research

ISSN: (e) 3007-1607 (p) 3007-1593

uncertainty across all models. This may highlight that the transformation or balancing techniques could amplify the inherent variability that affect confidence of model.

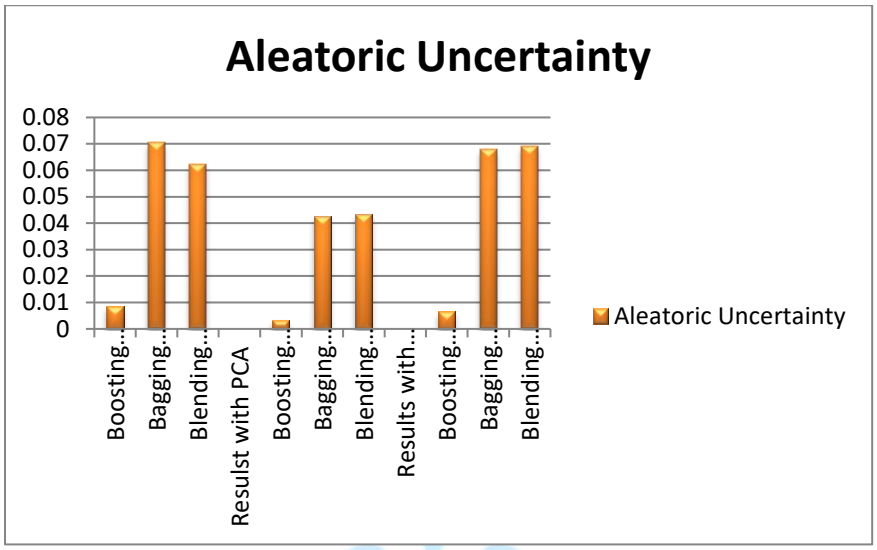



**Figure 15. Aleatoric Uncertainty Comparison**

(

**vii) Epistemic Uncertainty**

Boosting (AdaBoost) has zero epistemic uncertainty; shows model has high certainty about its predictions. Bagging (Random Forest) and Blending (Stacking) models show some level of epistemic uncertainty, particularly with Blending (Stacking) having most prominent uncertainty due to the complexity of combining multiple base models. The PCA and SMOTE transformations seem to induce more epistemic uncertainty, especially in the Blending (Stacking) model, in which uncertainty becomes more pronounced. This suggests that although these transformations can improve model generalization, they might also expose gaps in performance of model.

**(viii) Predictive Uncertainty**

Blending (Stacking) exhibits the highest predictive uncertainty, with PCA and SMOTE, indicating that the model is less confident in its predictions when complex transformations are used. This could be due to the increased complexity and variability introduced by the ensemble methods and feature transformation.

**Conclusion**

In conclusion, this study evaluates the utility of ensemble machine learning techniques in predicting hypothyroidism with emphasize on increasing the accuracy of model using with class balancing and dimensionality reduction. Despite use of PCA and SMOTW the performance of Blending (Stacking) remains strong that shows ability to handle complex medical data effectively and maintaining the level of precision and recall. On the other hand, Bagging (Random Forest) performance well in term of accuracy and provide high level of precision. It shows higher level of aleatoric uncertainty and predictive uncertainty as compare to Blending (Stacking) when PCA and SMOTE transformation are used. AdaBoost achieving the perfect precision fall behind in recall and accuracy and provide PCA leads to a noticeable drop in performance in accuracy and F1-score. The use of PCA and SMOTE enhance the model generalization by higher uncertainty in some cases. Blending (Stacking) maintains a better balance of uncertainty and overall performance. Blending (Stacking) is the most reliable model showing the best balance between performance and uncertainty making it most robust choice for predicting the hypothyroidism in medical

Frontier in Medical & Health Research

ISSN: (e) 3007-1607 (p) 3007-1593

application. Future work may evaluate optimizing Blending (Stacking) by addressing the uncertainty issues imposed by the PCA and SMOTE and using these models to larger diverse datasets for increased robustness.

**REFERENCES**

Graulich, E. (2024). Decoding the endocrine system, a few glands at a time. Nursing made Incredibly Easy, 22(3), 14-22.

Munoz-Robles, B., & Haney, B. (2024). Inventions and Patents for Treating Thyroid Disease. Ind. Health L. Rev., 21, 277.

Zamwar, U. M., & Muneshwar, K. N. (2023). Epidemiology, types, causes, clinical presentation, diagnosis, and treatment of hypothyroidism. Cureus, 15(9), 1-15.

Hegedüs, L., Bianco, A. C., Jonklaas, J., Pearce, S. H., Weetman, A. P., & Perros, P. (2022). Primary hypothyroidism and quality of life. Nature Reviews Endocrinology, 18(4), 230-242.

Patil, N., Rehman, A., Anastasopoulou, C., Jialal, I., & Saathoff, A. D. (2024). Hypothyroidism (nursing). StatPearls [Internet].

Pirahanchi, Y., Toro, F., & Jialal, I. (2018). Physiology, thyroid stimulating hormone.

Silvani, L. (2022). Fatigue To Fit: Connecting The Dots On Energy And Chronic Diseases. Lisa SIlvani.

de Virgilio, C. (2014). Question Sets and Answers. Surgery, 591–699.

Zhang, B., Tian, J., Pei, S., Chen, Y., He, X., Dong, Y., Zhang, L., Mo, X., Huang, W., Cong, S., et al. (2019). Machine learning-assisted system for thyroid nodule diagnosis. Thyroid, 29(6), 858–867.

Idarraga, A.J., Luong, G., Hsiao, V., Schneider, D.F. (2021). False negative rates in benign thyroid nodule diagnosis: Machine learning for detecting malignancy. J. Surg. Res. 268, 562–569.

Razia, S., Rao, M.N. (2016). Machine learning techniques for thyroid disease diagnosis-a review. Indian J. Sci. Technol., 9(28), 1–9.

Aversano, L., Bernardi, M.L., Cimitile, M., Iammarino, M., Macchia, P.E., Nettore, I.C., Verdone, C. (2021). Thyroid disease treatment prediction with machine learning approaches. Procedia Computer Science, 192, 1031–1040.

Usman AG, Ghali UM, Degm MA, Muhammad SM, Hincal E, Kurya AU, Işik S, Hoti Q, Abba SI. (2022). Simulation of liver function enzymes as determinants of thyroidism: a novel ensemble machine learning approach. Bulletin of the National Research Centre, 46(1), 73.

Dalal S, Lilhore UK, Faujdar N, Simaiya S, Agrawal A, Rani U, Mohan A. (2024). Enhancing thyroid disease prediction with improved XGBoost model and bias management techniques. Multimedia Tools and Applications. 2024 Jul 1:1-32.

Kour H, Singh B, Gupta N, Manhas J, Sharma V. (2023). Bagged based ensemble model to predict thyroid disorder using linear discriminant analysis with SMOTE. Research on Biomedical Engineering, 39(3), 733-746.

Sonuç E. (2021). Thyroid disease classification using machine learning algorithms. Journal of Physics: Conference Series, 1963(1), 012140.

Afshan N, Mushtaq Z, Alamri FS, Qureshi MF, Khan NA, Siddique I. (2023). Efficient thyroid disorder identification with weighted voting ensemble of super learners by using adaptive synthetic sampling technique. AIMS Mathematics, 8(10), 24274-309.

Devi MS, Kumar VD, Brezulianu A, Geman O, Arif M. (2023). A Novel Blunge Calibration Intelligent Feature Classification Model for the Prediction of Hypothyroid Disease. Sensors, 23(3), 1128.

Frontier in Medical & Health Research

ISSN: (e) 3007-1607 (p) 3007-1593

Obaido, G., Achilonu, O., Ogbuokiri, B., Amadi, C. S., Habeebullahi, L., Ohalloran, T., ... & Aruleba, K. (2024). An improved framework for detecting thyroid disease using filter-based feature selection and stacking ensemble. IEEE Access, 12(1), 1-15.

Akter, S., & Mustafa, H. A. (2024). Analysis and interpretability of machine learning models to classify thyroid disease. Plos one, 19(5), 030-067.

Alkhazzar, H. (2025). Enhanced Thyroid Disease Prediction Using Ensemble Machine Learning Techniques. Al-Salam Journal for Engineering and Technology, 4(1), 1-10.

Al-Hakim, R. R., Arief, Y. Z., Pangestu, A., Hidayah, H. A., Hamid, A. P., Andriand, A., ... & Alrahman, M. H. A. (2023). Predict the thyroid abnormality particular disease likelihood of the symptoms' certainty factor value and its confidence level: A regression model analysis. Sistemasi: Jurnal Sistem Informasi, 12(2), 415-424.