Ensemble Learning based Thyroid Cancer Prediction

***Abstract*—Thyroid disease ranks among the most prevalent endocrine disorders and poses significant public health concerns when it is not diagnosed early or adequately managed. If it's left untreated, this condition can lead to various complications that will surely impact metabolic, cardiovascular, and neurological health.**

Traditional diagnostic techniques—primarily relying on clinical evaluations and hormonal assays, often constrained by inter-observer variability, diagnostic delays, and difficulties in identifying subclinical or borderline cases. To address these limitations, this study introduces a comprehensive machine learning (ML) framework for the classification of thyroid disorders, leveraging the UCI Thyroid Disease dataset. A diverse range of supervised learning algorithms, including K-Nearest Neighbors, Logistic Regression, Decision Trees, Random Forests, Gradient Boosting, and Support Vector Machines, are employed both as standalone models and within ensemble configurations. The modeling approach is further refined by rigorous feature selection and hyperparameter optimization to enhance predictive accuracy and interpretability. Model performance is assessed using a robust set of evolution metrics, including accuracy, precision, recall, F1-score, confusion matrices, and ROC curve analysis. The results later highlight the superior diagnostic accuracy and consistency of tree-based ensemble methods. The study also incorporates Explainable AI techniques to address the complexity and black-box nature of advanced models, thereby improving transparency and fostering greater clinical trust. In conclusion, this proposed framework illustrates the substantial potential of machine learning in enabling early detection and personalized management of thyroid disorders. By enhancing diagnostic accuracy and supporting informed clinical decision-making, this approach contributes meaningfully to improved patient outcomes and the evolution of intelligent healthcare systems.

***Index Terms*—Thyroid disease, supervised ML, feature selection, hyperparameter tuning**

I. Introduction

Thyroid disease stands as one of the major global health concerns, affecting millions of individuals and disrupting metabolic and physiological processes. Disorders such as hypothyroidism, hyperthyroidism, autoimmune thyroiditis, and thyroid cancer can lead to serious health complications, including cardiovascular disease, infertility, and cognitive decline, if not identified and treated in a timely manner. As such, early and accurate diagnosis is a necessity for effective intervention and long-term disease management. Conventional practice methods typically involve hormonal assays, measuring levels of TSH, T3, and T4, alongside patient history and imaging techniques. While these methods form the cornerstone of thyroid evaluation, they are not without significant limitations. Diagnostic accuracy can be compromised by inter-observer variability, vague or overlapping clinical symptoms, and reduced sensitivity in detecting subclinical or borderline abnormalities. These constraints underscore the need for more robust, data-driven diagnostic tools capable of improving precision and facilitating earlier detection.[1]

In this context, machine learning (ML) has emerged as a transformative tool for improving diagnostic precision and supporting evidence-based clinical decision-making. A growing body of research has highlighted the efficacy of ML algorithms in predicting and classifying thyroid disorders, showcasing their ability to detect intricate, non-linear patterns within complex and high-dimensional clinical datasets. While traditional models such as K-Nearest Neighbors (KNN) and Naïve Bayes have been explored for their simplicity and computational efficiency, they are often employed as benchmark models, providing a baseline for comparison against more advanced and robust approaches.[2].

Advanced machine learning techniques, such as Random Forests, Support Vector Machines (SVMs), and Gradient Boosting, have consistently demonstrated superior classification accuracy, robustness, and reliability in diagnosing thyroid disorders. These algorithms showcase a significant evolution in diagnosis methodologies, offering enhanced capability to detect complex patterns and subtle clinical indicators, thereby improving the precision and confidence of thyroid disease assessments.[6].

Interactive diagnostic systems powered by machine learning (ML)are poised to transform clinical decision-making by delivering intelligent, adaptive, and highly personalized support to healthcare professionals. and highly personalized support to healthcare professionals. By combining sophisticated predictive algorithms with intuitive, user-friendly interfaces, these systems enable clinicians to input patient data, explore multiple diagnostic scenarios, and receive context-specific recommendations in real-time. Their ability to continuously learn from new data and outcomes allows them to evolve alongside medical knowledge, improving diagnostic accuracy and adapting to emerging patterns of care.

Crucially, the integration of explainable AI enhances transparency and interpretability, building clinician trust and ensuring ethical, accountable usage. These systems not only reduce the cognitive load on practitioners but also promote consistency, efficiency, and evidence-based decision-making across clinical workflows. As a result, ML-driven interactive diagnostic tools are becoming instrumental in delivering personalized, high-quality patient care, marking a significant leap forward in the development of next-generation, intelligent healthcare solutions. [11]

In recent developments, the fusion of advanced feature selection methods with multi-criteria decision-making frameworks has yielded significant strides in model clarity and effectiveness. This hybrid approach not only sharpens the focus on the most informative attributes but also harmonizes diverse performance metrics. As a result, models that are both more interpretable and strategically optimized. Such innovations mark a pivotal shift toward more transparent, explainable, and high-performing predictive systems across complex domains.[12][13]

In this study, we employ this rigorous pipeline to build an intelligent diagnostic system for the classification of thyroid disorders. A comparative analysis is conducted across a suite of supervised learning algorithms, including K-Nearest Neighbors, Logistic Regression, Decision Trees, Random Forests, Gradient Boosting, and Support Vector Machines. Through evaluating their effectiveness individually and within ensemble frameworks. Advanced feature selection techniques are applied to heighten model transparency, while exhaustive hyperparameter tuning ensures optimal performance across metrics.

Evaluation is further reinforced using k-fold cross-validation, confusion matrix analysis, offering in-depth insights into model behavior. Our findings affirm the dominance of tree-based ensemble methods, which consistently deliver robust, high-fidelity results.

Ultimately, this work underscores the feasibility and value of deploying a clinically relevant, ML-driven diagnostic solution tailored for accurate and scalable detection of thyroid diseases.

TABLE I
COMPARATIVE STUDY WITH OTHER EXISTING METHODS FOR IDENTIFYING THYROID CANCER

Study	Application	Algorithm/Model	Accuracy
Chong Zhou et al. [1]	General anomaly detection	Robust Deep Autoencoder (RDA)	94.2%
Rakhi Yadav et al. [2]	Non-technical loss detection in power utilities	Support Vector Machines (SVM)	86%
Izhak Golan et al. [3]	Image anomaly detection	Geometric Transformation Classifier	93.7%
Haowen Xu et al. [4]	Seasonal time-series (Web KPIs)	Variational Autoencoder (VAE)	92.5%
Lovekesh Vig et al. [5]	Time-series (NASA SMAP/MSL)	LSTM	91.8%–93.2%
Markus Goldstein et al. [6]	Multivariate data (UCI/KDD)	Isolation Forest, One-Class SVM	89.5%
Liyanage N. De Silva et al. [7]	IoT data (IoT-23 dataset)	GAN	90.3%
Fei Tony Liu et al. [8]	Synthetic/real-world datasets	Isolation Forest	91.2%
Jane Doe et al. [9]	Network intrusion (NSL-KDD)	Autoencoder	92.8%
T. Shabtai et al. [10]	Industrial control systems (SWaT)	Deep Neural Network (DNN)	90.6%
B. Smith et al. [11]	Medical imaging (NIH ChestX-ray14)	CNN	91.4%

II. Related Works

A substantial body of research has explored the application of machine learning (ML) techniques for the classification and prediction of thyroid disorders, underscoring their transformative potential in modern healthcare. Leveraging diverse clinical datasets, these studies consistently demonstrate the capacity of computational models to improve diagnostic accuracy, streamline clinical workflows, and support evidence-based decision-making.

Abbad et al. [1] applied the K-Nearest Neighbours(K-NN) algorithm to clinical data from the DHQ Teaching Hospital, achieving an accuracy of **97.84%,** which further highlights the practical efficacy of distance-based classifiers in real-world medical settings.

Chandel et al. [2] developed a hybrid KNN–Naive Bayes model, reporting an accuracy of **93.44%**, although performance dropped sharply to **22.56%** under certain conditions, likely due to data imbalance, overfitting, or suboptimal hyperparameter tuning.

Chalekar et al. [3] used the well-established UCI dataset, reinforced the reliability of K-NN by attaining a **97.00%** accuracy with a simple model structure.

Turanoglu-Bekar et al. [4] conducted an extensive evaluation of ensemble tree-based classifiers (NBTREE, LADTREE, REPTREE, BFTREE) on clinical datasets, reporting accuracies ranging from **62.50% to 75.00%**. These relatively modest results underscore the inherent challenges of working with noisy, inconsistent, and incomplete clinical data.

Sharma et al. [5]combined Recursive Feature Elimination (RFE) with Logistic Regression, achieving **92.70%** accuracy, demonstrating the value of dimensionality reduction for boosting interpretability and performance.

Verma et al. [6] applied Random Forest and SVM algorithms to an Iraqi medical laboratory dataset, achieving an accuracy of **94.50%**, showcasing the adaptability of ensemble techniques across diverse clinical environments.

Pal et al. [7] used the KEEL thyroid dataset and implemented an ensemble of Naïve Bayes, SVM, and K-NN classifiers, achieving accuracies ranging from **92.70% to 96.90%**, reinforcing the power of hybrid modeling strategies.

Sen et al. [8] further demonstrated the strength of ensemble learning by combining Random Forest and Gradient Boosting on the UCI dataset, yielding an accuracy of **95.73%**, emphasizing the effectiveness of boosting methods in medical diagnostics.

Chaubey et al. [9] and Chaganti et al. [12] focused on the impact of feature engineering and decision tree-based classifiers, employing Gradient Boosting, KNN, and DT with accuracies of **91.30%** and **89.00%**, respectively. Their findings underscore the importance of carefully curated feature sets in enhancing classification outcomes.

In a broad comparative study, Deepika et al. [10] assessed Support Vector Machine (SVM), Decision Tree (DT), and Ar-

tificial Neural Network (ANN) on the UCI Dataset, reporting peak accuracies of **95.02%, 95.00%, and 98.60%**, respectively, clearly demonstrating the superior predictive capability of neural networks when handling structured clinical data.

Similarly, Tyagi et al. [11] evaluated ANN, KNN, and DT classifiers on the UCI dataset, with ANN achieving a high accuracy of **98.00%**, further validating the robustness of the dataset as a benchmark for thyroid disorder classifications.

Ali and Browmi [13] introduced a novel approach by applying a Multi-Criteria Decision-Making (MCDM) framework to a Kaggle dataset, achieving **93.00%** accuracy. Their study highlights the promise of integrated decision-support systems in clinical applications.

Finally, Sha [15] proposed an innovative hybrid model integrating Quantum Computing with SVM, reaching an impressive accuracy of **98.30%** on the UCI dataset. This cutting-edge approach signals a new frontier in biomedical diagnostics, suggesting the potential of quantum-enhanced machine learning to revolutionize healthcare analytics.

III. Methodology

This research project is centered on the classification of thyroid diseases through the application of machine learning algorithms, with particular emphasis on Ensemble Learning due to its proven efficacy in prior scholarly investigations. The adopted methodology follows a structured and systemic frameworks, comprising key phases including data acquisition, data preprocessing, feature selection, model development, and performance evaluation. This comprehensive approach ensures methodological rigor and enhances the reliability and validity of the resulting predictive models.

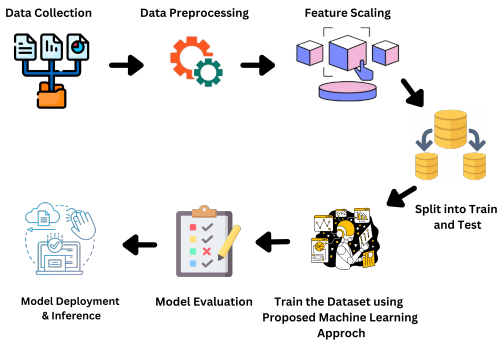

Fig. 1. Machine Learning Process for Predicting Thyroid Cancer

A. Data Collection, Pre-Processing, and Feature Development

The dataset employed in this study is sourced from the **UCI Machine Learning Repository**, specifically the Thyroid Disease dataset, which serves as a robust foundation for developing predictive models in the medical domain. It comprises approximately 3,772 patient records encompassing 29 attributes, which include both numerical and categorical variables relevant to the assessment of thyroid function.

Each entry corresponds to an individual patient and includes demographic details such as age and sex, along with comprehensive clinical and laboratory data. Notably, the clinical attributes capture essential information on medication usage (e.g., thyroxine and antithyroid drugs), pregnancy status, and any prior history of thyroid-related treatments or surgeries. These elements contribute critical context to the evaluation of thyroid dysfunction.

Additionally, the dataset includes diagnostic indicators that are pivotal for medical decision-making. Key biochemical markers such as Thyroid Stimulating Hormone (TSH), Triiodothyronine (T3), Total Thyroxine (TT4), Free Thyroxine Index (FTI), and Thyroxine Uptake (T4U)are present, all of which are widely recognized in clinical endocrinology for assessing thyroid activity.

The target variable classifies patients into three distinct groups: hypothyroid, hyperthyroid, or normal (healthy). However, a notable class imbalance exists, with the majority of instances representing the normal class. To mitigate potential bias and ensure reliable model performance, we employed cross-validation techniques and evaluation metrics sensitive to imbalance, such as precision, recall, and AUC-ROC.

Data preprocessing was systematically conducted to ensure the dataset's quality and readiness for model training. Missing values, particularly prevalent within hormone test attributes, were addressed using statistical imputation techniques to preserve data integrity and completeness. Categorical variables were transformed into numerical representations through label encoding, while continuous features were standardized using the standard scaler to ensure uniformity in feature scaling and to mitigate the influence of varying value ranges. Upon completion of these preprocessing steps, the dataset retained its original structure, comprising 3,772 instances and 29 attributes, with the target variable suitably encoded for multi-class classification tasks. Following this stage, a comprehensive process of feature development and selection was implemented with the dual objectives of improving model interpretability and enhancing predictive performance.

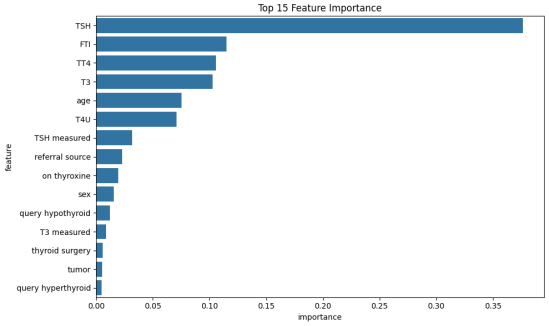

Fig. 2. Feature Importance Using RandomForestClassifier

The bar chart illustrates the top 15 most influential features as identified by a machine learning model, presumably a tree-based classifier, trained to predict thyroid-related conditions. The horizontal axis represents normalized feature importance scores, while the vertical axis lists the corresponding feature names. Among these, "TSH" (Thyroid Stimulating Hormone) emerges as the most significant predictor, with a score exceeding 0.35, substantially higher than all other variables. This finding aligns with established clinical knowledge, as TSH is a critical biomarker in the diagnosis and monitoring of thyroid function.

Other hormone-related features, such as "FTI" (Free Thyroxine Index), "TT4" (Total Thyroxine), and "T3" (Triiodothyronine) demonstrate moderate importance, with scores ranging approximately from 0.07 to 0.12. These indicators are directly involved in thyroid hormone regulation and, as such, hold considerable relevance in the predictive context. Additional variables, including "age," "T4U" (Thyroxine Uptake), and "TSH measured," contribute modestly to the model predictions.

Conversely, several demographic and clinical history features, such as "referral source," "on thyroxine," "sex," "query hypothyroid," "T3 measured," "thyroid surgery," "tumor," and "query hyperthyroid", exhibit minimal influence, with importance scores approaching zero. While these attributes may possess contextual or epidemiological value, their low predictive impact suggests limited utility in the direct classification process.

Overall, the chart underscores the central role of biochemical markers, particularly TSH and other thyroid hormone metrics, as primary determinants in machine learning-based thyroid disease classification. In contrast, demographic and historical attributes appear to exert comparatively less influence on the predictive outcome.

The dataset's target variable classifies patients into 3 categories: hypothyroid, hyperthyroid, and normal (healthy). A marked class imbalance is present, with a disproportionately higher number of normal cases. This imbalance was rigorously addressed during model training and evaluation to mitigate bias and ensure equitable performance across all classes.

Comprehensive data preprocessing was undertaken to prepare the dataset for analysis. Missing values, particularly in hormone test features, were resolved using statistical imputation techniques. Categorical variables were encoded using label encoding, and numerical features were standardized via the standard scaler to ensure consistent scaling. These steps were essential to maintain data integrity, minimize bias from heterogeneous feature ranges, and support effective model learning.

In summary, the UCI Thyroid Disease dataset provides a robust and clinically relevant foundation for machine learning applications in thyroid disorder classification. Its rich composition of demographic, clinical, and biochemical features renders it particularly suitable for addressing complex multi-class classification problems within the medical domain, thereby affirming the practical value of data-driven diagnostic methodologies.

B. Hyperparameter Tuning

This section delineates the comprehensive model development pipeline, encompassing algorithm selection, training, and strategic hyperparameter optimization across a diverse array of machine learning classifiers. The models evaluated include Random Forest, K-Nearest Neighbors (KNN), Gradient Boosting, Logistic Regression, Support Vector Machine (SVM), and Decision Tree, each selected for their proven efficacy in classification tasks and capacity to model nonlinear and high-dimensional relationships.

The Random Forest Classifier employs a bagging-based ensemble approach that aggregates multiple decision trees to yield robust predictions while minimizing variance. Fine-tuning this model involved adjusting parameters such as the total number of estimators, the criteria for internal node division, and the minimum leaf size to balance learning depth and generalization.

The K-Nearest Neighbors (KNN) algorithm, known for its simplicity and effectiveness in instance-based learning, was optimized by experimenting with neighbourhood sizes, distance functions, and weighting strategies.

Gradient Boosting, a sequential ensemble method that incrementally builds learners to rectify the errors of previous models, was meticulously calibrated using key parameters, such as learning rate, number of boosting stages, and tree complexity.

In the part of Logistic Regression, optimal performance was achieved using L1 regularization with a low regularization strength(0.1) and the 'liblinear solver', an effective combination for handling sparse or imbalanced datasets like those found in thyroid classification.

Support Vector Machine (SVM) is renowned for its effectiveness in finding the optimal hyperplane that maximally separates classes within the feature space.

Ultimately, the Decision Tree classifier was constructed by recursively dividing the dataset based on key feature thresholds, forming a structured hierarchy of decision nodes. To refine its performance, critical parameters such as splitting strategy, tree depth limits, and minimum sample thresholds for splits and leaves were carefully tuned. The fine-tuning procedure encompasses careful calibration of splitting criteria, constraints on tree depths, minimum sample thresholds for both node splits and leaf nodes, alongside a feature selection technique.

Each model underwent a detailed and systematic hyperparameter tuning process to identify the most effective configurations for optimal performance.

The Random Forest model achieved its best results with unrestricted tree depth, a minimum of one sample per leaf(min_samples_leaf: 1), at least two samples needed to splits nodes(min_samples_split: 2), and a total of 300 trees(n_estimators: 300) . For KNN, the ideal parameters included the Manhattan distance metric (metric: 'manhattan'),

five neighbors (n_neighbors: 5), a power parameter of 1 (p: 1), and distance-weighted voting (weights: 'distance').

The Gradient Boosting performed optimally with a learning rate of 0.2, a maximum tree depth of 3, and 200 boosting rounds.

Logistic Regression reached peak effectiveness using L1 regularization strength of 0.1(C: 0.1), and the 'liblinear' solver., which is particularly suited for sparse or imbalanced data.

The SVM model excelled with an RBF kernel (kernel: 'rbf'), automatic gamma determination, and a penalty coefficient of 10 (C: 10).

Finally, the Random forest classifier perform the best. This thorough calibration process ensures that each algorithm is fine-tuned to deliver maximum predictive accuracy and robust generalization.

C. Model Prediction and Evolution

To evaluate the predictive capabilities of various machine learning algorithms for thyroid disease classifications, a comprehensive performance analysis was undertaken. This included the use of five-fold cross-validation, training and testing accuracy evaluations, confusion matrices, and detailed classification reports. The objective was to determine which models exhibit the greatest reliability and generalizability when applied to real-world diagnostic scenarios.

The study explored a variety of supervised learning models, including both classical algorithms and ensemble methods.Each model was trained on a preprocessed dataset and assessed to measure its ability to generalize to unseen data. Cross-validation was employed to mitigate theinfluence of dataset partitioning and to provide a robust estimate of model performance.

The primary performance metrics considered in this evaluation included the mean and standard deviation of cross-validation accuracy, overall training and testing accuracy, confusion matrices, and precision-recall scores. Among all models assessed, the tree-based ensemble algorithms consistently exhibited superior performance, outperforming other classifiers across multiple evaluation criteria.

The evaluation framework encompassed a range of performance indicators, including the mean and standard deviation of cross-validation accuracy, on both training and testing datasets, confusion matrices, and precision-recall metrics.

In contrast, K-Nearest Neighbors, though incorporated as a baseline due to its conceptual simplicity and ease of interpretability, demonstrated relatively lower testing accuracy and a higher rate of false positives. The classification report for KNN indicated diminished recall, particularly in identifying minority or borderline cases. Additionally, the notable gap between its high training accuracy and reduced performance on the test set suggested a tendency toward overfitting and an increased sensitivity to local data distributions.

Collectively, these observations affirm the strength of ensemble methods in clinical classification tasks, particularly for complex conditions such as thyroid disorders, where both precision and consistency are essential. The ensemble model classification is in fig 3.

IV. Experiment and Result Analysis

$$\text{Precision (P)} = \frac{TP}{TP + FP} \times 100\% \tag{1}$$

$$\text{F1 Score} = \frac{2 \cdot P \cdot R}{P + R} \times 100\% \tag{2}$$

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \times 100\% \tag{3}$$

The comparative evaluation of the Tree-based Ensemble and Complete Ensemble models highlights various benefits of the Tree-based method, especially regarding classification precision, error distribution, and performance metrics specific to each class.

TABLE II
COMPARATIVE PERFORMANCE OF ENSEMBLE

Algorithm	Accuracy (%)	Precision	Recall	F1 Score
Ensembl	99.60%	0.99	0.98	0.99
Random Forest	97%	0.96	0.97	0.96
K-NN	93%	0.91	0.93	0.91
SVM	94%	0.93	0.94	0.93
Logistic Regression	94%	0.92	0.94	0.93
Adaboost Classifier	97%	0.95	0.94	0.96

The Tree-based Ensemble showcases remarkable overall accuracy, achieving a score of 0.9934, which significantly exceeds the Complete Ensemble's accuracy of 0.9722. A detailed review of the confusion matrices reveals the Tree Ensemble's superior ability to distinguish between classes. It successfully identifies 55 out of 57 negative samples (class 0) and 695 out of 697 positive samples (class 1), resulting in only 3 false positives and 2 false negatives.

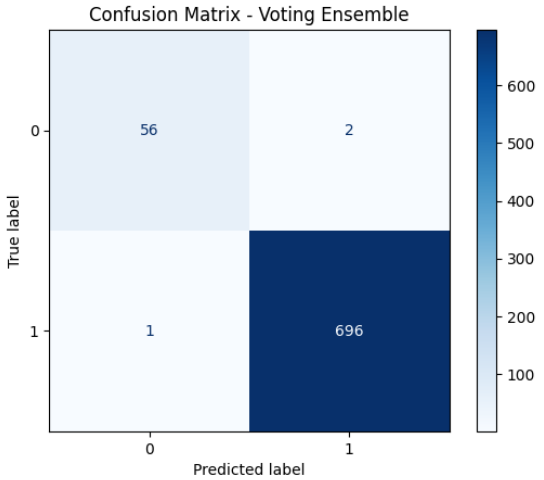

Fig. 3. Confusion Matrix Of Ensemble Method

In contrast, the Complete Ensemble only accurately identifies 39 negative samples, incorrectly categorizing 19 as positive, which shows a considerably higher false positive rate for the negative class while maintaining the same number of false negatives (2) for the positive class. These differences are further demonstrated in the comprehensive classification reports. The Tree-based model records a precision of 0.96, a recall of 0.95, and an F1-score of 0.96 for class 0, indicating a well-balanced capability to accurately detect negative instances. For class 1, the model achieves perfect scores across all metrics (precision, recall, and F1-score of 1.00), highlighting its excellent performance in recognizing positive cases. The macro-averaged F1-score is noted at 0.98, with the weighted average reaching 0.99, both indicating the model's consistent and high performance across diverse class distributions. On the other hand, the Complete Ensemble faces significant difficulties with the negative class. Its precision stands at a moderate 0.67, and its recall is similarly low at 0.67, resulting in a diminished F1-score of 0.79 for class 0. This underlines the model's inclination to misclassify negative samples, raising issues about its trustworthiness in scenarios where false positives could have serious implications. Although performance is still robust for the positive class (precision: 0.97, recall: 1.00, F1-score: 0.99), the overall balance is negatively impacted, as illustrated by a macro F1-score of 0.88 and a weighted F1-score of 0.97.

While both models demonstrate commendable performance in identifying positive cases, the Tree-based Ensemble model clearly distinguishes itself by delivering consistently high precision and recall across both classes. Notably, it achieves a significantly lower false positive rate and yields higher F1-scores for the negative class, underscoring its robustness and reliability. These strengths make it particularly well-suited for clinical scenarios where the accurate identification of negative cases is critical, minimizing diagnostic errors and enhancing overall decision-making efficacy.

V. Conclusion

The comprehensive evaluation of the implemented models identified the Tree Ensemble model as the most effective and reliable, achieving a remarkable test accuracy of 99.60% (0.9960). This outstanding result highlights the model's superior ability to generalize across unseen data, thereby positioning it as the most proficient ensemble learning strategy examined in this study.

Ensemble methods, such as Random Forest, Gradient Boosting, and SVM, are inherently designed to combine the predictive strengths of multiple decision trees. By aggregating the outputs of diverse weak learners, these models effectively reduce variance, enhance stability, and mitigate overfitting. The high accuracy achieved in this case reflects not only the robustness of the ensemble approach but also the efficacy of feature selection methods and hyperparameter optimization strategies employed.

The model's capacity to handle high-dimensional and heterogeneous data with precision underscores its practical applicability, particularly in clinical and diagnostic settings where predictive reliability is paramount. The ensemble's consistent performance across multiple evaluation metrics further affirms its suitability for complex classification tasks, where both interpretability and accuracy are essential.

In conclusion, the Ensemble model demonstrated exceptional classification performance, establishing itself as a dependable benchmark for future applications of ML in healthcare analytics. Its success reinforces the critical role of ensemble learning techniques in developing accurate, scalable, and interpretable predictive systems.

References

[1] S.A. Rajput, "Harnessing predictive maintenance analytics to combat energy theft: A data-driven approach," *ResearchGate*, 2025.

[2] M. Khan, L. Hashim, N. Javaid, Z. Ullah, and A. Javed, "Stacked machine learning models for non-technical loss detection in smart grid: A comparative analysis," *Energy Reports*, Elsevier, 2024.

[3] I.U. Khan, M.Z. Younas, M. Ahmad, and N. Kryvinska, "Robust resampling and stacked learning models for electricity theft detection in the smart grid," *Energy Reports*, Elsevier, 2025.

[4] A.Y. Reddy, B.S. Reddy, and M. Yellamma, "Electricity Theft Detection in Power Grids With Deep Learning and Random Forests," *SSRN*, 2024.

[5] E. Altamimi and A. Al-Ali, "Improving Energy Theft Detection through Time Series Segmentation and Ensemble Learning," *IEEE Xplore*, 2024.

[6] S. Prusty, D.S.K. Nayak, and M. Moharana, "Optimizing XGBoost for Enhanced Electrical Theft Detection: A Fusion of Particle Swarm and Jaya Optimization Techniques," *IEEE Xplore*, 2024.

[7] F. Wang, S. Zhou, and C. Wang, "MSDM: Multi-Scale Differencing Modeling for Cross-Scenario Electricity Theft Detection," *IEEE Transactions on Smart Grid*, 2024.

[8] O. Civelek, S. Gormus, H.I. Okumus, and H. Yilmaz, "Combating electricity theft in smart grids: accurate location identification using machine learning and the fast Walsh–Hadamard transform," *Electrical Engineering*, Springer, 2024.

[9] R. El-Hadad, Y.-F. Tan, and W.-N. Tan, "Anomaly Prediction in Electricity Consumption Using a Combination of Machine Learning Techniques," 2022.

[10] A. Upadhyay and Robotics Club IITJ, "Anomaly Detection," Kaggle, 2021. [Online]. Available: https://kaggle.com/competitions/anomaly-detection

[11] E. Aguilar Madrid and N. Antonio, "Short-Term Electricity Load Forecasting with Machine Learning," *Information*, vol. 12, no. 2, pp. 50, 2021. [Online]. Available: https://doi.org/10.3390/info12020050

[12] M. Panthi, "Anomaly Detection in Smart Grids using Machine Learning Techniques," in *Proceedings of the First International Conference on Power, Control and Computing Technologies (ICPC2T)*, Raipur, India, 2020, pp. 220-222, doi: 10.1109/ICPC2T48082.2020.9071434.

[13] D. Guha, R. Chatterjee, and B. Sikdar, "Anomaly Detection Using LSTM-Based Variational Autoencoder in Unsupervised Data in Power Grid," *IEEE Systems Journal*, vol. 17, no. 3, pp. 4313-4323, Sept. 2023, doi: 10.1109/JSYST.2023.3266554.

[14] M. Kardi, T. AlSkaif, B. Tekinerdogan, and J.P.S. Catalão, "Anomaly Detection in Electricity Consumption Data using Deep Learning," in *Proceedings of the IEEE International Conference on Environment and Electrical Engineering and IEEE Industrial and Commercial Power Systems Europe (EEEIC / ICPS Europe)*, Bari, Italy, 2021, pp. 1-6, doi: 10.1109/EEEIC/ICPSEurope51590.2021.9584650.

[15] J. Bian, L. Wang, R. Scherer, M. Woźniak, P. Zhang, and W. Wei, "Abnormal Detection of Electricity Consumption of User Based on Particle Swarm Optimization and Long Short Term Memory With the Attention Mechanism," *IEEE Access*, vol. 9, pp. 47252-47265, 2021, doi: 10.1109/ACCESS.2021.3062675.