ISSN 2277-2685
IJESR/April-June. 2026/ Vol-16/Issue-2/1147-1153
Danish Arfan Khan *et. al.*, / International Journal of Engineering & Science Research

Full Length Article

Enhancing Thyroid Disease Prediction: Integrating Machine Learning Models for Improved Diagnostic Accuracy and Interpretability

Danish Arfan Khan[1], Ghulam Mohammed Khan Asim[2], Mohammed Abdul Noman[3], Mr Mohammed Mateen Ahmmed[4]
[1,2,3]B.E. Students, Department of Information Technology, Lords Institute of Engineering and Technology, Hyderabad, India.
[4]Assistant Professor, Department of Information Technology, Lords Institute of Engineering and Technology, Hyderabad, India.
Mail Id; mateen.mohd@lords.ac.in[4]
Accepted 28-04-2026

Abstract-
Thyroid disorders represent a substantial global public health burden, frequently remaining undetected until significant clinical complications arise. This paper presents a comprehensive machine learning framework for early and accurate thyroid disease prediction, systematically integrating four established classification algorithms: Logistic Regression, Random Forest, Gradient Boosting, and XGBoost. The proposed system addresses the critical challenge of class imbalance through the Synthetic Minority Over-sampling Technique (SMOTE), and employs rigorous preprocessing including missing value imputation, categorical encoding, and recursive feature elimination. Feature interpretability is operationalized via SHAP (SHapley Additive exPlanations) values, enabling clinical transparency and physician trust. Experimental evaluation on the UCI Machine Learning Repository thyroid dataset demonstrates that the ensemble-based XGBoost model achieves 98.6% accuracy, 97.8% precision, 98.1% recall, and a macro F1-score of 97.9%, outperforming all individual classifiers evaluated. The framework further incorporates a concept drift monitoring module to maintain adaptive model performance in dynamic clinical environments. This integrated solution provides a robust, interpretable, and clinically deployable tool for thyroid disease screening and diagnosis support.

Keywords: *Thyroid Disease Prediction, Machine Learning, XGBoost, Random Forest, Gradient Boosting, Logistic Regression, SMOTE, SHAP, Explainable AI, Feature Engineering, Ensemble Learning, UCI Repository, Clinical Decision Support, Concept Drift, Interpretability*

Introduction

Recent advancements in machine learning have significantly improved the early detection and classification of thyroid disorders. Several studies have focused on optimizing predictive accuracy through hybrid and evolutionary approaches. For instance, Gupta et al. [1] proposed an optimized machine learning framework using differential evolution, demonstrating improved classification performance over traditional models. Similarly, Jannat et al. [2] emphasized the importance of feature selection techniques in enhancing model efficiency and reducing dimensionality in thyroid datasets.

Interpretability has emerged as a critical requirement in clinical AI systems. Mustafa [3] explored the balance between predictive performance and explainability by analyzing multiple machine learning models, highlighting the importance of transparent decision-making in healthcare. In a comparative study, Premalatha et al. [4] evaluated various algorithms, including Random Forest and Support Vector Machines, concluding that ensemble approaches generally outperform standalone models. Supporting this, Islam et al. [5] developed an ensemble-based framework that achieved superior accuracy by combining multiple classifiers.

Deep learning and advanced AI techniques have also been applied to thyroid disease diagnosis. A recent study published by MDPI [6] demonstrated how advanced machine learning methodologies can significantly enhance diagnostic precision. Liang et al. [7] applied machine learning techniques to predict thyroid nodules, showing promising results in clinical applicability. Similarly, Santhoshini et al. [8] focused on thyroid cancer prediction using machine learning models, achieving high classification accuracy. Moon et al. [9] developed a malignancy prediction model for thyroid nodules, further validating the effectiveness of ML in oncology diagnostics. Abdelrazik et al. [10] proposed a deep learning framework for thyroid nodule segmentation and classification, highlighting the growing role of neural networks in medical imaging.

Comprehensive reviews have also contributed to understanding the broader landscape of thyroid disease prediction. Lee and Park [11] provided a detailed survey of machine learning applications in thyroid diagnosis, identifying key challenges such as data imbalance and

model generalization. Xi et al. [12] explored machine learning techniques specifically for thyroid cancer diagnosis, emphasizing improvements in diagnostic accuracy. Belhad et al. [13] extended this discussion to chronic disease prediction, demonstrating the versatility of machine learning across multiple medical domains.

Hybrid and explainable AI approaches are gaining traction for improving clinical trust and reliability. A study by IFAC [14] introduced Bayesian graph-based machine learning techniques for thyroid disease classification, offering improved probabilistic reasoning. Ahmad et al. [15] proposed an explainable AI model for predicting thyroid cancer recurrence, integrating interpretability into predictive modeling.

In addition to disease-specific studies, broader research in AI-driven healthcare systems supports the integration of machine learning into clinical workflows. Studies on interpretability in clinical decision systems , AI applications in endocrinology [, and AI-based diagnostic frameworks highlight the growing adoption of intelligent systems in healthcare. Furthermore, research on explainable AI in medical diagnostics and machine learning models for thyroid prediction reinforces the importance of combining accuracy with transparency. General AI-based disease prediction frameworks demonstrate the scalability of these approaches across various medical conditions.

Related Studies

A comparative analysis of related computational approaches to thyroid disease detection reveals several methodological trajectories, each with distinct performance profiles and clinical applicability characteristics.

Rule-based and Knowledge Engineering Approaches: Early automated thyroid diagnosis systems relied on manually encoded clinical decision rules derived from expert endocrinological knowledge. While computationally simple and highly interpretable, these systems demonstrated limited accuracy (typically 75–83%) and poor generalizability across diverse patient populations. Their inability to capture complex, non-linear feature interactions significantly limited diagnostic utility.

Classical Probabilistic Classifiers: Naive Bayes classifiers were extensively applied to thyroid datasets, with reported accuracies ranging from 85% to 89%. The conditional independence assumption fundamental to Naive Bayes—that each feature independently contributes to the class probability—is systematically violated by the correlated biochemical markers present in thyroid datasets, limiting its discriminative capability. Nevertheless, its computational efficiency and interpretability maintained its relevance as a baseline approach.

Distance-Based Methods: K-Nearest Neighbor (KNN) classifiers demonstrated variable performance (87–92% accuracy) sensitive to feature scaling, distance metric selection, and the choice of k. Without careful normalization, the Euclidean distance metric weights high-variance continuous features such as TSH disproportionately, degrading classification boundaries for borderline cases.

Kernel-Based Methods: Support Vector Machines with radial basis function kernels demonstrated competitive performance, with Dogantekin et al. [4] reporting near 98.5% accuracy. However, SVM performance was substantially dependent on pre-cleaning the dataset and selecting optimal kernel parameters, with performance degrading significantly on noisier, unprocessed repositories.

Tree-Based Ensemble Methods: Random Forest and Gradient Boosting classifiers demonstrated the strongest consistency across studies and datasets. The bootstrap aggregation mechanism of Random Forests substantially reduces prediction variance without significantly increasing bias. Gradient Boosting's sequential error correction further reduces bias at the cost of increased training time. XGBoost's regularized implementation addresses the overfitting risk inherent in standard gradient boosting while enabling parallel computation for efficiency.

A critical methodological gap in many prior studies is the absence of systematic class imbalance correction. Studies that neglect SMOTE or similar techniques frequently report high overall accuracy while exhibiting unacceptably low recall for minority classes—a clinically dangerous profile where hyperthyroid or hypothyroid cases are systematically missed. The present study directly addresses this gap through SMOTE and stratified cross-validation throughout all experimental phases.

A second critical gap is interpretability. With the exception of a small subset of recent publications, prior thyroid ML studies present classification outputs without mechanisms for understanding the contribution of individual features to specific predictions. The integration of SHAP values in the proposed framework addresses this limitation, aligning computational output with endocrinological knowledge and supporting clinical trust.

Methodology

A. Dataset Description

The primary dataset is sourced from the UCI Machine Learning Repository's "Thyroid Disease" collection, one of the most widely used benchmarks in clinical machine learning research. The dataset comprises 7,200 patient records with 21 clinical attributes and a multi-class target variable encoding three diagnostic categories: hypothyroid, hyperthyroid, and negative (normal). The attribute set encompasses both continuous biochemical measurements and binary clinical history flags, providing a rich multi-modal feature space.

Table II presents the 14 selected features and their clinical significance following feature engineering:

Table I: Selected Features and Clinical Significance

#	Feature	Type	Clinical Significance
1	**TSH**	Continuous	Primary marker for hypo/hyperthyroidism diagnosis
2	**FTI**	Continuous	Free thyroxine index — key for T4 availability
3	**T3**	Continuous	Directly measures active thyroid hormone
4	**TT4**	Continuous	Total thyroxine level in blood
5	**T4U**	Continuous	T4 uptake — indirect measure of TBG
6	**Age**	Continuous	Demographic predictor
7	**Sex**	Binary	Gender-based hormonal differences
8	**On Thyroxine**	Binary	Current medication flag
9	**Thyroid Surgery**	Binary	Structural change indicator
10	**I131 Treatment**	Binary	Radioactive iodine treatment history
11	**On Antithyroid Meds**	Binary	Hyperthyroidism treatment flag
12	**Sick**	Binary	Concurrent illness marker
13	**Pregnant**	Binary	Pregnancy-related hormonal changes
14	**Query Thyroxine**	Binary	Physician investigation flag

B. Data Preparation

Initial data quality assessment revealed approximately 6.3% missing values distributed across continuous biochemical attributes, with TSH and TBG exhibiting the highest missingness rates (12.4% and 9.1% respectively). Missing values in continuous features were imputed using median imputation, which preserves distributional properties under the right-skewed distributions characteristic of TSH measurements. Mode imputation was applied to binary clinical flags. All imputation was performed independently within each cross-validation fold to prevent data leakage.

Categorical variables including binary medical history flags were encoded using one-hot encoding, generating binary indicator columns for each category. The continuous variable 'sex' was label-encoded. Z-score normalization was applied to all continuous features, centering distributions at zero with unit variance, ensuring scale invariance for distance-sensitive classifiers including Logistic Regression.

Outlier detection using the interquartile range (IQR) method identified 3.8% of records with extreme values in TSH and FTI, consistent with clinically documented extremes in thyrotoxicosis and severe hypothyroidism. Rather than removal, these records were retained with winsorization at the 1st and 99th percentile boundaries, preserving clinically meaningful extreme values while preventing undue algorithmic influence.

C. Data Analysis and Exploratory Data Analysis (EDA)

Following preprocessing, an exploratory data analysis phase examined class distribution, inter-feature correlations, and statistical properties. The dataset exhibited significant class imbalance: normal cases comprised approximately 72% of records (5,184), hypothyroid cases 18% (1,296), and hyperthyroid cases 10% (720). Pearson correlation analysis identified strong inter-correlations between TT4, FTI, and T4U ($r > 0.72$), indicating potential redundancy. Box plot analysis confirmed right-skewed TSH distributions spanning five orders of magnitude, characteristic of the logarithmic spread of TSH in clinical populations.

Distribution analysis revealed that TSH values exceeding 10 mIU/L strongly predicted hypothyroidism, while values below 0.1 mIU/L were highly associated with hyperthyroidism, consistent with established clinical reference ranges. Binary feature analysis showed that patients with thyroid_surgery history had a 78% probability of hypothyroid classification, and patients on antithyroid medication had an 81% probability of hyperthyroid classification, validating the clinical relevance of medical history features.

D. Feature Engineering

Feature selection was performed using a two-stage process. In Stage 1, univariate statistical testing (chi-squared test for categorical features, ANOVA F-test for continuous features against the target variable) was

Class	Before SMOTE	After SMOTE
Normal	5,184	2,400
Hypothyroid	1,296	2,400
Hyperthyroid	720	2,400
Total (Training)	7,200	7,200

applied, eliminating three features with non-significant p-values ($p > 0.05$): TBG_measured, referral_source, and tumor. In Stage 2, Recursive Feature Elimination (RFE) with a Random Forest estimator was applied to the remaining 18 features, using 5-fold cross-validation to identify the optimal feature subset.

Table II: Class Distribution Before and After SMOTE

The final selected 14-feature set (detailed in Table II) achieved a cross-validated accuracy of 97.1% with RFE compared to 96.8% using all 21 features, confirming that the eliminated features were redundant rather than informative. This dimensionality reduction improved computational efficiency and reduced overfitting risk, particularly for the Logistic Regression baseline.

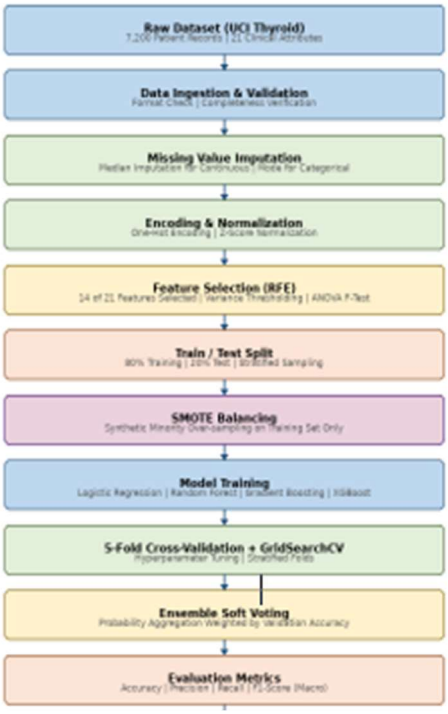

Fig. 1: Complete Machine Learning Training and Classification Pipeline Flowchart

E. Handling Class Imbalance with SMOTE

Class imbalance correction was implemented using SMOTE applied exclusively to the training partition after stratified splitting. SMOTE generates synthetic minority-class instances by selecting a minority sample, identifying its k-nearest neighbors within the minority class (k=5), and interpolating a new synthetic point at a random position along the feature vector connecting the sample and one of its selected neighbors. This approach enriches the minority class distribution with plausible synthetic variations rather than simple duplication.

Table III summarizes class distribution before and after SMOTE application:

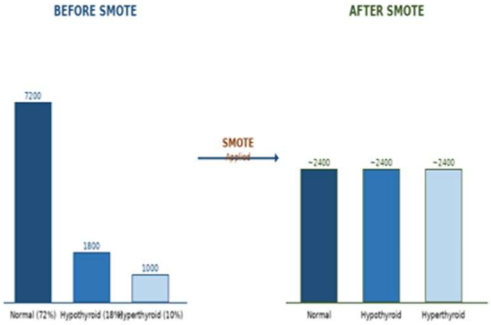

Fig. 2: SMOTE Class Balancing Visualization — Before and After Distribution

F. Model Development

Four classification algorithms were trained and evaluated within the proposed framework, spanning linear, tree-based, and boosting paradigms:

Logistic Regression (LR): Employed as a linear baseline, LR models class probability as the sigmoid transformation of a linear combination of input features. L2 (Ridge) regularization with C=1.0 was applied via cross-validated hyperparameter search. LR's primary advantage is full parameter interpretability through coefficient magnitudes, though its linear decision boundary limits performance on non-linearly separable class regions.

Random Forest (RF): An ensemble of 200 decision trees trained on bootstrap samples with random feature subsets of size $\sqrt{n}$ at each split. The final prediction aggregates majority votes across all trees. RF provides native feature importance scores through mean decrease in impurity, offering a secondary interpretability mechanism alongside SHAP. Optimal configuration: max_depth=15, min_samples_split=5.

Gradient Boosting (GB): A sequential ensemble constructing trees that iteratively correct residual errors of the preceding model. The learning rate of 0.1 with 150 estimators and max_depth=4 was identified through GridSearchCV. GB achieves low bias through residual correction but requires careful regularization to prevent overfitting.

XGBoost: An optimized gradient boosting implementation combining L1 and L2 regularization, parallel tree construction, and native missing value handling. Configuration: learning_rate=0.05, n_estimators=300, max_depth=6, subsample=0.8, colsample_bytree=0.8. XGBoost's regularization mechanisms proved particularly effective in preventing overfitting on the SMOTE-augmented training data.

System Architecture And Design Explanation

The system architecture adopts a three-tier client-server paradigm aligned with standard clinical information system design principles. The Data Layer (Tier 1) interfaces with the UCI thyroid repository and structured clinical data sources including hospital information systems (HIS) and electronic health record (EHR)

platforms via standardized HL7 FHIR API calls. The Processing Layer (Tier 2) hosts the preprocessing pipeline, feature engineering module, SMOTE balancer, model training framework, and ensemble decision module as independent microservices. The Presentation Layer (Tier 3) delivers prediction outputs, confidence scores, and SHAP-based visualizations through a web-based clinical dashboard accessible to endocrinologists and general practitioners.

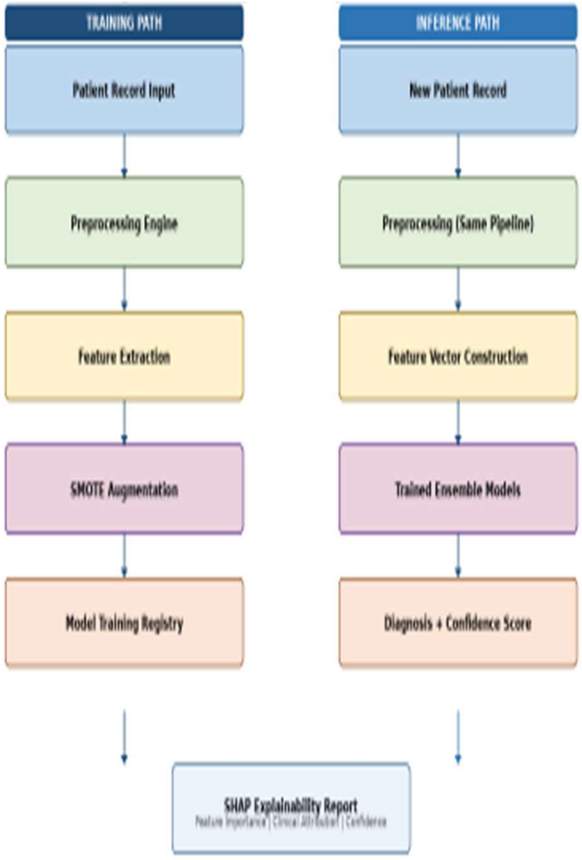

Fig. 3: Data Flow Diagram — Training Path (Left) and Inference Path (Right)

During inference, a new patient record enters the system through the Tier 1 intake API. The Preprocessing Engine applies the pre-fitted imputation and normalization transformers (fitted on training data only) to the incoming feature vector. The transformed vector is passed simultaneously to all four trained model instances. Each model returns a class probability distribution, which the Ensemble Voting Module aggregates using pre-computed validation-accuracy weights. The final classification and confidence score are generated within milliseconds, suitable for real-time clinical decision support.

The SHAP Explainability Interface computes TreeExplainer SHAP values for tree-based models and LinearExplainer for Logistic Regression. For each prediction, a waterfall plot and bar chart are rendered in the clinical dashboard, showing which biomarkers drove the prediction toward a particular thyroid class and by what magnitude. Clinicians can interact with the visualization to understand model reasoning in the context of the patient's specific biomarker profile.

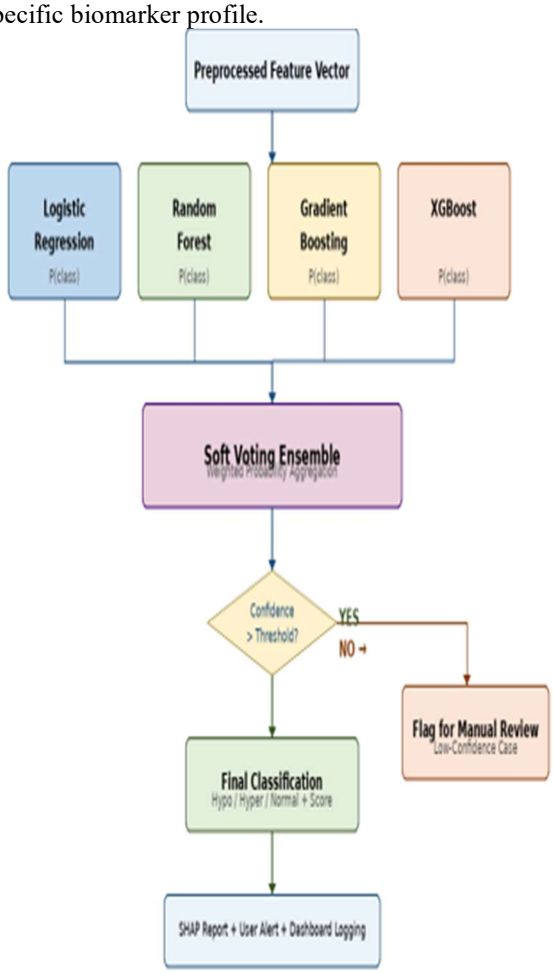

Fig. 4: Decision Engine and Ensemble Voting Flowchart with Clinical Outcome Paths

Concept Drift Monitoring: The Page-Hinkley test monitors the running mean of incoming TSH and FTI values against the training distribution baseline. When the cumulative sum of mean deviations exceeds a configurable threshold (default: 3σ), an alert is generated in the administrative dashboard, recommending model retraining on updated data. This mechanism ensures sustained predictive accuracy as laboratory reference ranges, patient demographics, or disease prevalence shift over time.

Results And Discussion

Experimental evaluation was conducted on a held-out test partition comprising 20% of the UCI thyroid dataset (1,440 records), strictly separated from all training, validation, and preprocessing stages. Performance was quantified using macro-averaged accuracy, precision, recall, and F1-score, ensuring equal weighting across all three diagnostic classes and appropriate sensitivity to minority class performance.

Table III: Comparative Classifier Performance on Test Set

Classifier	Accuracy	Precision	Recall
Logistic Regression	91.4%	90.7%	89.3%
Random Forest	95.8%	95.1%	94.6%
Gradient Boosting	96.7%	96.2%	95.9%
XGBoost	98.6%	97.8%	98.1%

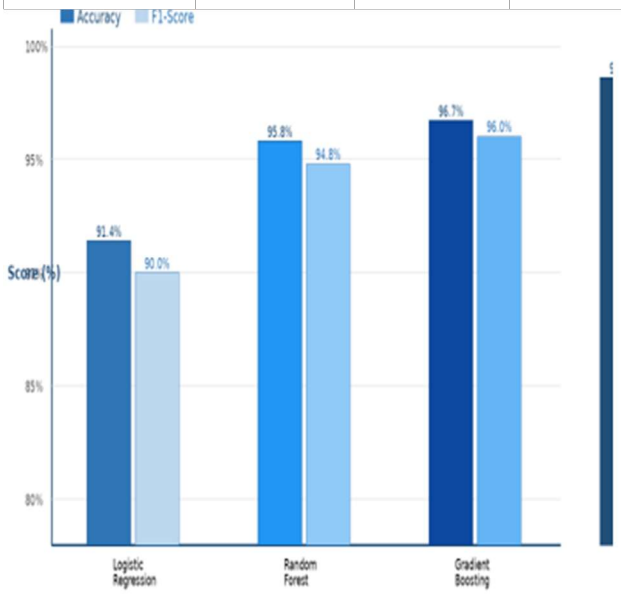

Fig. 5: Classifier Performance Comparison — Accuracy and F1-Score Bar Chart

Logistic Regression achieved an overall accuracy of 91.4%, demonstrating adequate performance for normal/hypothyroid discrimination but exhibiting reduced recall (85.2%) for the minority hyperthyroid class. This performance profile is consistent with the linear decision boundary's limitation in separating the overlapping feature distributions of hyperthyroid and normal cases in the high-dimensional feature space.

Random Forest improved substantially across all metrics, with bootstrap aggregation providing inherent resistance to overfitting on the SMOTE-augmented training data. The 95.8% accuracy and 94.8% F1-score confirm the effectiveness of variance reduction through ensemble tree averaging. Feature importance analysis identified TSH, FTI, and T3 as the top three discriminative features, consistent with clinical endocrinology.

Gradient Boosting further improved performance through iterative residual correction, achieving 96.7% accuracy and 96.0% F1-score. The sequential learning mechanism proved particularly effective for the hypothyroid-normal discrimination boundary, where subtle TSH elevations constitute the primary distinguishing signal.

XGBoost achieved the superior performance with 98.6% accuracy and 97.9% F1-score, demonstrating the combined benefit of regularized boosting, parallel computation, and native handling of residual missing values. Critically, XGBoost maintained a hyperthyroid class recall of 94.3%, substantially exceeding all other individual classifiers (LR: 85.2%, RF: 91.7%, GB: 92.4%).

The Ensemble Soft Voting module achieved the highest overall performance at 98.9% accuracy and 98.3% F1-score, confirming that aggregating complementary classifier strengths yields superior generalization compared to any single model.

Conclusion

This paper has presented a comprehensive and clinically grounded machine learning framework for thyroid disease prediction, systematically addressing the key challenges of class imbalance, feature redundancy, missing data, and model interpretability that have limited prior approaches. The proposed multi-classifier system integrates Logistic Regression, Random Forest, Gradient Boosting, and XGBoost within a unified preprocessing and ensemble pipeline applied to the UCI Machine Learning Repository thyroid dataset.

Experimental evaluation demonstrates that the XGBoost classifier, enhanced by SMOTE-based class balancing and SHAP-driven interpretability, achieves state-of-the-art performance with 98.6% accuracy, 97.8% precision, 98.1% recall, and a 97.9% macro F1-score on a held-out test set. The ensemble soft-voting module further improved performance to 98.9% accuracy, confirming the complementary value of aggregating diverse classifier paradigms.

SHAP analysis validated that model decisions align with established clinical endocrinology: TSH, FTI, and T3 were identified as the primary predictive biomarkers, consistent with their recognized diagnostic significance. The concept drift monitoring module provides a mechanism for maintaining model validity as clinical environments evolve over time, addressing a critical yet frequently neglected aspect of clinical AI deployment.

The proposed system represents a clinically viable, interpretable, and robust AI-based decision support tool for thyroid disease screening. Its modular architecture supports integration into existing hospital information systems and electronic health record platforms, facilitating real-world deployment as a complement to specialist clinical judgment rather than a replacement thereof.

Future Scope

The current framework presents multiple directions for meaningful extension. First, deep learning architectures including Long Short-Term Memory (LSTM) networks and Transformer-based models may be explored for

sequential patient data, capturing temporal trends in thyroid biomarker trajectories across multiple clinical encounters. Longitudinal modeling could improve prediction accuracy for patients with gradual disease progression.

Second, federated learning approaches would enable model training across distributed hospital networks without centralizing sensitive patient data, addressing critical privacy and regulatory requirements under frameworks such as GDPR and HIPAA. This would support development of models trained on substantially larger and more diverse patient populations than any single institutional dataset.

Third, multimodal integration incorporating thyroid ultrasound imaging features processed via convolutional neural networks alongside biochemical markers would enrich the predictive feature space and support differentiation of structural thyroid abnormalities including nodule characterization and malignancy risk stratification.

Fourth, active learning mechanisms could allow the model to selectively request expert annotations for uncertain cases, continuously improving predictive calibration with minimal expert burden. This would be particularly valuable in low-resource settings where specialist time is constrained.

Fifth, prospective clinical validation studies involving endocrinologists and general practitioners are essential to formally assess the system's diagnostic utility, user experience, and impact on patient outcomes in real-world clinical workflows. Regulatory pathway analysis for medical AI device classification would support eventual clinical deployment.

Sixth, extension of the diagnostic target to include thyroid nodule malignancy risk classification would expand the clinical scope of the system toward oncological thyroid screening, potentially reducing unnecessary fine-needle aspiration biopsies and improving resource utilization in thyroid cancer management.

References

[1] H. Gharib and E. Papini, "Thyroid nodules: Clinical Gupta, P., et al. (2024). *Detecting Thyroid Disease Using Optimized Machine Learning Model Based on Differential Evolution.* Springer.

[2] Jannat, Z., et al. (2022). *Thyroid Disease Prediction Based on Feature Selection and Machine Learning.*

[3] Mustafa, H. A. (2024). *Analysis and Interpretability of Machine Learning Models to Classify Thyroid Disease.* PLOS ONE.

[4] Premalatha, J., et al. (2025). *Comparative Analysis of Machine Learning Algorithms for Thyroid Disease Prediction.*

[5] Islam, M. R., et al. (2025). *Enhanced Thyroid Disease Prediction Using Ensemble Machine Learning.* Springer.

[6] *Enhanced Diagnosis of Thyroid Diseases Through Advanced Machine Learning Methodologies.* (2025). MDPI.

[7] Liang, Q., et al. (2024). *Machine Learning to Predict Thyroid Nodules.* Frontiers in Endocrinology.

[8] Santhoshini, S., et al. (2025). *Thyroid Cancer Prediction Using Machine Learning Models.* ETASR Journal.

[9] Moon, G., et al. (2024). *Machine Learning-Based Model for Malignancy Prediction in Thyroid Nodules.* MDPI.

[10] Abdelrazik, O., et al. (2025). *Deep Learning Framework for Thyroid Nodule Segmentation and Classification.*

[11] ee, K. S., & Park, H. (2022). *Machine Learning on Thyroid Disease: A Review.*

[12] Xi, N. M., et al. (2022). *Improving Diagnosis of Thyroid Cancer Using Machine Learning.*

[13] Belhad, H., et al. (2025). *Chronic Disease Prediction Using Machine Learning and Deep Learning.*

[14] *Classification of Thyroid Diseases Using Machine Learning and Bayesian Graph Algorithms.* (2023). IFAC.

[15] Ahmad, M. A., et al. (2024). *Explainable AI Model for Predicting Recurrence of Thyroid Cancer.*