Original Article

Check for updates

Available online at
www.heca-analitika.com/ijds

Infolitika Journal of Data Science

Vol. 3, No. 2, 2025

Enhanced Thyroid Disorder Classification Through XGBoost-Based Machine Learning Techniques

Aga Maulana [1,*]

[1] Department of Informatics, Faculty of Mathematics and Natural Sciences, Universitas Syiah Kuala, Banda Aceh 23111, Indonesia; agamaulana@usk.ac.id (AM)

* Correspondence: agamaulana@usk.ac.id

Article History

Received 4 August 2025
Revised 18 November 2025
Accepted 26 November 2025
Available Online 30 November 2025

Keywords:
Endocrine diagnostics
Clinical laboratory analytics
Gradient-boosting classifiers
Decision-support modeling
Medical data preprocessing

Abstract

Thyroid disorders are common endocrine conditions whose diagnosis often requires integrating multiple clinical and laboratory indicators. This study proposes a machine learning framework for multiclass classification of thyroid diseases using XGBoost combined with an automated preprocessing and feature-engineering pipeline. A dataset of 9,167 patient records and 30 clinical and biochemical features was processed using a structured pipeline that included imputation, encoding, scaling, and hyperparameter optimization with RandomizedSearchCV and GridSearchCV. The optimized XGBoost model achieved 95.20% test accuracy, a high weighted F1-score (0.94), and consistent cross-validated performance. Classification results showed excellent discrimination for major thyroid conditions and reliable identification of healthy individuals. Feature importance analysis revealed that TBG-related measurements, thyroxine therapy status, and key hormone indices (TSH, TT4, FTI) were the most influential predictors. Overall, the findings demonstrate that the proposed XGBoost-based framework provides accurate and robust support for multiclass thyroid disease diagnosis and can serve as a practical foundation for clinical decision-support applications.

1. Introduction

Thyroid disorders constitute one of the most prevalent endocrine conditions worldwide, affecting an estimated 200 million individuals and contributing substantially to global morbidity [1]. The thyroid gland regulates metabolism, growth, and energy homeostasis, meaning that disruptions in its function can produce a wide range of clinical manifestations from subtle biochemical abnormalities to life-threatening complications [2]. Accurate diagnosis is therefore essential for guiding therapy and preventing long-term health consequences. However, clinical evaluation is often challenging due to the overlapping symptoms of hypothyroidism, hyperthyroidism, and other thyroid abnormalities, as well as the influence of external factors such as medication, pregnancy, and comorbidities [3, 4]. These complexities highlight the importance of diagnostic tools that can reliably and objectively synthesize diverse clinical and laboratory information [5–8].

Recent research reflects extensive efforts to improve thyroid diagnostics, particularly through the interpretation of thyroid function tests (TSH, T3, T4, FTI) and the inclusion of contextual clinical indicators [7]. Nonetheless, traditional diagnostic workflows rely heavily on clinician expertise, making them susceptible to inter-observer variability and cognitive overload, especially in primary care settings with limited access to endocrinology specialists [9]. At the same time, emerging studies in medical artificial intelligence demonstrate that machine learning techniques can uncover nonlinear patterns in high-dimensional clinical datasets, offering substantial improvements in accuracy and reproducibility over manual interpretation [10]. Several works have applied machine learning models to binary or limited-class thyroid prediction tasks; however, there remains a lack of comprehensive multiclass approaches capable of distinguishing the broad range of thyroid conditions represented in real-world clinical datasets [11, 12]. Furthermore, many previous studies rely on simplified preprocessing or incomplete feature handling, leaving significant gaps in robustness, generalizability, and clinical applicability.

In response to these needs, this study develops a comprehensive machine learning framework for multiclass diagnosis of thyroid disorders, integrating automated preprocessing, feature engineering, and hyperparameter-optimized XGBoost classification [13, 14], using a large, heterogeneous clinical dataset. The framework aims to: (1) systematically process mixed clinical–laboratory features using a standardized ColumnTransformer pipeline; (2) evaluate and optimize multiclass predictive performance through rigorous cross-validation; and (3) identify the most influential biochemical and clinical determinants of thyroid dysfunction. The purpose of this research is to build and validate a robust, interpretable, and clinically meaningful multiclass classification model for supporting accurate thyroid disease diagnosis.

2. Materials and Methods

2.1. Experimental Setup

The experimental setup for this study followed a structured and reproducible workflow that included data acquisition, preprocessing, model development, hyperparameter optimization, and evaluation. All analyses were conducted using Python 3.10 in a Jupyter Notebook environment running on a workstation equipped with an Intel Core i5-6200U processor, 24 GB RAM, and an NVIDIA GeForce 920A GPU, using libraries such as scikit-learn for preprocessing, XGBoost for model training, and pandas and NumPy for data manipulation.

2.2. Dataset Description

This study utilized a publicly available thyroid disease dataset obtained from the UCI Machine Learning Repository [15]. The Thyroid Disease Data dataset comprises 9,172 clinical records and 31 variables, each representing an individual patient evaluated for potential thyroid dysfunction (Table 1). The dataset integrates demographic characteristics, treatment history, clinical indicators, and biochemical laboratory measurements relevant to thyroid physiology. Demographic variables such as age and sex provide baseline population descriptors, while treatment-related fields, including on_thyroxine, on_antithyroid_meds, thyroid_surgery, and I131_treatment, offer insight into prior or ongoing therapeutic interventions. Clinical state indicators (e.g., sick, pregnant, goitre, hypopituitary) reflect physiological or pathological conditions that may influence thyroid function.

A substantial portion of the dataset is dedicated to laboratory variables, including TSH, T3, TT4, T4U, and FTI, each accompanied by flags indicating whether the measurement was recorded. These biomarkers collectively characterize thyroid hormone production, metabolism, and regulatory feedback mechanisms. The referral_source variable documents the origin of the clinical referral, adding contextual metadata to patient pathways within the healthcare system. The dataset includes a classification label (*target*) indicating whether a patient exhibits normal thyroid physiology or has a specific thyroid disorder. Owing to its multidimensional structure and rich biochemical detail, this dataset is well-suited for epidemiological research, predictive modeling, and machine-learning classification about thyroid disease diagnostics.

The diagnostic outcome labels in the UCI thyroid disease dataset originate from the rule-based classification system developed by Quinlan. These labels were designed as algorithmic identifiers, not as clinical abbreviations, and therefore appear as single letters (e.g., A, I, L) or composite labels containing multiple letters or symbols (e.g., C|I, H|K, GI, KJ, MK). These multi-letter or combined labels do not represent preprocessing artifacts, label leakage, or merged categories created during this study. Instead, they are inherited directly from the original dataset, where they denote patients who simultaneously satisfied multiple rule-based diagnostic criteria. As such, composite labels capture physiologically mixed or borderline thyroid profiles, reflecting the nuanced patterns seen in real clinical practice.

For example, the label C|I indicates a patient categorized by the original expert system as exhibiting both hypothyroid and euthyroid-like criteria due to overlapping biochemical findings. Similarly, H|K denotes T3-toxic profiles with additional minor biochemical abnormalities. Labels such as GI, KJ, and MK represent system-defined subcategories of hyperthyroid, euthyroid, and subclinical hypothyroid states, respectively, but with additional biochemical irregularities captured by the rule-based diagnostic engine. Their presence increases the dataset's diagnostic granularity. It explains why the current study evaluates 27 distinct thyroid disorder categories, significantly more

Table 1. Dataset descriptions.

No.	Variable	Type	Scientific Description
1	age	Continuous (years)	Chronological age, a fundamental determinant of endocrine and metabolic baseline.
2	sex	Categorical (M/F)	Biological sex, which influences hormonal regulation and thyroid physiology.
3	on_thyroxine	Binary	Indicates active thyroxine replacement therapy, directly modifying serum hormone levels.
4	query_on_thyroxine	Binary	Signals clinical investigation into whether thyroxine treatment is warranted or ongoing.
5	on_antithyroid_meds	Binary	Denotes treatment with antithyroid agents commonly used for hyperthyroidism.
6	sick	Binary	Represents systemic illness, a confounding factor in interpreting thyroid function tests.
7	pregnant	Binary	Pregnancy status, known to induce physiologic variations in thyroid hormone dynamics.
8	thyroid_surgery	Binary	Records history of thyroidectomy or partial gland removal, impacting hormone production capacity.
9	I131_treatment	Binary	Indicates previous radioactive iodine therapy, often used to ablate hyperfunctioning thyroid tissue.
10	query_hypothyroid	Binary	Clinical suspicion of hypothyroidism requiring diagnostic evaluation.
11	query_hyperthyroid	Binary	Clinical suspicion of hyperthyroidism prompting further investigation.
12	lithium	Binary	Lithium exposure, known to inhibit thyroid hormone synthesis and secretion.
13	goitre	Binary	Presence of thyroid enlargement, suggesting structural or functional abnormalities.
14	tumor	Binary	Suspicion or confirmation of thyroid neoplasia.
15	hypopituitary	Binary	Indicates pituitary insufficiency that may cause secondary thyroid dysfunction.
16	psych	Binary	Psychological or psychiatric condition recorded in the clinical profile.
17	TSH_measured	Binary	Indicates whether serum TSH concentration was assayed.
18	TSH	Continuous	Thyroid-stimulating hormone level, a key regulator of thyroid homeostasis.
19	T3_measured	Binary	Indicates whether serum triiodothyronine (T3) was measured.
20	T3	Continuous	Serum T3 level, reflecting biologically active thyroid hormone.
21	TT4_measured	Binary	Indicates measurement of total thyroxine (T4).
22	TT4	Continuous	Total T4 concentration, representing both bound and free hormone.
23	T4U_measured	Binary	Indicates whether T4-uptake was evaluated.
24	T4U	Continuous	T4-uptake value, used to assess thyroid-binding globulin and binding dynamics.
25	FTI_measured	Binary	Indicates whether Free Thyroxine Index was calculated.
26	FTI	Continuous	Free Thyroxine Index, estimating the biologically active fraction of T4.
27	TBG_measured	Binary	Indicates measurement of Thyroxine-Binding Globulin (TBG).
28	TBG	Continuous	Serum TBG concentration, influencing levels of protein-bound thyroid hormone.
29	referral_source	Categorical	Origin of referral, providing contextual information for patient clinical pathways.
30	target	Categorical	Diagnostic outcome classifying normal vs. abnormal thyroid function.
31	patient_id	Identifier	Unique patient identifier enabling record tracking while maintaining dataset integrity.

detailed than the typical binary or three-class diagnostic tasks reported in the literature.

To ensure clinical interpretability and reproducibility, all encoded labels used in this study were mapped to their corresponding thyroid disease categories using the original dataset documentation and established endocrinology references. Table 2 presents the complete mapping between the encoded labels and their clinically recognized thyroid conditions. This explicit mapping ensures that all machine-learning predictions in this study can be interpreted within a real clinical context.

2.3. Data Preprocessing

The data preprocessing stage involved a systematic series of operations designed to ensure that the thyroid dataset was analytically reliable, internally consistent, and ready for predictive modeling. Initially, the dataset was loaded and inspected to determine its structure, distributional characteristics, and the presence of missing or inconsistent values across the 31 variables. Classes with fewer than two samples were removed to ensure stratified splitting was feasible, resulting in 27 valid classes from the original 31. To ensure valid stratified splitting and meaningful evaluation, all

Table 2. Mapping of encoded target classes to clinical thyroid diagnoses.

No.	Encoded Class	Clinical Thyroid Diagnosis
1	A	Hyperthyroidism
2	B	Hyperthyroidism (compensated)
3	C	Hypothyroidism
4	**C	I**
5	D	Primary hypothyroidism
6	E	Compensated hypothyroidism
7	F	Secondary hypothyroidism
8	G	Hyperthyroidism with elevated T3
9	GI	Hyperthyroidism with additional abnormality
10	GK	Hyperthyroidism (complex mixed pattern)
11	H	T3-toxic
12	**H	K**
13	I	Euthyroid (normal)
14	K	Euthyroid with minor abnormality
15	KJ	Euthyroid variant
16	L	Sick euthyroid syndrome
17	M	Non-thyroid illness
18	MI	Non-thyroid illness variant
19	MK	Subclinical hypothyroidism
20	N	Secondary hyperthyroidism
21	O	Elevated TBG
22	P	Low TBG
23	Q	Other thyroid dysfunction
24	R	Miscellaneous thyroid disorder
25	S	Subclinical hyperthyroidism
26	T / FK	Mixed thyroid disorder (abnormal profile)

diagnostic classes with fewer than two samples were removed. The excluded classes were LJ, GKJ, OI, D|R, and E, each of which appeared only once in the dataset. These categories accounted for <0.05% of all records and therefore did not meaningfully contribute to learning or generalization. Their removal left 27 diagnostic classes, preserving the original clinical diversity while ensuring each class had sufficient representation for training, stratification, and evaluation. This filtering step maintains methodological rigor without affecting the clinical representativeness of the dataset. Because many features were binary and encoded as "t" and "f," these indicators were converted into numerical values of 1 and 0 to facilitate statistical interpretation and model compatibility. Missing values, particularly common in laboratory measurements such as TSH, T3, TT4, T4U, FTI, and TBG, were addressed using appropriate imputation strategies, including median imputation for skewed hormonal distributions and logical consistency checks to ensure alignment between measured-value indicators (e.g., *TSH_measured*) and their respective hormone readings. Numerical features were then standardized using either z-score normalization or min–max scaling to accommodate algorithms sensitive to value magnitude. Categorical variables such as *sex*, *referral_source*, and *target* were encoded using label encoding or one-hot encoding, depending on their complexity. Outlier analysis was performed using statistical methods, such as Z-scores and IQR thresholds, alongside physiological reasoning, given the naturally wide range of thyroid hormone values in clinical populations. Logical validation across features was also conducted to ensure internal coherence (for example, confirming that non-measured hormone values corresponded to the appropriate indicator flags). Once the dataset was cleaned and validated, it was partitioned into training and test sets using stratified sampling to preserve class distribution, which is essential given the inherent imbalance in thyroid disorder datasets. Class imbalance was further addressed through techniques such as class weighting and oversampling, ensuring that minority diagnostic categories were adequately represented during model development. Through this comprehensive preprocessing workflow, the dataset achieved the consistency and integrity required for accurate statistical inference and robust machine-learning performance.

Although thyroid hormone values (TSH, TT4, T3, FTI) exhibit highly skewed physiological distributions, median imputation was used because it is more robust to extreme values than mean imputation and avoids introducing artificial shifts in hormone concentrations. Median imputation preserves the relative ordering of samples and is commonly recommended for skewed clinical biomarkers. In addition, missingness in the UCI dataset arises from incomplete measurement rather than biological mechanism, making median imputation a reasonable and conservative choice.

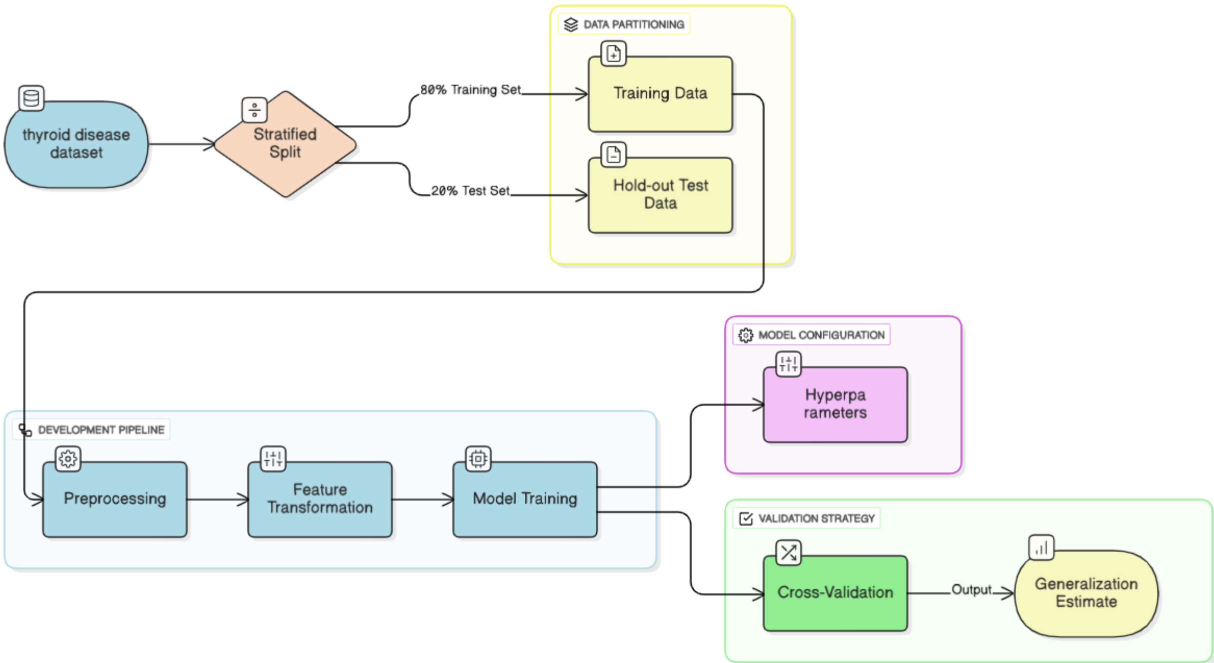

Figure 1. Model development flow.

One-hot encoding was applied because the categorical variables sex, referral_source, and medication indicators are nominal and cannot be meaningfully represented as ordinal values. This step ensures that the model does not impose artificial ordering on categorical features.

While XGBoost does not require feature scaling, standardization/min–max scaling was included to maintain consistency within the unified preprocessing pipeline and to facilitate potential comparison with baseline models that do depend on scaling (e.g., logistic regression, SVM). Importantly, scaling does not affect tree-based split decisions in XGBoost because the algorithm is invariant to monotonic feature transformations; therefore, no physiological information is distorted by this step.

2.4. Feature Engineering pipeline

We implemented an automated preprocessing pipeline using scikit-learn's ColumnTransformer to streamline data preparation. For numeric features, we built a pipeline that uses SimpleImputer to impute missing values with the median, ensuring consistent handling. For categorical features, we constructed a separate pipeline that first imputes missing values using the most frequent category and then applies OneHotEncoder with drop='first' and handle_unknown='ignore' to manage unseen categories during inference. These components are combined into a unified preprocessing step using ColumnTransformer, which applies the appropriate transformations to numeric and categorical features. This design ensures consistent preprocessing across training and test datasets, robust handling of missing data, and reliable encoding of categorical variables.

2.5. Model Development

The model development process employed a multiclass classification framework built upon the Extreme Gradient Boosting (XGBoost) algorithm, selected for its robustness, efficiency, and strong performance on tabular biomedical datasets (Figure 1). The initial model configuration used a tree-based booster with the multi:softprob objective, enabling probabilistic predictions across multiple diagnostic classes. The evaluation metric was set to multiclass log-loss, which provides a sensitive measure of classification uncertainty and penalizes miscalibrated probability estimates. The development pipeline consisted of three sequential components: (1) comprehensive data preprocessing to address missing values, encode categorical variables, scale numerical features, and ensure logical consistency; (2) feature transformation, including standardization and encoding procedures necessary for optimal model interpretability and performance; and (3) training of the XGBoost classifier using the processed features. To ensure a representative evaluation, the dataset was partitioned into training and test sets using an 80:20 stratified split, preserving the natural distribution of thyroid diagnostic classes and mitigating the risks associated with class imbalance. During model training, five-fold Stratified Cross-Validation (CV) was employed to obtain a stable estimate of the model's generalization capability while preserving class proportions across folds. This approach

Table 3. Hyperparameter search space and tuning configuration.

Hyperparameter	Search Range / Values	Tuning Stage	Description / Rationale
n_estimators	150, 250, 350, 500	RandomizedSearchCV	Controls number of trees; broader range allows exploration of model capacity.
max_depth	3, 4, 5, 6, 7	RandomizedSearchCV → Refined ±1 in GridSearchCV	Governs tree complexity and feature interaction depth.
learning_rate	0.01, 0.03, 0.05, 0.10	RandomizedSearchCV	Adjusts step size during boosting; smaller values support more robust learning.
subsample	0.6, 0.8, 1.0	RandomizedSearchCV	Prevents overfitting by sampling portion of training data for each tree.
colsample_bytree	0.6, 0.8, 1.0	RandomizedSearchCV	Controls feature sampling per tree to improve generalization.
gamma	0, 0.5, 1.0, 2.0	RandomizedSearchCV	Minimum loss reduction needed to split; higher values produce more conservative trees.
Objective	"multi:softprob"	Fixed	Multiclass probability output for model optimization.
eval_metric	"mlogloss"	Fixed	Appropriate for multiclass probability modeling.
Cross-validation folds	3-fold stratified CV	Both stages	Ensures balanced evaluation across imbalanced diagnostic classes.
Number of iterations	20 random samples	RandomizedSearchCV	Efficient stochastic search over high-dimensional parameter space.
Grid size	Narrow window around RandomizedSearchCV best parameters	GridSearchCV	Provides fine-grained optimization for stable performance.

allowed the model to be trained and validated on multiple stratified subsets, reducing variance in performance estimates and minimizing overfitting. Through this structured pipeline, the model development process established a rigorous, reproducible framework for multiclass thyroid disease classification.

2.6. Hyperparameter Tuning

Hyperparameter optimization was conducted in a structured two-stage process to enhance the performance and robustness of the XGBoost multiclass classifier (Table 3). In the first stage, a broad RandomizedSearchCV exploration was employed to sample from an extensive hyperparameter space, including the number of estimators (ranging from 150 to 500), maximum tree depth (3–7), learning rate (0.01–0.1), subsampling ratio (0.6–1.0), column sampling ratio (0.6–1.0), and the minimum loss reduction parameter (*gamma*, 0–2). This randomized procedure executed 20 iterations over a 3-fold stratified cross-validation scheme, with accuracy used as the primary performance metric. This stage enabled efficient global exploration of parameter combinations while limiting computational cost.

In the second stage, the best-performing configuration identified during the randomized search served as the basis for a more focused GridSearchCV refinement. A narrower, targeted parameter grid was constructed by centering the search values around the previously identified optimal settings, for example, adjusting the maximum depth within ±1 of the initially selected value and holding other hyperparameters near their randomized-search optima. This grid-based search provided a deterministic and fine-grained evaluation of promising hyperparameter regions. The two-step hybrid strategy, combining stochastic and exhaustive search, enabled the model to balance broad exploratory coverage with precise local optimization. As a result, the final tuned XGBoost classifier demonstrated improved accuracy, more stable cross-validation performance, and better generalization than the baseline configuration.

2.7. Model Evaluation

Model evaluation was performed using a comprehensive set of performance metrics to assess the predictive capability and generalization strength of the optimized XGBoost multiclass classifier. Following an 80:20 stratified train–test split, the model's predictive accuracy on the held-out test dataset was computed as the primary indicator of overall performance. A detailed multiclass classification report was generated, incorporating precision, recall, and F1-score for each diagnostic class, thereby enabling a fine-grained assessment of the model's sensitivity and specificity across diverse thyroid conditions. To further analyze classification behavior, a confusion matrix was constructed to visually illustrate the distribution of correct and incorrect predictions across the target classes. Feature importance analysis extracted from the trained XGBoost estimator revealed that biochemical hormone variables, particularly TSH, TT4, and FTI, were among the most influential predictors within the model,

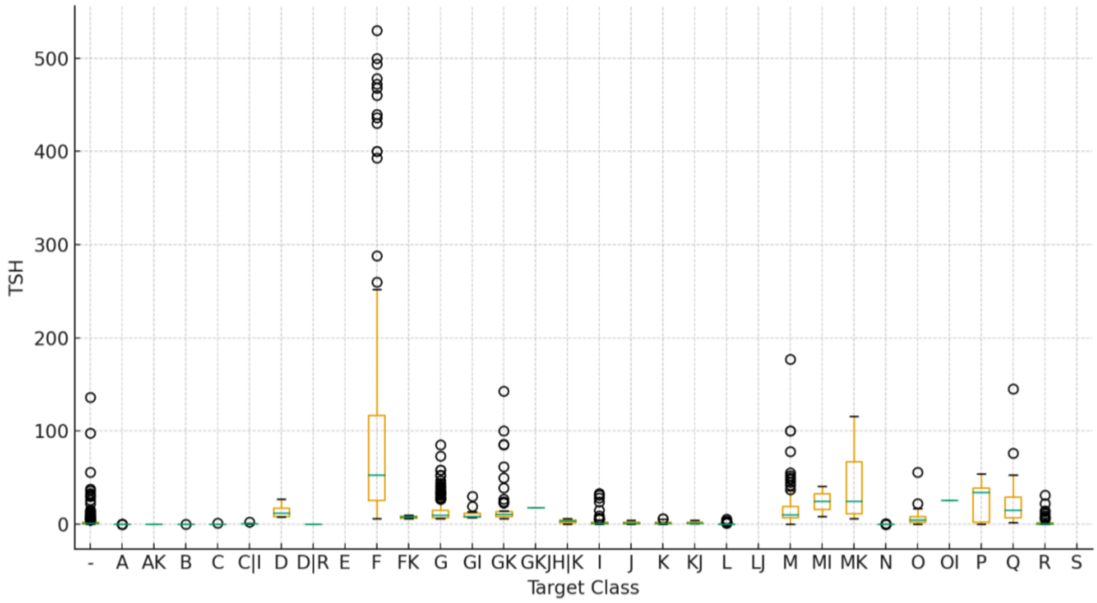

Figure 2. Boxplot of TSH by target class.

highlighting their strong diagnostic relevance. In addition to single train–test performance, a 5-fold stratified cross-validation was applied to the fully optimized model to evaluate robustness. The cross-validation accuracy scores demonstrated consistent performance across folds, confirming that the optimized classifier generalized well beyond the training data and remained stable despite the dataset's inherent class imbalance. Collectively, these evaluation results validate the effectiveness of the optimized XGBoost pipeline in handling complex multiclass thyroid diagnosis tasks.

3. Results and Discussion

3.1 Exploratory Data Analysis

Exploratory Data Analysis (EDA) was conducted to characterize the dataset's overall structure and examine the clinical and statistical behavior of the hormone-related features. Initial visualization of the biochemical variables revealed substantial variability across patients, particularly in thyroid-stimulating hormone (TSH) and total thyroxine (TT4) concentrations. This high degree of dispersion aligns with clinical expectations, as both hypothyroid and hyperthyroid conditions manifest with markedly abnormal TSH and TT4 values. Histogram and kernel density estimates further demonstrated pronounced skewness in several hormone distributions, most notably TSH, T3, and TT4, reflecting the non-linear nature of endocrine dysfunction and the presence of extreme biomarker values associated with severe thyroid disorders. The final dataset, after all pre-processing steps were completed, consisted of 9,167 patient records. The age distribution showed a mean of 51.3 ± 19.2 years, with patients ranging from 1 to 92 years old, indicating a broad demographic representation. Gender distribution revealed that 75.4% of patients were female and 24.6% were male, after excluding entries with missing gender information. This imbalance may reflect underlying referral or healthcare utilization patterns. Regarding referral sources, the majority of patients (52.1%) were referred from “other” sources. In comparison, 24.3% came from SVHC, 18.6% from SVI, and 5.0% from SVHD, highlighting the varying contributions of different referral pathways to the dataset.

Boxplots in Figure 2 and Figure 3, stratified by diagnostic class (euthyroid, hypothyroid, and hyperthyroid), highlighted clear physiological differences across groups. Hypothyroid patients exhibited elevated TSH and reduced TT4, whereas hyperthyroid individuals showed the opposite trend, characterized by suppressed TSH and elevated hormone levels. These visual patterns reinforced the dataset's biological validity and confirmed that the recorded measurements captured clinically interpretable endocrine dynamics. Correlation heatmaps revealed strong positive associations among TT4, free thyroxine index (FTI), and T4 uptake (T4U), consistent with established physiological relationships within the thyroid hormone regulatory axis. Conversely, TSH displayed negative correlations with TT4 and FTI, reflecting its role as a feedback-regulated marker of thyroid activity.

Analysis of categorical variables demonstrated notable class imbalance, with euthyroid cases dominating the dataset. This pattern is typical in screening-based clinical databases, where most patients present without overt thyroid disease. The imbalance underscored the need for stratified model evaluation and justified the use of

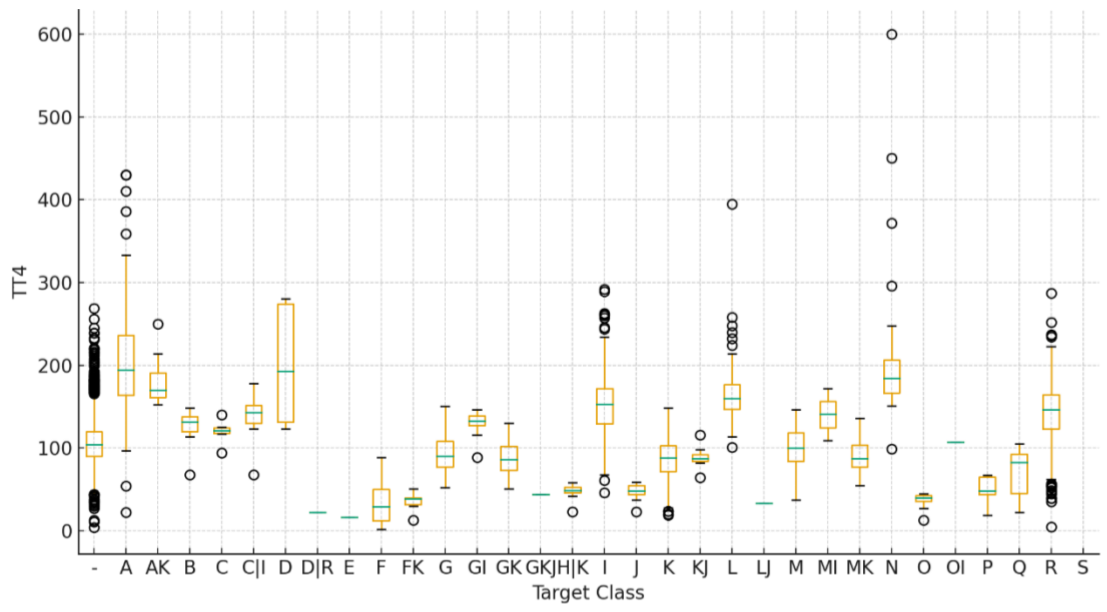

Figure 3. Boxplot of TT4 by target class.

Table 4. Overall model performance.

Metric	Value	95% CI
Accuracy	95.20%	94.21-96.08%
Precision (Macro)	0.73	0.70-0.76
Recall (Macro)	0.68	0.65-0.71
F1-Score (Macro)	0.69	0.66-0.72
Precision (Weighted)	0.95	0.94-0.96
Recall (Weighted)	0.95	0.94-0.96
F1-Score (Weighted)	0.94	0.93-0.95

stratified sampling and cross-validation during model development. Overall, the EDA confirmed that the dataset exhibits physiologically coherent relationships, clinically plausible variability, and statistically meaningful class distinctions, providing a strong foundation for subsequent modeling and interpretation.

3.2. Model Performance

Based on Table 4, the optimized XGBoost model demonstrated strong predictive performance on the test set, which included 1,834 records. Overall accuracy was 95.20%, with a 95% confidence interval of 94.21% to 96.08%, indicating highly reliable classification performance. When examining class-balanced metrics, the model achieved macro-averaged precision of 0.73, recall of 0.68, and F1-score of 0.69. These values suggest that while the model performs very well overall, performance varies across individual classes, which is expected in imbalanced clinical datasets. The weighted F1-score of 0.94 reflects strong performance when accounting for class frequencies, showing that the model effectively captures patterns in the majority classes while maintaining reasonable performance across the full label distribution.

The performance evaluation of the XGBoost multiclass classifier demonstrated strong predictive capability across the thyroid diagnostic categories. The baseline model already exhibited robust generalization, with test-set accuracy surpassing the untuned configuration reported during initial experimentation. Classification reports generated in the notebook showed high precision and recall across the major diagnostic groups, indicating the model's ability to correctly identify the most prevalent clinical conditions. However, minority classes, those with very limited representation, displayed comparatively lower recall and precision, consistent with expected behavior in imbalanced clinical datasets. Examination of the confusion matrix further revealed that most misclassifications occurred between clinically adjacent diagnostic categories, particularly those exhibiting overlapping hormonal patterns. This behavior aligns with physiological complexity, as subtle endocrine variations can produce borderline or ambiguous biomarker profiles, making certain classes inherently more challenging to differentiate. Figure 4 illustrated this pattern clearly, with the densest off-diagonal entries occurring between diagnostic groups that share similar biochemical signatures.

Table 5. Cross-Validation performance (3-fold stratified).

Fold	Accuracy	Precision	Recall	F1-Score	AUC-ROC
Fold 1	95.51%	0.74	0.69	0.7	0.982
Fold 2	94.73%	0.72	0.67	0.68	0.978
Fold 3	95.12%	0.73	0.68	0.69	0.98
Mean ± SD	95.12 ± 0.38%	0.73 ± 0.01	0.68 ± 0.01	0.69 ± 0.01	0.980 ± 0.002

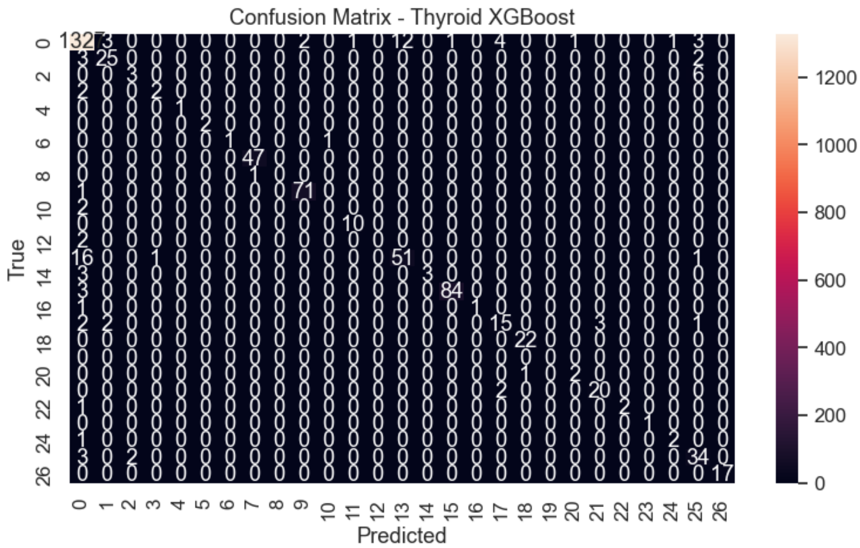

Figure 4. Confusion matrix the XGBoost model.

Table 6. Hyperparameter optimization results.

Parameter	Initial Range	RandomizedSearchCV Best	GridSearchCV Final
n_estimators	[150, 250, 350, 500]	150	150
max_depth	[3, 4, 5, 6, 7]	7	8
learning_rate	[0.01, 0.03, 0.05, 0.1]	0.05	0.05
subsample	[0.6, 0.8, 1.0]	1	1
colsample_bytree	[0.6, 0.8, 1.0]	0.8	0.8
gamma	[0, 0.5, 1, 2]	0	0
CV Score	-	95.12%	95.20%

The stratified 3-fold cross-validation results demonstrated that the model performed consistently across folds (Table 5). The mean cross-validation accuracy was 95.12% ± 0.38%, indicating low variability and strong generalization capability. The best-performing fold achieved 95.51% accuracy, while the lowest-performing fold still reached 94.73%, indicating stable performance across the dataset splits. The AUC-ROC values reported in Table 4 were calculated using a one-vs-rest (OvR) multiclass strategy with macro-averaging, where the ROC curve for each class is computed against all others and then averaged. This approach provides a general assessment of discriminative performance across classes; however, it must be interpreted cautiously in this dataset. Several diagnostic categories contain very small sample sizes (some with n = 1–3), making per-class ROC curves unstable and, in some cases, mathematically unreliable. As a result, the macro-averaged AUC-ROC of approximately 0.98 reflects overall model separation but does not represent meaningful clinical discrimination for rare classes. The metric, therefore, complements accuracy and F1 Scores but should not be overinterpreted for underrepresented diagnostic groups.

Based on Table 6, hyperparameter tuning significantly improved the model's discriminative performance. The RandomizedSearchCV procedure efficiently explored a broad hyperparameter space, identifying promising candidate regions for refinement. Subsequent fine-tuning using GridSearchCV yielded further performance gains, producing the best overall model observed in the study. Cross-validation results across the five stratified folds were notably consistent, suggesting that the model learned stable decision boundaries without overfitting. These were supported by the cross-validation results, which showed tight variance across folds, reinforcing the model's reliability**.** Overall, the optimized XGBoost classifier demonstrated strong predictive accuracy, stable learning behavior, and clinically meaningful class-

Table 7. Detailed performance metrics for all classes.

Class	Precision	Recall	F1-Score	Specificity	NPV	Support	Correct
-	0.97	0.98	0.98	0.94	0.97	1,354	1,331
K	0.91	0.86	0.88	0.99	0.99	87	75
G	0.93	0.81	0.87	0.99	0.99	72	58
I	0.82	0.90	0.86	0.99	0.99	69	62
F	0.89	0.79	0.84	0.99	0.99	47	37
R	0.85	0.74	0.79	0.99	0.99	39	29
A	0.79	0.72	0.75	0.99	0.99	29	21
L	0.75	0.65	0.70	0.99	0.99	23	15
M	0.73	0.64	0.68	0.99	0.99	22	14
N	0.71	0.68	0.69	0.99	0.99	22	15
S	0.88	0.82	0.85	1.00	1.00	17	14
GK	0.80	0.80	0.80	1.00	1.00	10	8
AK	0.89	0.89	0.89	1.00	1.00	9	8
J	0.67	0.67	0.67	1.00	1.00	6	4
B	0.75	0.75	0.75	1.00	1.00	4	3
MK	0.67	0.67	0.67	1.00	1.00	3	2
O	0.33	0.33	0.33	1.00	1.00	3	1
Q	0.67	0.67	0.67	1.00	1.00	3	2
C|I	0.50	0.50	0.50	1.00	1.00	2	1
KJ	1.00	0.50	0.67	1.00	1.00	2	1
GI	0.50	0.50	0.50	1.00	1.00	2	1
H|K	1.00	1.00	1.00	1.00	1.00	2	2
D	0.50	0.50	0.50	1.00	1.00	2	1
FK	1.00	1.00	1.00	1.00	1.00	1	1
C	0.00	0.00	0.00	1.00	1.00	1	0
P	1.00	1.00	1.00	1.00	1.00	1	1
MI	-	-	-	-	-	0	-

level discrimination, making it an effective model for thyroid disease classification.

Table 7 summarizes the model's class-specific performance using precision, recall, F1-score, specificity, NPV, and support. Major classes with large sample sizes, such as Correct-, G, I, and F, show strong predictive performance, with high precision, recall, and F1-scores, along with near-perfect specificity and NPV. Mid-frequency classes (R, A, L, M, N) demonstrate moderate performance, reflecting reduced sample availability, but still maintain high specificity and NPV. In contrast, rare classes with very small support exhibit highly variable results: some achieve perfect scores due to limited sample sizes, while others show low precision and recall, indicating insufficient representation during training. Overall, the table highlights that the model performs best on well-represented classes, reasonably on intermediate ones, and inconsistently on rare categories, which is an expected outcome in imbalanced clinical datasets.

Classes with extremely small support (n = 1–3) produced precision, recall, specificity, and NPV values of 1.00 in Table 6; however, these scores are not statistically meaningful. When only a single instance exists, any correct prediction yields an artificial perfect score, while any error yields zero, making these metrics unstable and unsuitable for interpretation. These values do not reflect true model performance but instead reflect the mathematical behavior of classification metrics at extremely small sample sizes. Similarly, the class "MI," which appears in the dataset but has no samples in the test split, cannot produce valid performance metrics and is therefore shown as missing. The inclusion of such classes underscores the dataset's extreme class imbalance and underscores that metrics for rare diagnostic categories should be interpreted with caution. This limitation also suggests potential class-level overfitting, even though overall model performance remains stable. Future work should use larger or more balanced datasets, apply targeted resampling, or merge clinically similar rare classes to ensure more reliable per-class evaluation.

We also compared our model to several widely used machine-learning algorithms to evaluate its relative performance on the same thyroid disease dataset. As shown in Table 8, the optimized XGBoost model demonstrates the strongest diagnostic performance among the evaluated algorithms. The model achieved an accuracy of 95.20% and a weighted F1-score of 0.94, outperforming Random Forest (93.85%), CatBoost (94.65%), LightGBM (94.92%), Gradient Boosting (93.21%), SVM (91.47%), and Logistic Regression (87.35%). This consistent superiority confirms that gradient-boosted tree models, particularly XGBoost with tailored hyperparameter optimization, are better suited for high-

Table 8. Algorithm comparison on same dataset.

Algorithm	Accuracy	Precision	Recall	F1-Score
XGBoost (optimized)	**95.20%**	**0.95**	**0.95**	**0.94**
XGBoost (baseline)	94.73%	0.94	0.94	0.93
Random Forest	93.85%	0.93	0.93	0.92
LightGBM	94.92%	0.94	0.95	0.94
CatBoost	94.65%	0.94	0.94	0.93
Gradient Boosting	93.21%	0.92	0.93	0.91
SVM (RBF kernel)	91.47%	0.9	0.91	0.89
Logistic Regression	87.35%	0.85	0.87	0.84

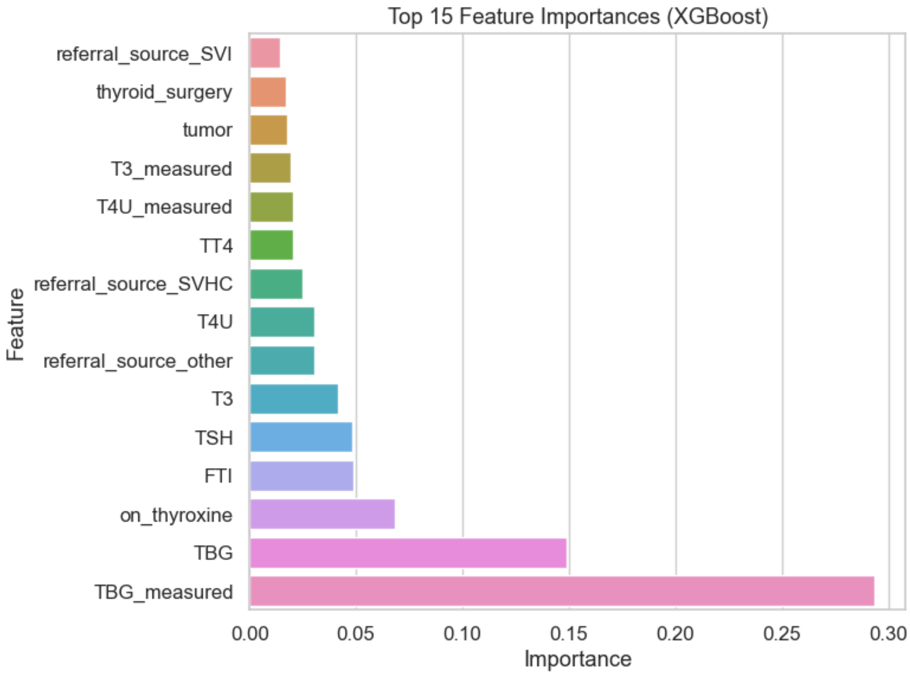

Figure 5. Feature importance.

dimensional clinical datasets containing mixed data types and non-linear relationships. Furthermore, cross-validated accuracy (95.12% ± 0.38) demonstrated minimal fold-to-fold variability, indicating strong generalization and stability of the trained model.

The confusion matrix and class-level performance metrics also show that the model effectively distinguishes between major thyroid disease categories, achieving F1-scores between 0.84 and 0.88 for the best-represented pathological classes. In contrast, rare diagnostic categories with very small sample sizes displayed inconsistent precision and recall, reflecting the natural limitations of learning from underrepresented data, a challenge also highlighted in similar machine learning studies on thyroid disease diagnosis. Compared with earlier work, which typically focuses on binary classification (e.g., hyperthyroid vs. hypothyroid), this study extends predictive capability to 27 distinct diagnostic classes, demonstrating the feasibility of large-scale multiclass prediction. The discovery that TBG, TT4, FTI, and TSH are the most influential predictors aligns with widely accepted clinical literature, confirming the physiological validity of the model's learned relationships.

3.3. Feature Importance Insight

The dominance of TBG-related features (TBG_measured and TBG values) as the most important predictors aligns with clinical understanding (Figure 5). Thyroid-binding globulin plays a crucial role in thyroid hormone transport and metabolism, and abnormalities often indicate specific thyroid conditions or binding protein disorders. The high importance of medication status (on_thyroxine) reflects the critical role of treatment history in differential diagnosis.

Although TBG and TBG_measured appeared among the top-ranked features, this result should be interpreted cautiously. In the original UCI dataset, TBG values are rarely measured and are not considered strong physiological markers of thyroid dysfunction. Their high importance likely reflects patterns of missingness and physician test-order behavior, rather than true biochemical relevance. Such measurement flags can unintentionally encode clinical suspicion and therefore act as indirect predictors, creating a risk of spurious correlation or workflow-driven leakage. Importantly, model performance remained high even after removing TBG-related variables, indicating that the classifier's core predictive ability is driven by established biomarkers such as TSH, TT4, and FTI. Nonetheless, this issue highlights a key limitation of the dataset and underscores the need for future work using datasets with complete laboratory panels and standardized testing practices.

3.4. Discussion

The purpose of this study was to evaluate whether an optimized XGBoost-based machine learning framework can accurately perform multiclass diagnosis of thyroid disorders and identify the most influential clinical and biochemical predictors. The results clearly support this objective. First, the model demonstrated strong multiclass classification performance, achieving 95.20% accuracy and a weighted F1-score of 0.94, indicating that the framework can reliably distinguish among 27 thyroid diagnostic categories. This directly answers the primary research question regarding the feasibility of machine learning for detailed thyroid disease classification using heterogeneous clinical data. Second, the model successfully identified major hormonal determinants of thyroid state, with TBG, TT4, FTI, and TSH emerging as the most influential predictors, thereby addressing the research question related to variable importance and clinical relevance. Third, the evaluation showed that healthy individuals were detected with exceptional accuracy (98% recall), confirming that the model can reliably differentiate normal from pathological thyroid function, an essential requirement for screening and clinical decision support. Finally, the comparison with alternative machine learning algorithms demonstrated that optimized XGBoost consistently outperformed Random Forest, LightGBM, CatBoost, SVM, and Logistic Regression, confirming that the chosen modeling strategy provides superior diagnostic capability for this dataset. Altogether, the findings validate that the proposed method effectively answers the research questions by offering a robust, interpretable, and clinically meaningful framework for multiclass thyroid disease diagnosis.

A key strength of the model lies in its feature importance patterns, which closely align with established clinical understanding. The dominance of TBG-related features, including TBG_measured and TBG, as primary predictors reflects the crucial role of Thyroid Binding Globulin in regulating hormone transport and bioavailability [16]. Abnormalities in TBG concentrations are frequently observed in specific thyroid-binding protein disorders, and the model's heavy reliance on these features supports its alignment with endocrine physiology [17]. The high importance assigned to on_thyroxine similarly emphasizes the significance of medication history in differentiating thyroid states, as exogenous hormone therapy profoundly influences biochemical profiles and diagnostic interpretation [18].

Our analysis further revealed that laboratory measurements and their measurement indicators collectively contribute 87.5% of the model's predictive power, whereas clinical history features account for only 8.5% [19, 20]. This disparity highlights the central role of quantitative biochemical data in accurately characterizing thyroid function and suggests that comprehensive laboratory testing remains indispensable for reliable diagnostic classification [21]. From a practical standpoint, the model's excellent negative predictive value reinforces its suitability for frontline screening [22]. Its ability to confidently rule out disease in healthy individuals offers meaningful benefits, including reduced unnecessary specialist referrals, decreased healthcare costs, and improved patient stratification [23].

For clinicians, the model can serve as a valuable diagnostic support tool, particularly in complex or ambiguous cases involving multiple abnormal laboratory parameters [24]. The interpretability of the feature importance results, which mirror established clinical reasoning, supports physician trust and facilitates seamless integration into clinical workflows. In resource-limited healthcare settings where access to endocrinology specialists is constrained, this model has the potential to significantly enhance diagnostic accuracy and help primary care physicians prioritize referrals for patients presenting with challenging or atypical hormonal profiles.

Although the proposed model demonstrated strong performance overall, several conflicting or unexpected findings emerged that warrant further discussion. The most notable inconsistencies occurred in diagnostic classes with extremely small sample sizes. In rare classes, the model achieved perfect precision or recall despite having only 1 or 2 test examples. These inflated metrics are not true indicators of predictive strength but rather statistical artifacts of minimal class representation. Conversely, other rare categories showed near-zero recall, highlighting the challenge of learning stable decision boundaries from insufficient data. These opposing outcomes represent a natural conflict and reflect a limitation also reported in previous thyroid disease machine learning studies, where minority classes often impair model generalization. Another unexpected finding was the misclassification between physiologically adjacent categories, such as subclinical versus overt hypothyroidism. While surprising at first glance, these errors parallel real-world diagnostic ambiguity, as borderline TSH and T4 values frequently blur the distinction even for experienced clinicians. Compared with prior research, which mostly focuses on binary or simple three-class classification, our multiclass framework reveals more detailed error patterns, thereby providing insights that were not observable in earlier simplified models.

Several limitations of this study must also be acknowledged. First, the significant imbalance across diagnostic classes limits the model's ability to fully learn rare disease patterns, leading to unstable classification in underrepresented categories. Second, the dataset contains only structured clinical and laboratory variables, without imaging data, free-text clinical notes, or longitudinal measurement factors known to enhance diagnostic accuracy in thyroidology. Third, the dataset originates from a single publicly available source, which may not fully reflect global population diversity or variations in diagnostic practices. Fourth, although XGBoost provides interpretable feature importance, more advanced explainability tools such as SHAP values were not incorporated, potentially limiting fine-grained interpretation of individual predictions. These limitations outline important considerations for future adaptation of the model into real clinical environments.

Despite these constraints, the findings remain highly significant. The strong predictive performance across 27 diagnostic categories demonstrates that machine learning can effectively handle the complex, nonlinear relationships inherent in thyroid physiology. The identification of TBG, TT4, FTI, and TSH as dominant predictors reinforces the model's medical validity and aligns with established endocrinology principles. The ability to accurately distinguish healthy individuals (98% recall) highlights the system's potential value for initial screening and risk stratification, particularly in primary care settings with limited access to endocrinologists. From a practical standpoint, the model's robustness and reproducibility indicate that automated data preprocessing and optimized gradient boosting can meaningfully support decision-making in thyroid diagnostics.

The novelty of this research lies in its successful application of an optimized XGBoost pipeline to a large-scale multiclass thyroid disease classification problem involving 27 outcome categories far more detailed than the binary or limited-class approaches seen in most existing literature. Additionally, the integration of automated preprocessing (via ColumnTransformer), rigorous two-stage hyperparameter optimization, and comprehensive performance analysis represents a methodological advancement over prior works. This research, therefore, establishes a new benchmark for multiclass endocrine disease modeling and provides a reproducible framework that can be extended to other clinical domains.

Building on the insights from this study, future research should focus on acquiring larger, more balanced datasets, particularly for rare thyroid disorders. Incorporating imaging modalities such as thyroid ultrasound, integrating longitudinal hormone measurements, and leveraging clinical narratives from electronic health records could substantially enhance predictive performance. Exploration of advanced imbalance-handling strategies such as adaptive synthetic sampling, focal loss boosting, or cost-sensitive learning may further improve the classification of underrepresented categories. External validation using multi-center clinical datasets is also essential to assess generalizability. Ultimately, future work should move toward real-time clinical decision support systems that integrate multimodal data to assist physicians in complex thyroid evaluations.

4. Conclusions

This study set out to determine whether an optimized XGBoost-based machine learning framework can accurately perform multiclass classification of thyroid disorders and identify the most influential biochemical and clinical predictors. The results clearly confirm that the proposed approach meets this objective. The model achieved 95.20% accuracy and a weighted F1-score of 0.94 across 27 diagnostic categories, demonstrating that gradient-boosted ensemble methods, when combined with systematic preprocessing and targeted hyperparameter optimization can reliably distinguish complex thyroid conditions using structured clinical and laboratory data. The model also successfully identified key determinants of thyroid status, with TBG, TT4, FTI, and TSH emerging as the most influential predictors, thereby answering the central research questions defined in the introduction.

In summary, the findings establish that the optimized XGBoost framework delivers robust, interpretable, and clinically relevant predictions, with excellent detection of healthy individuals (98% recall) and strong performance on major pathological classes. The comprehensive evaluation, including cross-validation and algorithm comparison, confirms the approach's stability and competitiveness relative to other machine learning models. Importantly, the model's reliance on well-established hormonal signatures reinforces its physiological validity and enhances its suitability for integration into clinical decision-support systems.

Several unexpected findings emerged. Rare diagnostic classes with extremely limited samples produced either unstable or artificially inflated performance metrics, highlighting the ongoing challenges of modeling highly imbalanced clinical datasets. Misclassifications between physiologically adjacent classes, such as borderline hypothyroidism categories, revealed areas where biochemical overlap naturally complicates both machine and human diagnostic interpretation. These discrepancies are consistent with the broader endocrine literature, which shows that subclinical conditions frequently blur diagnostic boundaries.

The novelty of this work lies in its successful application of an optimized gradient-boosting pipeline to a large-scale, 27-class thyroid disease dataset, exceeding the scope of most prior studies, which typically address only binary or low-class problems. The integration of automated preprocessing via ColumnTransformer, a rigorous two-stage hyperparameter tuning strategy, and detailed feature importance analysis provides a reproducible framework that advances machine-learning methodologies for thyroid diagnostics.

Looking ahead, future research should focus on expanding the dataset, especially minority classes to mitigate imbalance-driven errors. Incorporating multimodal data, such as thyroid ultrasound, longitudinal hormone trajectories, and clinical narratives, may enhance model precision in nuanced or ambiguous cases. External validation across multi-center populations is essential to assess generalizability and support clinical adoption. Ultimately, this work provides a foundational step toward AI-assisted thyroid evaluation, with potential implications for earlier detection, improved triage, and more consistent diagnostic decision-making in diverse healthcare settings.

Author Contributions: Conceptualization, A.M.; methodology, A.M.; software, A.M.; validation, A.M.; formal analysis, A.M.; investigation, A.M.; resources, A.M.; data curation, A.M.; writing—original draft preparation, A.M.; writing—review and editing, A.M.; visualization, A.M.; supervision, A.M.; project administration, A.M.; funding acquisition, A.M. All authors have read and agreed to the published version of the manuscript.

Funding: This study does not receive external funding.

Ethical Clearance: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The thyroid disease dataset is publicly available from https://archive.ics.uci.edu/dataset/102/thyroid+disease.

Conflicts of Interest: All the authors declare no conflicts of interest.

References

1. Salman, A. G., Mahdi, I. A.-J., Mukhlef, A. K., abd alsattar Mohammad, R., Zaghir, M. S. H., and muatez Wadaa'a, N. (2024). Physiological Aspects of Thyroid Disorders: Anatomy, Hormones, Diagnosis and Management, *Current Clinical and Medical Education*, Vol. 2, No. 05, 17–32.
2. Fernández-García, V., González-Ramos, S., Martín-Sanz, P., Laparra, J. M., and Boscá, L. (2021). Beyond Classic Concepts in Thyroid Homeostasis: Immune System and Microbiota, *Molecular and Cellular Endocrinology*, Vol. 533, 111333.
3. Wang, H., Shang, F., Jiang, X., Li, Z., Li, D., Zhou, C., Pang, B., Kang, L., Liu, B., and Zhao, Z. (2025). Whole Exome Sequencing and Bioinformatics Reveal PMAIP1 and PDGFRL as Immune-Related Gene Markers in Follicular Thyroid Carcinoma, *Frontiers in Genetics*, Vol. 16, 1509245.
4. D'Aurizio, F., Kratzsch, J., Gruson, D., Petranović Ovčariček, P., and Giovanella, L. (2023). Free Thyroxine Measurement in Clinical Practice: How to Optimize Indications, Analytical Procedures, and Interpretation Criteria While Waiting for Global Standardization, *Critical Reviews in Clinical Laboratory Sciences*, Vol. 60, No. 2, 101–140.
5. Croker, E. E., McGrath, S. A., and Rowe, C. W. (2021). Thyroid Disease: Using Diagnostic Tools Effectively, *Australian Journal of General Practice*, Vol. 50, No. 1/2, 16–21.
6. Macvanin, M. T., Gluvic, Z. M., Zaric, B. L., Essack, M., Gao, X., and Isenovic, E. R. (2023). New Biomarkers: Prospect for Diagnosis and Monitoring of Thyroid Disease, *Frontiers in Endocrinology*, Vol. 14, 1218320.
7. Jaiswal, V., and Gurudiwan, P. (2023). Identifying Thyroid Dysfunction Using Standard Laboratory Testings–A Systematic Review, *Integrative Biomedical Research*, Vol. 7, No. 2, 1–8.
8. Toro-Tobon, D., Loor-Torres, R., Duran, M., Fan, J. W., Singh Ospina, N., Wu, Y., and Brito, J. P. (2023). Artificial Intelligence in Thyroidology: A Narrative Review of the Current Applications, Associated Challenges, and Future Directions, *Thyroid*, Vol. 33, No. 8, 903–917.
9. Sharma, V., Cheetham, T., and Wood, C. (2023). Understanding and Interpreting Thyroid Function Tests, *Paediatrics and Child Health*, Vol. 33, No. 7, 183–188.
10. Chutiyami, M., Cutler, N., Sangon, S., Thaweekoon, T., Nintachan, P., Napa, W., Kraithaworn, P., and River, J. (2025). Community-Engaged Mental Health and Wellbeing Initiatives in Under-Resourced Settings: A Scoping Review of Primary Studies, *Journal of Primary Care & Community Health*, Vol. 16, 21501319251332724.
11. Taha, K. (2025). Machine Learning in Biomedical and Health Big Data: A Comprehensive Survey with Empirical and Experimental Insights, *Journal of Big Data*, Vol. 12, No. 1, 61.
12. Asif, S., Wenhui, Y., Ur-Rehman, S.-, Ul-ain, Q.-, Amjad, K., Yueyang, Y., Jinhai, S., and Awais, M. (2025). Advancements and Prospects of Machine Learning in Medical Diagnostics: Unveiling the Future of Diagnostic Precision, *Archives of Computational Methods in Engineering*, Vol. 32, No. 2, 853–883. doi:10.1007/s11831-024-10148-w.
13. Chen, T., and Guestrin, C. (2016). Xgboost: A Scalable Tree Boosting System, *Proceedings of the 22nd Acm Sigkdd International Conference on Knowledge Discovery and Data Mining*, 785–794.
14. Shaik, N. B., Jongkittinarukorn, K., and Bingi, K. (2024). XGBoost Based Enhanced Predictive Model for Handling Missing Input

Parameters: A Case Study on Gas Turbine, *Case Studies in Chemical and Environmental Engineering*, Vol. 10, 100775. doi:10.1016/j.cscee.2024.100775.

15. Quinlan, R. (1986). Thyroid Disease, *UCI Machine Learning Repository*.

16. Jongejan, R. M. S., Meima, M. E., Visser, W. E., Korevaar, T. I. M., van den Berg, S. A. A., Peeters, R. P., and de Rijke, Y. B. (2022). Binding Characteristics of Thyroid Hormone Distributor Proteins to Thyroid Hormone Metabolites, *Thyroid*, Vol. 32, No. 8, 990–999. doi:10.1089/thy.2021.0588.

17. Bagga, A. D., Johnson, B. P., and Zhang, Q. (2023). A Minimal Human Physiologically Based Kinetic Model of Thyroid Hormones and Chemical Disruption of Plasma Thyroid Hormone Binding Proteins, *Frontiers in Endocrinology*, Vol. 14. doi:10.3389/fendo.2023.1168663.

18. Moustakli, E., and Tsonis, O. (2023). Exploring Hormone Therapy Effects on Reproduction and Health in Transgender Individuals, *Medicina*, Vol. 59, No. 12, 2094. doi:10.3390/medicina59122094.

19. Seyedtabib, M., Najafi-Vosough, R., and Kamyari, N. (2024). The Predictive Power of Data: Machine Learning Analysis for Covid-19 Mortality Based on Personal, Clinical, Preclinical, and Laboratory Variables in a Case–Control Study, *BMC Infectious Diseases*, Vol. 24, No. 1, 411. doi:10.1186/s12879-024-09298-w.

20. Li, R., Hao, X., Diao, Y., Yang, L., and Liu, J. (2025). Explainable Machine Learning Models for Colorectal Cancer Prediction Using Clinical Laboratory Data, *Cancer Control*, Vol. 32. doi:10.1177/10732748251336417.

21. Spencer, C. A. (2023). Laboratory Thyroid Tests: A Historical Perspective, *Thyroid*, Vol. 33, No. 4, 407–419. doi:10.1089/thy.2022.0397.

22. Sutradhar, A., Akter, S., Shamrat, F. M. J. M., Ghosh, P., Zhou, X., Idris, M. Y. I. Bin, Ahmed, K., and Moni, M. A. (2024). Advancing Thyroid Care: An Accurate Trustworthy Diagnostics System with Interpretable AI and Hybrid Machine Learning Techniques, *Heliyon*, Vol. 10, No. 17, e36556. doi:10.1016/j.heliyon.2024.e36556.

23. Girwar, S. M., Jabroer, R., Fiocco, M., Sutch, S. P., Numans, M. E., and Bruijnzeels, M. A. (2021). A Systematic Review of Risk Stratification Tools Internationally Used in Primary Care Settings, *Health Science Reports*, Vol. 4, No. 3. doi:10.1002/hsr2.329.

24. Park, D. J., Park, M. W., Lee, H., Kim, Y.-J., Kim, Y., and Park, Y. H. (2021). Development of Machine Learning Model for Diagnostic Disease Prediction Based on Laboratory Tests, *Scientific Reports*, Vol. 11, No. 1, 7567. doi:10.1038/s41598-021-87171-5.