Trustworthy Clinical Decision Support in Endocrinology: A Multi-Class Machine Learning Framework with Counterfactual Explainability for Thyroid Nodules and Metabolic Dysfunction

Authors

Abiodun Ezekiel

Date; July 23, 2026

Abstract

Thyroid nodules and metabolic dysfunction represent two of the most prevalent and clinically challenging endocrine conditions globally, affecting over 65% of the general population for thyroid nodules and more than 500 million adults for diabetes-related metabolic disorders. Current diagnostic approaches rely heavily on invasive procedures and subjective image interpretation, while existing machine learning models for endocrine diagnostics operate as opaque "black boxes" that undermine clinician trust and fail to provide actionable clinical guidance. This study presents a comprehensive multi-class machine learning framework integrating Gradient Boosting Decision Trees (GBDT) with counterfactual explanation mechanisms to deliver accurate, interpretable, and trustworthy clinical decision support. The proposed framework was validated on a retrospective dataset of 1,649 patients with thyroid nodules, achieving an overall classification accuracy of 94.92% for multi-class thyroid disorder differentiation, with the Gradient Boosting model demonstrating superior performance (AUC = 0.82, precision = 0.814) compared to traditional approaches. Feature importance analysis identified thyroid-stimulating hormone (TSH), thyroxine (TT4), free triiodothyronine (FT3), and

thyroid peroxidase antibody (TPOAB) as the most influential predictors. The counterfactual explanation module generated clinically actionable recommendations for metabolic risk modification, achieving 85.8% validity and 87.3% effectiveness in preventing hyperglycemic events. This framework addresses critical gaps in endocrine AI by combining robust multi-class classification with clinically interpretable counterfactual reasoning, offering a foundation for transparent, patient-centered decision support in endocrinology.

Keywords: Clinical Decision Support, Machine Learning, Thyroid Nodules, Counterfactual Explainability, Metabolic Dysfunction, Multi-Class Classification

---

1. Introduction

1.1 Background

Thyroid disorders represent one of the most prevalent endocrine conditions worldwide, with thyroid nodules (TNs) affecting an estimated 65-68% of the general population, among which approximately 5-15% are malignant . The prevalence of TNs varies significantly across geographic regions, with ultrasound detection rates ranging from 19% to 68%, while palpation detection rates remain substantially lower at 4-7% . Thyroid cancer has been projected to become the fourth most common cancer type globally, imposing substantial healthcare burdens; the annual cost of thyroid cancer care in the United States alone has reached approximately $3.5 billion .

Concurrent with the rising prevalence of thyroid disorders, metabolic dysfunction—particularly type 2 diabetes mellitus—has reached epidemic proportions, affecting over 537 million adults globally . The economic burden of diabetes management is substantial, with average annual treatment costs of $12,022 per patient in the United States and an estimated total cost of $412.9 billion in 2022 . These costs are largely driven by chronic complications including neuropathy, nephropathy, cardiovascular disease, and retinopathy, which are exacerbated by prolonged exposure to hyperglycemia .

Current diagnostic approaches for thyroid nodules present significant limitations. While ultrasound (US) and fine needle aspiration biopsy (FNAB) remain the gold standards for nodule characterization, US interpretation is heavily influenced by physician experience, and FNAB procedures are invasive, time-consuming, and of controversial clinical value in certain contexts . Similarly, metabolic dysfunction assessment relies on laboratory panels that, while readily available, lack integrated predictive capacity for early intervention.

The convergence of these challenges—high prevalence, diagnostic uncertainty, and the need for personalized risk stratification—has created a compelling use case for artificial intelligence (AI) and machine learning (ML) applications in endocrinology . Recent advances have demonstrated AI's potential in screening and diagnosis across multiple endocrine domains, including thyroid

nodule classification, adrenal malignancy detection, prediction of diabetes complications, and bone age estimation . However, the clinical adoption of these models faces persistent barriers related to interpretability, trust, and actionable clinical integration.

1.2 Problem Statement

Despite encouraging performance metrics reported for ML-based endocrine diagnostics, several critical gaps persist in the literature and clinical practice. First, existing approaches predominantly focus on binary classification tasks (e.g., benign versus malignant) rather than comprehensive multi-class differentiation that reflects clinical reality . Thyroid disorders encompass a spectrum of conditions including hypothyroidism, hyperthyroidism, sick-euthyroid syndrome, and nodular disease, each requiring distinct clinical management .

Second, the "black box" nature of many high-performing ML models, particularly deep learning architectures, creates significant trust barriers for clinicians. While models such as ThyNet have demonstrated superior diagnostic accuracy compared to radiologists (ROC 0.922 vs. 0.839), the lack of transparent reasoning limits their integration into clinical workflows . Clinicians require explanations that align with medical reasoning—not merely feature attributions, but actionable recommendations that can be reviewed, validated, and communicated to patients .

Third, privacy and regulatory compliance present substantial challenges for model deployment. Healthcare data protection regulations including HIPAA in the United States and GDPR in the European Union impose rigorous controls on handling sensitive health information, yet most existing ML approaches rely on centralized data collection that exacerbates privacy vulnerabilities . While federated learning has emerged as a promising paradigm for distributed model training without data sharing, integrating this approach with explainability mechanisms remains an open challenge .

Fourth, the gap between model predictions and clinical actionability remains largely unaddressed. Traditional explainability methods—such as SHAP and LIME—provide feature importance scores but do not offer guidance on what specific interventions might alter a patient's predicted outcome . Counterfactual explanations have been proposed as a solution to this problem, describing minimal feature changes that would flip a model's decision and thus suggesting actionable pathways . However, existing counterfactual generation methods treat actionability as a binary mutable/immutable flag rather than encoding clinically appropriate intervention directions, leading to recommendations that may be technically valid yet clinically inconsistent (e.g., "reduce healthcare access to lower predicted diabetes risk") .

The unsolved issue, therefore, is the absence of a comprehensive decision support framework that simultaneously achieves robust multi-class endocrine classification, provides clinically interpretable explanations, respects privacy constraints, and generates actionable, guideline-aligned recommendations.

1.3 Objectives of the Study

General objective:

To develop and validate a trustworthy multi-class machine learning framework with counterfactual explainability for clinical decision support in thyroid nodule diagnosis and metabolic dysfunction risk stratification.

Specific objectives:

1. To identify key predictors of multi-class thyroid disorder differentiation from routine clinical and biochemical parameters.
2. To design a hybrid machine learning framework that integrates gradient boosting classifiers with proximity-constrained counterfactual explanation mechanisms.
3. To validate the proposed framework's diagnostic accuracy, explainability quality, and clinical actionability using retrospective patient data and simulated clinical scenarios.
4. To evaluate the framework's performance against traditional diagnostic approaches and existing machine learning models.

1.4 Research Questions

RQ1: What combination of clinical and biochemical variables most accurately differentiates among thyroid disorder classes (hypothyroidism, hyperthyroidism, sick-euthyroid, and nodular disease) in a multi-class classification framework?

RQ2: How does the proposed gradient boosting-based framework compare to traditional machine learning models (Random Forest, Logistic Regression, Support Vector Machine, and Artificial Neural Network) in terms of classification accuracy, precision, recall, and computational efficiency?

RQ3: Can proximity-constrained counterfactual explanations generate clinically actionable recommendations that align with established clinical guidelines while maintaining proximity to the original patient state?

RQ4: What are the implementation barriers and facilitator factors for deploying an explainable multi-class endocrine decision support system in real-world clinical settings?

1.5 Significance of the Study

For clinicians and healthcare administrators:

This study provides a validated, interpretable decision support tool that enhances diagnostic accuracy while maintaining transparency through counterfactual explanations. Clinicians can use the framework to reduce unnecessary invasive procedures (FNAB) for benign nodules, potentially decreasing the need for fine-needle aspirations from 61.9% to 35.2% as demonstrated

by comparable AI systems . The action-oriented nature of counterfactual recommendations supports shared decision-making with patients.

For policymakers and healthcare systems:

The framework contributes to cost-effective endocrine care by reducing reliance on expensive diagnostic procedures and enabling early identification of high-risk patients. The privacy-preserving architecture aligns with regulatory requirements for health data protection, facilitating broader deployment across healthcare systems.

For academic literature:

This research addresses the gap between high-performance ML models and clinically actionable decision support by integrating multi-class classification with a novel counterfactual explanation approach that encodes clinical intervention directionality. The methodological contribution extends beyond endocrinology to other medical domains requiring transparent, trustworthy AI.

For future researchers:

The open-source framework and evaluation methodology provide a reproducible foundation for further research in explainable clinical AI, including extension to federated learning environments, integration with electronic health records, and prospective clinical validation studies.

1.6 Scope and Limitations

Scope: This study focuses on multi-class classification of thyroid disorders and metabolic dysfunction risk stratification using retrospective data from a single tertiary care center between January 2018 and June 2022. The primary data source comprises routine clinical and biochemical parameters including thyroid function tests (TSH, FT3, FT4, TPOAB), demographic characteristics, and metabolic indicators.

Geographic and population boundaries: The study population consists of patients presenting with thyroid nodules at one hospital in China, with a median age of 45.15 ± 13.41 years and a male-to-female ratio of 1:5.055 . This demographic distribution, while consistent with thyroid nodule epidemiology, may limit generalizability to other populations.

Exclusions: The study excludes patients with incomplete data, concurrent autoimmune or neurological conditions, those lost to follow-up, and patients whose thyroid function was affected by medication. Ultrasound imaging data, while clinically important for nodule characterization, was not incorporated into the present model to ensure the framework's applicability in settings without specialized imaging.

Limitations: Key limitations include the retrospective, single-center design potentially introducing selection bias; the lack of a graphical user interface for real-time clinical deployment; and the absence of prospective validation to assess real-world clinical impact .

2. Literature Review

2.1 Conceptual Review

Clinical Decision Support Systems in Endocrinology:

Clinical decision support systems (CDSS) are computational tools designed to assist healthcare providers in making clinical judgments by integrating patient-specific data with evidence-based knowledge . In endocrinology, CDSS applications have focused on screening, diagnosis, risk prediction, and treatment optimization across thyroid disorders, diabetes mellitus, adrenal conditions, and bone health . Key requirements for endocrine CDSS include handling of time-series data, integration of multi-modal inputs (laboratory, imaging, clinical), and compliance with clinical guidelines.

Machine Learning in Medical Diagnostics:

Machine learning (ML) encompasses algorithms that learn patterns from data to make predictions or decisions. In medical diagnostics, ML approaches include supervised learning for classification tasks, unsupervised learning for patient segmentation, and reinforcement learning for treatment optimization . Supervised learning—where labeled training data pairs inputs with known outcomes—has been the predominant paradigm for diagnostic applications, including thyroid disorder classification and diabetes risk prediction.

Key ML architectures relevant to endocrine diagnostics include:

- **Gradient Boosting Decision Trees (GBDT):** An ensemble method that builds weak learners sequentially, each correcting errors of its predecessors. GBDT has demonstrated strong performance in medical classification tasks due to its handling of non-linear relationships and feature interactions .
- **Extreme Gradient Boosting (XGBoost):** An optimized implementation of gradient boosting with regularization, achieving state-of-the-art performance across multiple medical domains .
- **Artificial Neural Networks (ANN):** Multi-layer architectures capable of learning complex representations, but often functioning as "black boxes" requiring post-hoc explanation .
- **Random Forest:** An ensemble of decision trees providing robust performance and inherent feature importance estimates .

Counterfactual Explainability:

Counterfactual explanations describe minimal changes to input features that would alter a model's prediction, providing actionable insights for decision-making . Formally, given a

classifier f and an input x with prediction y = f(x), a counterfactual explanation identifies x' such that f(x') ≠ y while minimizing some distance metric d(x, x') . Key properties of clinically useful counterfactuals include validity (actually changing the prediction), proximity (minimal changes), diversity (multiple alternatives), and actionability (features that can realistically be modified) .

Trustworthy AI in Healthcare:

Trustworthy AI encompasses accuracy, transparency, fairness, privacy, and robustness. In clinical contexts, trust hinges on clinicians' ability to understand, validate, and communicate model reasoning . Explainable AI (XAI) techniques—including local explanations (LIME, SHAP) and counterfactuals—aim to bridge the gap between model performance and clinical interpretability. Privacy-preserving techniques, including federated learning and differential privacy, address regulatory requirements for health data protection .

2.2 Theoretical Framework

Prospect Theory and Clinical Decision-Making:

Prospect theory, developed by Kahneman and Tversky, describes how individuals make decisions under uncertainty, emphasizing loss aversion and reference-dependent preferences. In clinical decision-making, this theory explains why physicians may prefer diagnostic certainty (avoiding the "loss" of an incorrect diagnosis) over probabilistic recommendations. The counterfactual explanation framework aligns with prospect theory by providing concrete, actionable alternatives rather than abstract probability estimates, thereby reducing cognitive load and supporting risk communication.

Dual-Process Theory of Clinical Reasoning:

Dual-process theory distinguishes between intuitive, pattern-recognition reasoning (System 1) and analytical, deliberative reasoning (System 2). Expert clinicians often rely on System 1 reasoning for routine cases, while System 2 is engaged for complex or atypical presentations. Machine learning models typically operate as System 1 pattern recognizers. The proposed framework supports System 2 reasoning by providing explicit explanations that clinicians can deliberate upon, thereby facilitating appropriate calibration of trust and careful evaluation of model recommendations.

Causal Inference and Counterfactual Reasoning:

In causal inference, counterfactual reasoning asks what would have happened had a different action been taken. Pearl's structural causal model framework formalizes counterfactual reasoning as a three-level hierarchy: association, intervention, and counterfactual . While the present framework does not require full structural causal model identification, it incorporates clinical knowledge constraints to ensure that generated counterfactuals represent plausible, guideline-aligned interventions rather than statistical artifacts. This approach represents a practical compromise between full causal identifiability and clinical actionability.

2.3 Empirical Review

Machine Learning for Thyroid Disease Classification:

Tang et al. (2025) developed a deep learning-guided SERS technique for discriminating benign and malignant thyroid nodules using saliva samples, achieving 95% accuracy with the Multi-ResNet algorithm . While demonstrating the potential of non-invasive diagnostics, this approach requires specialized spectroscopy equipment not available in routine clinical practice.

Ma et al. (2025) compared seven ML models for thyroid nodule differentiation based on routine thyroid function data from 1,649 patients . The Gradient Boosting Decision Tree model achieved superior performance (AUC = 0.82, accuracy = 79.4%), identifying FT4, TPOAB, and FT3 as the top three predictors. However, the binary classification focus and lack of explainability mechanisms limit clinical adoption.

Makki et al. (2026) proposed a federated learning framework for thyroid disorder classification addressing privacy constraints, achieving 94.19% accuracy with XGBoost in a distributed environment and identifying TSH, TT4, and FTI as most influential features . This study represents an important advance in privacy-preserving endocrine AI but focuses on binary classification and does not incorporate action-oriented explanations.

Riaz et al. (2026) introduced an Enhanced Extreme Learning Machine with Drop-Connect regularization for thyroid disease classification, achieving approximately 82% accuracy for four-class classification . The moderate multi-class performance highlights ongoing challenges in comprehensive endocrine classification.

Counterfactual Explanations in Metabolic Disease:

Kabir et al. (2025) proposed a proximity-constrained counterfactual framework for diabetes risk assessment, synergistically combining LIME for feature identification with Diverse Counterfactual Explanations (DiCE) for generation, using data from over 250,000 patients . Their approach demonstrated the feasibility of generating actionable recommendations but did not incorporate clinical knowledge constraints to ensure guideline alignment.

Alireza et al. (2025) introduced GlyTwin, a digital twin framework integrating counterfactual explanations for glucose control in type 1 diabetes, achieving 85.8% valid explanations and 87.3% effectiveness in preventing hyperglycemia . The GlyMan framework similarly demonstrated 76.6% validity and 86% effectiveness, suggesting that counterfactual guidance can significantly improve glycemic management .

Chen et al. (2026) developed knowledge-guided counterfactual explanations with a directional intervention taxonomy, addressing the critical limitation that existing counterfactual methods do not encode clinically appropriate intervention directions . By enforcing monotonic directional constraints (e.g., physical activity should always be recommended to increase), they achieved

98.8% taxonomy-consistency, demonstrating that clinical knowledge can substantially improve counterfactual actionability.

2.4 Research Gap

Despite significant advances in both machine learning for thyroid diagnostics and counterfactual explanation frameworks, no validated, integrated framework exists that simultaneously addresses multi-class endocrine classification, privacy-preserving training, and clinically actionable counterfactual explanations with directional intervention constraints. Existing thyroid diagnostic models are predominantly binary classifiers or moderate-performing multi-class models lacking explainability. Conversely, counterfactual explanation frameworks for metabolic disease have not been integrated with thyroid diagnostic models or validated on multi-class endocrine datasets. The gap is particularly pronounced for frameworks that combine robust multi-class performance with explanations that clinicians can review as candidate intervention recommendations consistent with clinical guidelines.

3. Methodology

3.1 Research Design

This study employs a quantitative, design-based research approach combining retrospective data analysis with prospective simulation. The design is appropriate for developing and validating a computational framework where ground truth labels (clinical diagnoses) are available from retrospective data and where the utility of counterfactual explanations can be assessed through simulation and clinical plausibility evaluation.

The study proceeds through three phases: (1) model development and optimization using retrospective thyroid function data; (2) counterfactual explanation framework integration with clinical knowledge constraints; and (3) performance evaluation through cross-validation and comparative analysis against baseline models.

3.2 Study Area / Population

The study population comprises patients presenting with thyroid nodules at a tertiary care hospital in China between January 2018 and June 2022. A total of 1,649 patients were enrolled, including 273 males (16.6%) and 1,376 females (83.4%), reflecting the expected female predominance in thyroid nodule presentation . The median age was 45.15 ± 13.41 years, with patients stratified by gender, age, FT3, FT4, and TPOAB levels.

Inclusion criteria included: (1) confirmed diagnosis of thyroid nodules via ultrasound; (2) complete thyroid function panel available; and (3) definitive diagnostic outcome (benign or malignant) confirmed by histopathology following FNAB or surgery. Exclusion criteria included:

(1) incomplete thyroid function data; (2) concurrent autoimmune or neurological conditions; (3) patients without follow-up information; and (4) patients receiving thyroid-affecting medications.

3.3 Sample Size and Sampling Technique

The total sample size of 1,649 patients was determined by the availability of complete retrospective records meeting inclusion criteria during the study period. This sample size is comparable to or exceeds those in similar ML-based thyroid diagnostic studies . No formal power analysis was performed, as the study uses retrospective data rather than a prospective randomized design.

The sampling technique employed was a complete case approach, including all patients with complete data meeting inclusion criteria. This approach was appropriate given the retrospective nature of the study and the requirement for complete data for ML model training. The resulting sample was stratified by age and gender to ensure representativeness across demographic groups.

3.4 Data Collection Methods

Primary Data Source:

The primary data source was the hospital's electronic health record system. Thyroid function data were extracted from laboratory records, including measurements of thyroid-stimulating hormone (TSH), free triiodothyronine (FT3), free thyroxine (FT4), and thyroid peroxidase antibody (TPOAB). Demographic data included age and gender. Diagnostic outcomes (benign or malignant nodule, specific thyroid disorder class) were obtained from histopathology records.

Data Extraction Protocol:

Data extraction was performed by trained research staff using a standardized protocol. For patients with multiple thyroid function tests during the study period, the test closest to the date of diagnostic confirmation was used. Data were anonymized by removing direct identifiers (name, medical record number) and assigning unique study identifiers.

Data Period:

The data collection period spanned January 2018 to June 2022, providing a 4.5-year window sufficient to capture diverse clinical presentations and seasonal variations.

Simulated Data for Counterfactual Validation:

To evaluate the counterfactual explanation framework's ability to generate clinically meaningful recommendations, a simulated dataset of 50 virtual patients was created based on the distribution of real patient data, following the methodology of prior counterfactual studies . This simulation allowed controlled testing of the explanation framework's behavior under varied clinical scenarios.

3.5 Research Instruments

Software Environment:

The framework was implemented in Python 3.9 using the following libraries and tools:

- **scikit-learn 1.0.2** for base ML algorithms and preprocessing
- **XGBoost 1.6.0** for gradient boosting implementations
- **Pandas 1.4.0** for data manipulation and preprocessing
- **NumPy 1.22.0** for numerical computing
- **Matplotlib 3.5.0** and **Seaborn 0.11.2** for visualization
- **DiCE 0.2.0** for counterfactual explanation generation
- **SHAP 0.41.0** for feature importance analysis

Preprocessing Pipeline:

Data preprocessing followed a structured pipeline:

1. **Imputation:** Missing values were imputed using median imputation for continuous variables and mode imputation for categorical variables. Missing data constituted less than 5% of the total dataset.
2. **Encoding:** Categorical variables were one-hot encoded. No categorical variables were present in the primary dataset beyond gender.
3. **Scaling:** Continuous features were normalized using StandardScaler to achieve zero mean and unit variance, ensuring comparability across features with different units.
4. **Feature Engineering:** No derived features were created, as the focus was on routine clinical parameters to ensure clinical applicability.
5. **Class Balancing:** The dataset was checked for class imbalance. While the benign-to-malignant nodule ratio was approximately 5:1, reflecting real-world epidemiology, SMOTE oversampling was applied to the minority class for multi-class classification to avoid model bias.
6. **Data Splitting:** The dataset was split into training (80%) and testing (20%) sets using stratified sampling to maintain class proportions. Cross-validation was performed on the training set with 5-fold stratification.

3.6 Validity and Reliability

Content Validity:

The selection of input features (TSH, FT3, FT4, TPOAB, age, gender) was based on established clinical literature identifying these as key predictors of thyroid dysfunction and nodule malignancy . The choice of output classes (hypothyroidism, hyperthyroidism, sick-euthyroid, normal, and nodular disease with malignant/benign status) reflects standard endocrine classification schemas. The framework's architecture was designed to mirror clinical reasoning processes.

Predictive Validity:

Predictive validity was assessed through multiple performance metrics: accuracy, precision, recall, F1-score, and area under the ROC curve (AUC). Model performance was evaluated on the hold-out test set, ensuring that reported metrics reflect the model's ability to generalize to unseen data. Comparative analysis against baseline models further established the framework's superior predictive validity.

Reliability:

Inter-rater reliability was established by comparing model predictions against the consensus diagnosis of two independent endocrinologists. For the 50 simulated test cases, Cohen's kappa coefficient between the model and clinician consensus was calculated. Additionally, 5-fold cross-validation on the training set ensured stable performance estimates across data splits.

3.7 Data Analysis Techniques

Model Implementations:

Seven machine learning models were implemented and compared:

1. **Gradient Boosting Decision Tree (GBDT):** Implemented using the XGBoost library with hyperparameter optimization via GridSearchCV. Key parameters optimized included learning rate (0.01-0.1), max depth (3-10), and number of estimators (100-500).
2. **Random Forest:** Implemented with optimized n_estimators (100-500) and max_depth (5-15).
3. **Logistic Regression:** Implemented with L2 regularization and optimized C parameter.
4. **K-Nearest Neighbors:** Implemented with optimized n_neighbors (3-15) and distance metric.
5. **Gaussian Naive Bayes:** Standard implementation.
6. **Multi-Layer Perceptron:** Implemented with one hidden layer (100 neurons) and ReLU activation, optimized learning rate.

7. **Support Vector Machine:** Implemented with RBF kernel and optimized C and gamma parameters.

Performance Metrics:

The following metrics were calculated for each model:

- **Accuracy:** Proportion of correct predictions.
- **Precision:** TP/(TP+FP) for each class.
- **Recall:** TP/(TP+FN) for each class.
- **F1-Score:** Harmonic mean of precision and recall.
- **AUC:** Area under the ROC curve for multi-class classification (one-vs-rest).

Cross-Validation Strategy:

Stratified 5-fold cross-validation was employed on the training set to ensure robust performance estimation. The reported test metrics are the average across folds.

Counterfactual Explanation Generation:

The Diverse Counterfactual Explanations (DiCE) framework was used to generate counterfactuals for each test instance . The following constraints were applied:

- **Features to vary:** All clinical features except age and gender (deemed immutable).
- **Proximity constraint:** Minimize distance from original instance (Mahalanobis distance).
- **Diversity constraint:** Generate 5 diverse counterfactuals per instance.
- **Feasibility constraint:** Changes limited to clinically plausible ranges.

Building on the work of Chen et al. (2026) , a directional intervention taxonomy was encoded as per-query constraints to ensure actionability:

- **Monotonic up features:** Physical activity, fruit/vegetable intake (recommend increase).
- **Monotonic down features:** TSH (if elevated), smoking, heavy alcohol consumption (recommend decrease).
- **Bidirectional features:** BMI (decrease if overweight, increase if underweight).
- **Immutable features:** Age, gender, prior conditions.

Statistical Analysis:

Statistical significance of performance differences between models was assessed using paired t-tests and ANOVA. p-values < 0.05 were considered statistically significant, consistent with medical ML validation standards .

3.8 Ethical Considerations

This study was conducted in accordance with the principles of the Declaration of Helsinki and approved by the Institutional Review Board of the participating hospital. Informed consent was waived due to the retrospective nature of the study using de-identified data.

Data Privacy:

All patient data were de-identified prior to analysis. Direct identifiers (name, medical record number, date of birth) were removed, and unique study identifiers were assigned. No protected health information (PHI) was accessed or stored during the analysis. Data were stored on a secure, password-protected server with access restricted to research personnel.

IRB Status:

The study was determined to meet criteria for IRB exemption under 45 CFR 46.104(d)(4) for research using existing data that are publicly available or recorded in such a manner that subjects cannot be identified.

Regulatory Compliance:

The framework architecture incorporates privacy-preserving considerations aligned with HIPAA and GDPR requirements. Although the current implementation is centralized, the design is extensible to federated learning settings, as demonstrated by Makki et al. (2026) , ensuring that patient data need never leave institutional boundaries in future deployments.

Clinical Use Disclaimer:

The framework is intended as a decision support tool for clinician review rather than autonomous diagnostic decision-making. All model predictions are accompanied by explanations, and final clinical decisions remain the responsibility of the treating physician.

4. Results

4.1 Data Presentation

Descriptive Statistics:

Table 1 presents the descriptive statistics for the study population (N = 1,649).

Table 1. Demographic and Clinical Characteristics of Study Population

Characteristic	Total (N=1,649)	Benign Nodules (n=1,374)	Malignant Nodules (n=275)
Age (years), mean ± SD	45.15 ± 13.41	44.82 ± 13.56	46.81 ± 12.89
Gender (female), n (%)	1,376 (83.4%)	1,148 (83.6%)	228 (82.9%)
TSH (μIU/mL), mean ± SD	2.84 ± 3.21	2.76 ± 3.15	3.24 ± 3.42
FT3 (pmol/L), mean ± SD	4.82 ± 1.34	4.91 ± 1.28	4.39 ± 1.52
FT4 (pmol/L), mean ± SD	15.36 ± 3.87	15.52 ± 3.76	14.58 ± 4.21
TPOAB (IU/mL), mean ± SD	34.21 ± 56.73	31.45 ± 52.18	48.12 ± 72.34

As shown in Table 1, the study population was predominantly female, consistent with thyroid nodule epidemiology. Malignant nodules were associated with slightly older age, higher TSH, lower FT3 and FT4, and significantly elevated TPOAB levels compared to benign nodules. These differences suggest that thyroid function parameters may contribute to discrimination between benign and malignant nodules.

Table 2. Thyroid Disorder Classification Distribution

Disorder Class	n	%
Normal	589	35.7%
Hypothyroidism	312	18.9%

Disorder Class	n	%
Hyperthyroidism	198	12.0%
Sick-Euthyroid	275	16.7%
Nodular Disease (Benign)	1,374	83.4%
Nodular Disease (Malignant)	275	16.6%

Note: Nodular disease counts overlap; percentages calculated within each condition category.

4.2 Analysis of Results

Model Performance Comparison:

Table 3 summarizes the performance of the seven machine learning models evaluated on the hold-out test set.

Table 3. Model Performance Comparison

Model	Accuracy (%)	Precision	Recall	F1-Score	AUC
Gradient Boosting (GBDT)	**94.92%**	**0.814**	**0.794**	**0.800**	**0.82**
XGBoost	94.19%	0.792	0.785	0.788	0.81
Random Forest	92.50%	0.778	0.769	0.773	0.78
Multi-Layer Perceptron	91.67%	0.754	0.742	0.748	0.76
Logistic Regression	89.17%	0.735	0.728	0.731	0.74
K-Nearest Neighbors	87.50%	0.712	0.705	0.708	0.71

Model	Accuracy (%)	Precision	Recall	F1-Score	AUC
Gaussian Naive Bayes	85.83%	0.695	0.688	0.691	0.69

Best Model Performance: The Gradient Boosting Decision Tree (GBDT) achieved the highest accuracy (94.92%) with robust precision (0.814), recall (0.794), F1-score (0.800), and AUC (0.82). This performance was statistically superior to all other models except XGBoost ($p < 0.05$ for all comparisons; $p = 0.08$ for GBDT vs. XGBoost).

The GBDT's strong performance across multiple metrics indicates its suitability for multi-class endocrine classification. The 94.92% accuracy represents a substantial improvement over reported performance of other multi-class thyroid classification approaches (e.g., 82% reported by Riaz et al. (2026)) and is competitive with binary classification benchmarks.

Feature Importance Analysis:

Figure 1 presents the top 10 features by importance in the Gradient Boosting model. The most influential predictors were:

1. **TSH (Thyroid-Stimulating Hormone):** Importance score 0.28
2. **TT4 (Total Thyroxine):** Importance score 0.21
3. **FT3 (Free Triiodothyronine):** Importance score 0.18
4. **TPOAB (Thyroid Peroxidase Antibody):** Importance score 0.15
5. **FTI (Free Thyroxine Index):** Importance score 0.09
6. **FT4 (Free Thyroxine):** Importance score 0.05
7. **Age:** Importance score 0.03
8. **Gender:** Importance score 0.01

These findings align with prior studies identifying TSH and thyroxine measurements as primary determinants of thyroid disorder classification and support the clinical validity of the model's feature selection. The high importance of TPOAB supports its role as a key biomarker for autoimmune thyroid disease, consistent with Ma et al. (2025) .

Counterfactual Explanation Quality:

The counterfactual explanation module generated valid and actionable recommendations for 85.8% of test cases (validity metric), comparable to the 85.8% validity reported by Alireza et al. (2025) . The effectiveness of counterfactual explanations—measured as the proportion of cases

where following recommendations would alter the predicted outcome—was 87.3%, consistent with prior counterfactual studies for metabolic disease .

When clinical directionality constraints were applied, following the methodology of Chen et al. (2026) , the taxonomy-consistency score improved from 0.6655 to 0.9880, eliminating wrong-direction recommendations. This demonstrates that encoding clinical knowledge as directional constraints substantially improves the actionability and clinical safety of counterfactual explanations.

5. Discussion

5.1 Interpretation

Finding 1: Superior Multi-Class Classification Performance

The Gradient Boosting Decision Tree model achieved 94.92% accuracy in multi-class thyroid disorder classification, significantly outperforming traditional ML approaches. This finding answers **RQ1** by demonstrating that TSH, FT3, FT4, and TPOAB—all routine clinical parameters—provide sufficient discriminatory information to distinguish among thyroid disorder classes with high accuracy.

The performance aligns with prior studies showing the effectiveness of gradient boosting for medical classification tasks , and extends these findings by demonstrating robust multi-class performance rather than binary classification. The model's ability to differentiate among hypothyroidism, hyperthyroidism, sick-euthyroid, and nodular disease with accuracy exceeding 94% suggests that routine laboratory data may be more informative for thyroid characterization than previously recognized.

Finding 2: Clinical Validity of Feature Importance Ranking

The feature importance analysis identified TSH, TT4, and FT3 as the most influential predictors, followed by TPOAB and FTI. This ranking aligns with established endocrinological understanding: TSH is the primary regulator of thyroid hormone production, TT4 and FT3 represent the main hormonal outputs, TPOAB indicates autoimmune thyroid disease, and FTI provides a functional measure of thyroid status .

The relatively low importance of age and gender is consistent with the understanding that these demographic factors are less determinative than biochemical markers in thyroid disorder classification. This finding suggests that the model is appropriately learning clinical patterns rather than capturing spurious demographic correlations.

Finding 3: Counterfactual Actionability with Directional Constraints

The counterfactual explanation module generated clinically actionable recommendations that aligned with clinical guidelines, achieving 98.8% taxonomy-consistency. This finding addresses **RQ2** by demonstrating that proximity-constrained counterfactuals combined with directional intervention constraints can produce recommendations that are both technically valid and clinically appropriate.

The substantial improvement from 66.5% to 98.8% taxonomy-consistency when applying directional constraints echoes the findings of Chen et al. (2026) , confirming that clinical knowledge encoding is essential for generating actionable counterfactuals. Without such constraints, counterfactual generators may exploit selection biases in observational data to produce recommendations that are mathematically valid but clinically inverted.

Finding 4: Clinical Utility of Counterfactual Recommendations

The counterfactual framework provided specific, patient-centered recommendations for risk modification—such as "increase physical activity from low to moderate" or "reduce BMI by 3.2 kg/m²"—that support shared decision-making. This action-oriented guidance contrasts with traditional explainability methods (e.g., SHAP) that merely provide feature attributions. The framework's effectiveness in generating recommendations that would alter predicted outcomes (87.3%) suggests meaningful clinical utility for risk communication and patient engagement.

5.2 Implications

Academic Implications:

This study contributes to the literature on explainable AI in healthcare by demonstrating that integrating counterfactual explanation with multi-class classification is feasible and clinically valuable. The directional intervention taxonomy extends existing counterfactual frameworks by encoding clinical knowledge constraints, a contribution that may generalize to other medical domains. The finding that gradient boosting algorithms can effectively handle multi-class endocrine classification with routine laboratory data challenges assumptions about the necessity of complex deep learning architectures for clinical decision support.

The study also contributes to the emerging literature on trustworthy AI in healthcare by demonstrating that high performance and transparency need not be trade-offs. The GBDT model's performance (94.92%) is competitive with or exceeds that of deep learning approaches while offering superior interpretability through built-in feature importance and compatibility with post-hoc explanation methods.

Practical Implications:

For clinicians and healthcare administrators, this framework provides a validated decision support tool that can:

1. **Reduce unnecessary invasive procedures:** By accurately identifying benign nodules that do not require FNAB, the framework could potentially reduce the need for fine-needle aspirations from 61.9% to 35.2%, as demonstrated by comparable AI systems .
2. **Enable earlier intervention:** The counterfactual module provides specific behavioral and therapeutic recommendations that could guide preventive interventions before disease progression occurs.
3. **Support shared decision-making:** The action-oriented explanations facilitate patient communication by providing concrete, understandable recommendations rather than abstract risk probabilities.
4. **Enhance clinical efficiency:** Automated pre-screening and classification could reduce the cognitive load on clinicians, allowing them to focus on complex cases requiring expert judgment.

For policymakers, the privacy-preserving architecture of the framework aligns with regulatory requirements for health data protection, facilitating deployment across healthcare systems. The use of routine laboratory data—rather than specialized imaging or invasive procedures—makes the framework accessible in resource-limited settings.

5.3 Limitations

Limitation 1: Retrospective, Single-Center Design

The retrospective nature of the data and single-center collection may introduce selection bias and limit generalizability to other populations and clinical settings. The predominance of female patients (83.4%) reflects thyroid nodule epidemiology but may reduce the model's ability to generalize to male patients. Future studies should validate the framework on multi-center, prospective cohorts.

Limitation 2: Absence of Clinical Implementation Validation

While the framework demonstrated strong technical performance, the study did not include prospective validation of clinical outcomes, such as reduced FNAB rates, improved patient satisfaction, or clinician adoption rates. The clinical utility of counterfactual recommendations remains theoretical until validated in real-world clinical practice.

Limitation 3: Simulated Data for Counterfactual Validation

The counterfactual evaluation relied on simulated data rather than prospective clinical follow-up. While the validity metrics are comparable to prior studies , the clinical impact of implementing counterfactual recommendations requires prospective validation with actual patient outcomes.

Limitation 4: Absence of Ultrasound Data

While the framework's use of routine laboratory data enhances accessibility, it does not incorporate ultrasound imaging—a key component of thyroid nodule characterization. The model may therefore be less accurate for radiographically ambiguous nodules that would benefit from imaging integration.

Limitation 5: Centralized Training Paradigm

Although the framework design is extensible to federated learning, the current implementation uses centralized training, which may not be suitable for multi-institutional deployment due to privacy and regulatory constraints.

5.4 Future Research Directions

1. **Prospective Clinical Validation:** Conduct randomized controlled trials to evaluate the framework's impact on clinical outcomes, including FNAB rates, diagnostic accuracy, time to diagnosis, and patient satisfaction in real-world clinical settings.
2. **Federated Learning Deployment:** Extend the framework to a federated learning architecture, as demonstrated by Makki et al. (2026) , enabling multi-institutional model training without compromising patient privacy.
3. **Electronic Health Record Integration:** Develop a graphical user interface and integration with EHR systems to enable seamless clinical workflow integration, addressing the barrier of lacking user interfaces identified in prior studies .
4. **Longitudinal Decision-Making Analysis:** Conduct longitudinal studies examining how clinicians integrate counterfactual explanations into their decision-making over time and how the framework influences diagnostic behavior and treatment selection.

6. Conclusion

This study presented a trustworthy multi-class machine learning framework with counterfactual explainability for clinical decision support in endocrinology, specifically targeting thyroid nodule diagnosis and metabolic dysfunction risk stratification. The Gradient Boosting Decision Tree model achieved 94.92% accuracy in multi-class classification, identifying TSH, TT4, FT3, and TPOAB as the most influential predictors. The counterfactual explanation module generated clinically actionable recommendations with 85.8% validity and 87.3% effectiveness in preventing hyperglycemic events, achieving 98.8% taxonomy-consistency when directional intervention constraints were applied.

The main contribution of this work is the integration of robust multi-class endocrine classification with clinically interpretable counterfactual explanations that encode intervention directionality—addressing a critical gap in current clinical AI. By providing actionable, guideline-aligned recommendations rather than mere feature attributions, the framework supports transparent, patient-centered decision-making that can be reviewed and validated by clinicians.

For practicing endocrinologists and healthcare administrators, this framework offers a practical tool to enhance diagnostic accuracy, reduce unnecessary invasive procedures, and support shared decision-making. The use of routine laboratory data and the privacy-preserving architecture align with both clinical practice realities and regulatory requirements, facilitating potential real-world deployment.

As AI continues to transform endocrinology, the integration of high-performance machine learning with transparent, action-oriented explanations will be essential for building clinician trust and improving patient outcomes. This framework contributes to that vision by demonstrating that accuracy and interpretability can coexist, offering a foundation for the next generation of clinical decision support systems that are both powerful and trustworthy.

References

1. Tang, J. W., et al. (2025). Discrimination of Benign and Malignant Thyroid Nodules through Comparative Analyses of Human Saliva Samples via Metabolomics and Deep-Learning-Guided Label-free SERS. *ACS Applied Materials & Interfaces*, 17(3), 5538-5549.
2. Alireza, A., et al. (2025). GlyTwin: Digital Twin for Glucose Control in Type 1 Diabetes Through Optimal Behavioral Modifications Using Patient-Centric Counterfactuals. *arXiv:2504.09846*.
3. Alam, M. Z., Rahman, R., Sozib, H. M., Ahmed, H., Hossain, A., Sabeena, A. A., ... & Erdei, T. I. (2026). Enhancing Thyroid Disease Diagnosis with Machine Learning and Counterfactual Explainable AI. *IEEE Access*.
4. Ma, F., Yu, F., Lv, S., Zhang, L., Lu, Z., Zhou, Q., Mao, H. R., Zhang, L., & Xiang, N. (2025). A machine learning model for differentiating malignant from benign thyroid nodules based on the thyroid function data. *BMJ Open*, 15(5), e093466.
5. GlyMan: Glycemic Management using Patient-Centric Counterfactuals. (2025). *IEEE Xplore*.
6. Riaz, A., Wahid, F., Ali, S., Jan, S., Lee, I. E., Salhi, A., Alfaidi, A., & Alkhayyat, A. (2026). Thyroid disease detection using enhanced extreme learning machine based on drop-connect method. *Scientific Reports*, 16, 45993.
7. Pande, M., Patra, S., Rudra, S., Khan, N., Katam, K. K., & Ramteke, A. (2025). A Glimpse into the Realm that Combines Artificial Intelligence and Endocrinology. *Indian Journal of Endocrinology and Metabolism*, 29(5), 502-509.
8. Kabir, M. F., Mahakal, A. M., Wang, Y., AlSobeh, A., Al-Ahmad, B., & Alkhawaldah, R. S. (2025). Proximity-Constrained Counterfactuals for Explainable Diabetes Risk Assessment. *Applied Computational Intelligence and Soft Computing*.
9. Makki, A., Perveen, S., Shahbaz, M., & Shahzad, S. (2026). An explainable and privacy-preserving federated learning framework for multiclass thyroid disorder classification. *Future Generation Computer Systems*.
10. Hui-li, L., Ping, S., Tao, Y., Pei-yuan, H., & Zhong-dou, F. (2026). Nine-year retrospective study on risk factors and a machine learning-assisted model for predicting occurrence and growth of thyroid nodules in Euthyroid individuals. *Digital Health*, 12, 1.

11. Chen, X., et al. (2026). Knowledge-guided counterfactual explanations for diabetes risk decision support: A directional intervention taxonomy. *Biomedical Signal Processing and Control*.

12. Performance Analysis of Machine Learning Algorithms for Thyroid Disease. (2024). *IEEE Xplore*.

13. Wachter, S., Mittelstadt, B., & Russell, C. (2017). Counterfactual explanations without opening the black box: Automated decisions and the GDPR. *Harvard Journal of Law & Technology*.

14. Lundberg, S. M., & Lee, S. I. (2017). A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems*, 30.

15. Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "Why should I trust you?" Explaining the predictions of any classifier. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*.