

أحمد عبدالله

AI-Assisted_Thyroid_CDSS_Paper_Combined.docx

 Midocean University

Document Details

Submission ID

trn:oid::2945:400569127

Submission Date

Jul 21, 2026, 8:10 PM GMT+3

Download Date

Jul 23, 2026, 6:47 PM GMT+3

File Name

AI-Assisted_Thyroid_CDSS_Paper_Combined.docx

File Size

3.4 MB

13 Pages

3,732 Words

23,434 Characters

15% Overall Similarity

The combined total of all matches, including overlapping sources, for each database.

Filtered from the Report

- Bibliography
- Quoted Text

Match Groups

-  **46 Not Cited or Quoted 13%**
Matches with neither in-text citation nor quotation marks
-  **8 Missing Quotations 2%**
Matches that are still very similar to source material
-  **0 Missing Citation 0%**
Matches that have quotation marks, but no in-text citation
-  **0 Cited and Quoted 0%**
Matches with in-text citation present, but no quotation marks

Top Sources

- 8%  Internet sources
- 7%  Publications
- 10%  Submitted works (Student Papers)

Integrity Flags

0 Integrity Flags for Review

No suspicious text manipulations found.

Our system's algorithms look deeply at a document for any inconsistencies that would set it apart from a normal submission. If we notice something strange, we flag it for you to review.

A Flag is not necessarily an indicator of a problem. However, we'd recommend you focus your attention there for further review.

Match Groups

- 46 Not Cited or Quoted** 13%
Matches with neither in-text citation nor quotation marks
- 8 Missing Quotations** 2%
Matches that are still very similar to source material
- 0 Missing Citation** 0%
Matches that have quotation marks, but no in-text citation
- 0 Cited and Quoted** 0%
Matches with in-text citation present, but no quotation marks

Top Sources

- 8% Internet sources
- 7% Publications
- 10% Submitted works (Student Papers)

Top Sources

The sources with the highest number of matches within the submission. Overlapping sources will not be displayed.

- 1** **Publication**
Liangbin Wu, Chuanqi Wang, Yong Yu. "Data-augmented machine learning for pr... <1%
- 2** **Internet**
pmc.ncbi.nlm.nih.gov <1%
- 3** **Publication**
Aysel Topşır, Ferdi Güler, Ecesu Çetin, Mehmet Furkan Burak, Melih Ağraz. "Thyro... <1%
- 4** **Internet**
acnsci.org <1%
- 5** **Submitted works**
KICT on 2026-05-26 <1%
- 6** **Submitted works**
University of Sunderland on 2026-06-25 <1%
- 7** **Internet**
arxiv.org <1%
- 8** **Internet**
pubmed.ncbi.nlm.nih.gov <1%
- 9** **Publication**
Jianhua Xie, Zhong Cao. "Carbon Star Identification in LAMOST DR12 Based on a ... <1%
- 10** **Submitted works**
University of Sunderland on 2026-01-01 <1%

11	Internet	repub.eur.nl	<1%
12	Submitted works	Far Eastern University on 2025-12-06	<1%
13	Submitted works	University of Greenwich on 2024-04-22	<1%
14	Publication	B. G. Nagaraja, S. Kannadhasan. "Information and Communication Systems", CRC...	<1%
15	Submitted works	City University on 2024-08-09	<1%
16	Publication	Shoujie Jia, Qiong Li, Wenji Zhao, Yanhui Wang. "How Do Rural Households Achiev...	<1%
17	Submitted works	Berlin School of Business and Innovation on 2025-08-11	<1%
18	Internet	arccjournals.com	<1%
19	Internet	github.com	<1%
20	Submitted works	UCL on 2025-07-23	<1%
21	Submitted works	KPR Institute of Engineering and Technology on 2026-03-23	<1%
22	Internet	repositorio.comillas.edu	<1%
23	Submitted works	Liverpool John Moores University on 2026-02-23	<1%
24	Submitted works	Trident University International on 2026-01-23	<1%

25	Submitted works	Universität Bayreuth on 2026-07-06	<1%
26	Internet	journal.nsps.org.ng	<1%
27	Publication	"Biomedical Engineering Science and Technology", Springer Science and Business...	<1%
28	Publication	Agostinho Sousa Pinto, António Abreu, Manuel Pérez Cota, Jerónimo Paiva, Laís G...	<1%
29	Submitted works	University of Stellenbosch, South Africa on 2025-03-19	<1%
30	Internet	mdpi-res.com	<1%
31	Internet	www.iieta.org	<1%
32	Internet	www.mdpi.com	<1%
33	Publication	Bhabani Prasanna Pattanaik, Dharendra Kumar Shukla, Sambit Satpathy, Kamala...	<1%
34	Publication	Min Yang, Wanzhong Shen, Junbo Hou. "Electrochemical Energy Storage for the G...	<1%
35	Submitted works	University of Exeter on 2024-12-10	<1%
36	Internet	jaiet.com	<1%
37	Internet	ps-rc.org	<1%
38	Internet	researchbriefings.files.parliament.uk	<1%

39

Submitted works

Higher Education Commission Pakistan on 2026-04-23

<1%

40

Submitted works

UC, San Diego on 2022-02-05

<1%

41

Internet

en.wikipedia.org

<1%

AI-Assisted Clinical Decision Support for Thyroid Risk Assessment Using an XGBoost Classifier

Ahmed Abdullah

253000598@st.midocean.edu.km

Dr. Amjad Qtaish

Abstract

Thyroid disorders are among the most common endocrine conditions worldwide, and delayed or inconsistent diagnosis can contribute to significant health complications. Conventional assessment relies on interpreting several laboratory biomarkers together, a process that can be time-consuming and subject to variation between clinicians.

This study presents an AI-assisted Clinical Decision Support System (CDSS) that supports, rather than replaces, clinical judgment in thyroid risk assessment. Five routinely measured biomarkers, Thyroid Stimulating Hormone (TSH), Triiodothyronine (T3), Total Thyroxine (TT4), Thyroxine Uptake (T4U), and the Free Thyroxine Index (FTI) were drawn from the UCI ANN Thyroid Disease dataset and used to train an Extreme Gradient Boosting (XGBoost) classifier for three-class prediction of Normal, Hypothyroidism, and Hyperthyroidism status. On an independent test set, the model reached an overall accuracy of 97.23%.

The trained classifier was then integrated into a Streamlit-based prototype that returns a predicted class, a confidence score, class probabilities, and supporting laboratory value summaries to the user. These results suggest that combining a gradient-boosted classifier with a decision-support interface is a feasible approach to assisting not automating, thyroid risk assessment, and the study outlines directions for validating the approach on larger, more diverse clinical data.

Keywords: *Artificial Intelligence, Clinical Decision Support System, Machine Learning, XGBoost, Thyroid Risk Assessment.*

1. Introduction

Thyroid disorders are among the most common endocrine conditions worldwide and can affect metabolism, cardiovascular function, and neurological activity when hormone production is disrupted. Hyperthyroidism and hypothyroidism, the two principal patterns of dysfunction, are associated with substantial morbidity when diagnosis is delayed.

Diagnosis of thyroid disorders is based on interpreting several laboratory biomarkers together — principally TSH, T3, TT4, T4U, and FT, rather than on any single measurement. In practice, this combined interpretation can be time-consuming, and results that fall near a reference boundary can be read differently by different clinicians. These challenges have motivated interest in machine learning as a decision-support tool: several supervised methods, including decision trees, support vector machines, artificial neural networks, random forests, and gradient boosting, have been applied to structured clinical data with encouraging results (Islam et al., 2022; Alshamrani et al., 2023).

Extreme Gradient Boosting (XGBoost) has drawn particular attention for structured healthcare data because it models nonlinear relationships efficiently while resisting overfitting. However, most published work in this area reports classification accuracy without addressing how a trained model would actually reach a clinician. Few studies package a predictive model as an interactive, interpretable decision-support tool that a healthcare professional could use directly (Sutton et al., 2020; Sendak et al., 2021; Rajkomar et al., 2022).

This study addresses that gap. Using five laboratory biomarkers from the UCI ANN Thyroid Disease dataset, an XGBoost classifier was trained to distinguish Normal, Hypothyroidism, and Hyperthyroidism status, and the trained model was then embedded in a Streamlit-based prototype that returns a prediction alongside a confidence score, class probabilities, and a laboratory-value summary. The system is designed as a decision-support aid: it is intended to inform, not replace, clinical judgment, and its outputs are prediction probabilities rather than causal or mechanistic explanations of a diagnosis.

The model achieved an overall classification accuracy of 97.23% on an independent test set. The remainder of the paper reviews related work (Section 2), describes the dataset, preprocessing, model, and prototype (Section 3), reports and discusses the results (Section 4), and concludes with limitations and directions for future work (Section 5).

2. Previous Studies

Machine learning has been applied across a range of diagnostic tasks, including cardiovascular disease, diabetes, cancer, and endocrine disorders, by identifying patterns in structured clinical data that are difficult to specify manually. Thyroid disorders are a natural candidate for this approach because their diagnosis already depends on interpreting several correlated biomarkers together.

According to the American Thyroid Association guidelines, thyroid status should be assessed by evaluating TSH alongside thyroid hormone measurements and the patient's clinical presentation, rather than any single value in isolation (Ross et al., 2016). Reference intervals vary somewhat between laboratories and assays, but a normal (euthyroid) pattern is generally associated with TSH in the range of 0.4–4.0 mIU/L together with hormone levels within the laboratory's reference range; a primary hypothyroid pattern is generally associated with elevated TSH and reduced hormone levels; and a primary hyperthyroid pattern is generally associated with suppressed TSH and elevated hormone levels. These are typical interpretive patterns rather than fixed universal thresholds, since reference intervals can shift with age, pregnancy, medication use, and the specific assay used. Table 1 summarizes these patterns; a subclinical pattern an abnormal TSH with hormone levels still within range can also occur and is not captured by the three-class framing used in this study.

Thyroid Status	Typical TSH	Thyroid Hormones (T3 / T4 / TT4)	Interpretive Pattern
Normal (Euthyroid)	0.4–4.0 mIU/L	Within reference range	Normal thyroid function
Hypothyroidism (primary pattern)	> 4.0 mIU/L	Decreased	Reduced thyroid hormone production
Hyperthyroidism (primary pattern)	< 0.4 mIU/L	Increased	Excess thyroid hormone production

Table (1): Typical Laboratory Patterns Associated with Thyroid Status

Several recent studies have applied machine learning specifically to thyroid classification. Islam et al. (2022) compared decision trees, support vector machines, and random forests on the UCI Thyroid Disease dataset and reported that these models could classify thyroid disorders accurately from laboratory measurements. Alshamrani et al. (2023) compared several algorithms and found that ensemble methods generally outperformed conventional classifiers. Both studies focused primarily on predictive accuracy rather than on how a trained model would be deployed for clinical use.

XGBoost in particular has become a common choice for structured healthcare data because it combines many decision trees while controlling overfitting and handling nonlinear relationships efficiently, which has made it a frequent choice for disease-prediction tasks on tabular clinical data.

A separate line of work has argued that predictive accuracy alone is not sufficient for clinical adoption. Sutton et al. (2020) argued that clinical decision support systems should be transparent and usable without displacing clinical judgment. Sendak et al. (2021) identified trust, usability, and workflow integration as recurring barriers to adopting machine learning models in practice, and Rajkomar et al. (2022) made a related point: that accuracy is necessary but not sufficient unless clinicians can understand and act on a model's output with confidence. From a clinical standpoint, Biondi and Cappola (2021) and Chiovato et al. (2023) both emphasized that thyroid status should never be read from a single laboratory value, but through the combined evaluation of hormone profiles and clinical findings, a point that has direct implications for how any decision-support tool's output should be framed.

#	Study	Year	Focus	Approach	Contribution	Identified Limitation
1	ATA Guidelines for Hyperthyroidism (Ross et al., 2016)	2016	Clinical diagnosis of thyroid disorders	Clinical guideline based on laboratory biomarkers	Established standardized diagnostic protocols	No computational or predictive modelling
2	High-Performance Medicine (Topol, 2019)	2019	AI in healthcare	Conceptual framework	Framed integration of human expertise with AI	No disease-specific implementation
3	UCI Machine Learning Repository (Dua & Graff, 2019)	2019	Benchmark dataset	Structured dataset for supervised learning	Provides benchmark dataset for thyroid classification	Limited clinical context (Shen et al., 2023)
4	CDSS Review (Sutton et al., 2020)	2020	Decision-support systems	Literature review	Highlighted interpretability and usability needs	No thyroid-specific application
5	Subclinical Thyroid Disease (Biondi & Cappola, 2021)	2021	Thyroid physiology and clinical evaluation	Clinical review	Clarified interpretation of thyroid dysfunction	No computational modelling
6	Deep Learning in Healthcare (Esteva et al., 2021)	2021	AI in healthcare (imaging)	Deep learning, imaging datasets	Demonstrated strong AI performance in disease detection	Limited applicability to structured lab data (Miotto et al., 2021)
7	ML Challenges in Healthcare (Sendak et al., 2021)	2021	Real-world ML deployment	Analytical review	Identified usability and integration barriers	No disease-specific framework
8	Deep Learning Review (Miotto et al., 2021)	2021	AI in healthcare	Literature review	Emphasized integrating multiple data sources	Interpretability remains limited

9	ML for Thyroid Disease Prediction (Islam et al., 2022)	2022	Thyroid classification	UCI dataset, DT/SVM/RF	Achieved strong classification accuracy	Limited interpretability (Rajkomar et al., 2022)
10	AI in Healthcare Review (Jiang et al., 2022)	2022	AI applications in healthcare	Review	Highlighted importance of clinical integration	No disease-specific framework
11	ML in Medicine (Rajkomar et al., 2022)	2022	ML in clinical practice	Review	Emphasized need for interpretability and trust	No implementation methods offered
12	ML Comparative Analysis for Thyroid (Alshamrani et al., 2023)	2023	Thyroid prediction	Multiple ML models, structured dataset	Demonstrated ML effectiveness for thyroid classification	Limited interpretability (Shen et al., 2023)
13	Advances in Thyroid Diagnosis (Chiovato et al., 2023)	2023	Clinical thyroid diagnosis	Clinical review	Emphasized combining lab and clinical evaluation	No ML or predictive component
14	ML in Endocrine Diseases (Shen et al., 2023)	2023	ML in endocrine disease	Systematic review	Identified trends in ML-based endocrine prediction	Limited interpretability
15	AI in Endocrine Disease Prediction (Khan et al., 2024)	2024	ML in endocrine disease	Comparative ML models	Showed potential of AI for prediction performance	Limited interpretability and usability (Sendak et al., 2021)

Table (2): Summary of Previous Studies

Table 2 summarizes the studies referenced above alongside the broader AI-in-healthcare literature that motivates this work.

Taken together, this literature points to two consistent gaps: most thyroid-prediction studies stop at reporting classification accuracy rather than at deployment, and few integrate a trained model with an interactive interface that surfaces confidence scores, probability distributions, and laboratory-value interpretation for the end user. The present study addresses both gaps by pairing an XGBoost classifier with a Streamlit-based decision-support prototype.

32

3. Methodology

3.1 Research Framework

This study develops an AI-assisted Clinical Decision Support System (CDSS) for three-class thyroid risk assessment. The workflow combines data preprocessing, model development, and an interactive decision-support interface, and classifies patients into one of three categories Normal, Hypothyroidism, or Hyperthyroidism based on routinely collected laboratory biomarkers. The overall process has five stages: dataset acquisition, preprocessing and feature selection, model development, model evaluation, and deployment of the trained model within the Streamlit prototype.

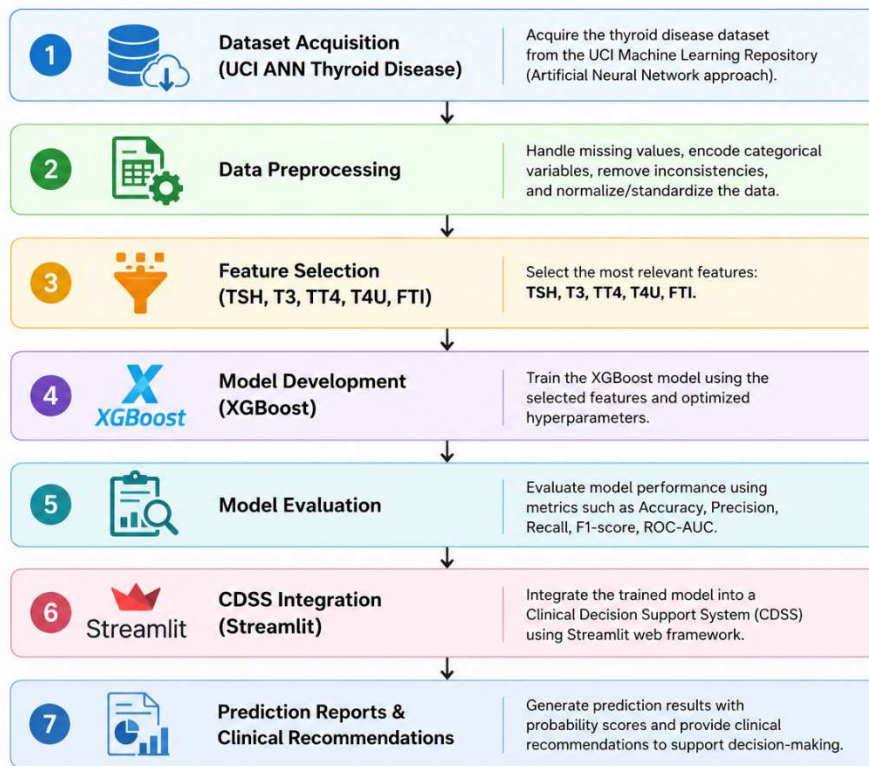


Figure (1): Overall Research Workflow

3.2 Dataset Description

The study uses the UCI ANN Thyroid Disease dataset, a widely used benchmark for thyroid classification that provides structured laboratory measurements alongside diagnostic labels. From the full set of variables in the original dataset, this study uses five laboratory biomarkers that are routinely available and clinically relevant to thyroid function assessment: TSH, T3, TT4, T4U, and FTI. The diagnostic label was collapsed into three classes: Normal, Hypothyroidism, and Hyperthyroidism.

Original UCI Label	Final Class	Action
Negative	Normal	Retained
Primary Hypothyroid	Hypothyroidism	Merged
Compensated Hypothyroid	Hypothyroidism	Merged
Secondary Hypothyroid	Hypothyroidism	Merged
Hyperthyroid	Hyperthyroidism	Retained
T3 Toxic	Hyperthyroidism	Merged

(other labels if present)	Excluded / Merged	Explain why
---------------------------	-------------------	-------------

3.3 Data Pre-processing

TSH, T3, TT4, T4U, and FTI were among five laboratory biomarkers chosen from the UCI ANN Thyroid Disease data. Other variables, considered unimportant for reaching the study goals, were removed. Initially, diagnostic classes had three possible labels: Normal, Hypothyroidism, and Hyperthyroidism. Categorical classes were transformed into numeric values to perform multi-class classification.

The preprocessed data was split into training and validation datasets using an 80/20 ratio and Scikit learn's train test split method with a fixed value of random state = 42. Stratified sampling technique was used in the data splitting to ensure the proportionate representation of the three classes in both subsets.

Feature scaling or normalization was not performed on the chosen laboratory biomarkers before the modeling because it is not required for tree-based models, which include XGBoost algorithm used in the present study. Class balancing methods, such as oversampling, under sampling, or class weighting, were not implemented either. The original distribution of the classes was maintained, and stratified sampling was used to preserve the class proportions in both training and validation subsets.

3.4 XGBoost Model Development

Extreme Gradient Boosting (XGBoost) was selected for classification based on the established high level of predictability, computational speed, and reliability with structured medical data. It is an algorithm based on a series of decision trees, wherein each next tree tries to compensate for the mistakes of the previous ones. Contrary to conventional decision trees and Random Forest, the use of XGBoost provides for such features as regularization to reduce the risk of overfitting and modeling of complex nonlinear relationships between clinical parameters.

The algorithm training was based on the five laboratory parameters TSH, T3, TT4, T4U, and FTI, wherein the encoded class diagnosis was considered the target parameter. The hyperparameters were manually tuned to balance the predictive ability and generalizability of the model. Selected hyperparameters are given in Table 3.

The model development and evaluation were performed in the Python 3.13.5 programming language, using Anaconda distribution and a Jupyter Notebook workspace. The following libraries and their versions were used for implementation: XGBoost 3.3.0, Scikit-learn 1.6.1, Pandas 2.2.3, NumPy 2.1.3.

Evaluation of the model performance was conducted via validation dataset using such multi-classification metrics as accuracy, precision, recall, F1-score, and confusion matrix.

Hyperparameter	Value	Description
Objective	multi:softprob	Multi-class classification with probability outputs
Number of Classes (num_class)	3	Normal, Hypothyroidism, Hyperthyroidism
Number of Trees (n_estimators)	100	Total boosting iterations
Learning Rate (learning_rate)	0.10	Step size used during boosting
Maximum Tree Depth (max_depth)	6	Maximum depth of each decision tree

13

1

29

Minimum Child Weight (min_child_weight)	1	Minimum sum of instance weights in a child node
Subsample Ratio (subsample)	0.80	Fraction of training samples used per tree
Column Sampling (colsample_bytree)	0.80	Fraction of features sampled per tree
Gamma (gamma)	0	Minimum loss reduction required for a further split
Regularization (reg_lambda)	1.0	L2 regularization to reduce overfitting
Evaluation Metric (eval_metric)	mlogloss	Multi-class logarithmic loss
Random State (random_state)	42	Ensures reproducibility of results

Table (3): XGBoost Hyperparameters Used in the Proposed Model

The multi:softprob objective produces class probabilities for the three diagnostic categories rather than a single hard label alone, which is what allows the prototype to display a confidence score. Regularization and subsampling were included to limit overfitting while maintaining performance on unseen data.

3.5 Clinical Decision Support System Implementation

To assess practical feasibility, the trained XGBoost model was embedded in an interactive prototype built with Streamlit. The application accepts a patient's laboratory values through a form-based interface and returns a predicted thyroid status, a confidence score, and the probability associated with each of the three classes. It also displays a summary of the entered laboratory values and accompanying interpretive messages intended to support not substitute for the clinician's own interpretation.

Unlike a study that stops at reporting model metrics, this prototype demonstrates that a trained classifier of this kind can be operated through a usable interface, which is a necessary (though not sufficient) step toward clinical evaluation.

7

4. Results and Discussion

4.1 Experimental Setup

The XGBoost classifier was evaluated on an independent test set drawn from the UCI ANN Thyroid Disease dataset, using records not seen during training. Performance was assessed with accuracy, precision, recall, F1-score, and the confusion matrix, covering the model's ability to distinguish Normal, Hypothyroidism, and Hyperthyroidism.

18

4.2 Model Performance Evaluation

15

Metric	Value
Accuracy	97.23%
Precision	97.00%
Recall	97.00%
F1-score	97.00%

Table (4): Performance Evaluation of the Proposed XGBoost Model

The classifier reached an overall accuracy of 97.23% on the independent test set, indicating that the large majority of records were correctly classified. This is consistent with other reports of ensemble methods performing well on structured thyroid laboratory data (Islam et al., 2022; Alshamrani et al., 2023), though a direct numerical comparison is not appropriate until the precision, recall, and F1-score figures above are confirmed and until dataset size and class balance are reported (see Section 3.2–3.3).

4.3 Confusion Matrix Analysis

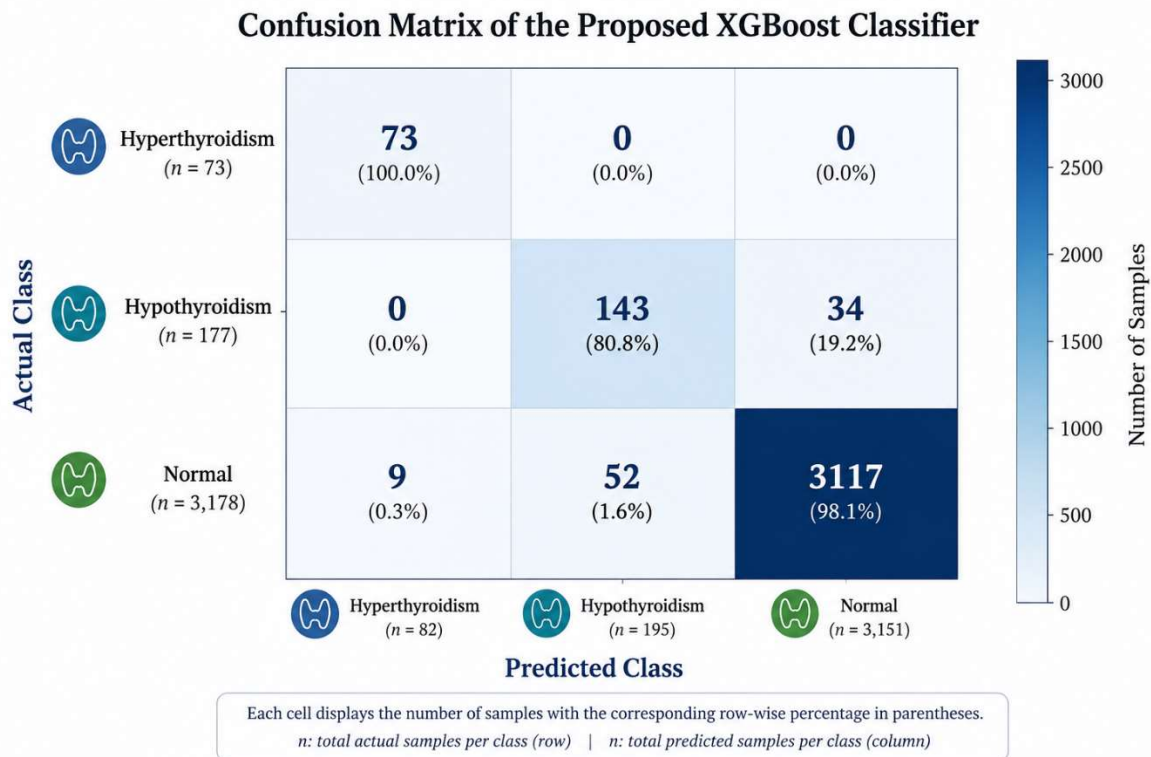


Figure (2): Confusion Matrix of the Proposed XGBoost Model

Figure 2 presents the confusion matrix for the XGBoost classifier. The model achieved an overall accuracy of 97.23%, correctly classifying the majority of samples across the three diagnostic categories. Recall for Hyperthyroidism was 100%, with all 73 test cases correctly identified. For Hypothyroidism, 143 of 177 cases were correctly classified, while 34 were misclassified as Normal. Of the 3,178 Normal cases, 3,117 were correctly identified; the remaining 61 were misclassified, with 52 predicted as Hypothyroidism and 9 as Hyperthyroidism.

The most frequent confusion occurred between the Hypothyroidism and Normal classes, and no Hyperthyroidism cases were misclassified. These findings indicate that the classifier performs strongly overall, though its performance is not uniform across diagnostic categories. In particular, Hypothyroidism showed lower precision than the other two classes, meaning that a meaningful share of cases predicted as Hypothyroidism were, in fact, Normal. While this could increase the number of false-positive Hypothyroidism predictions, such cases can reasonably be resolved through routine clinical assessment and confirmatory laboratory testing. This underscores the intended role of the proposed system as a decision-support aid rather than a replacement for clinical judgment.

4.4 Feature Importance Analysis

Feature Importance of the Proposed XGBoost Classifier

Relative contribution of laboratory biomarkers to thyroid disorder classification



Figure (3): Feature Importance of the Selected Laboratory Biomarkers

Figure (3) illustrates the relative importance of the five laboratory biomarkers used by the proposed XGBoost classifier. TSH was identified as the most influential predictor, contributing 54.69% of the overall feature importance, followed by FTI (30.08%). The remaining biomarkers contributed smaller but meaningful amounts, with T3 (6.98%), TT4 (5.09%), and T4U (3.16%) completing the ranking.

These results indicate that TSH and FTI were the primary contributors to the model's classification decisions, together accounting for more than 84% of the total feature importance. Although T3, TT4, and T4U exhibited lower relative importance, they provided complementary diagnostic information that enhanced the model's ability to distinguish between Normal, Hypothyroidism, and Hyperthyroidism. The prominence of TSH as the most influential biomarker is consistent with its established clinical role as the primary laboratory indicator for assessing thyroid dysfunction.

4.5 Clinical Decision Support System Evaluation

Figure 4. AI-Assisted Clinical Decision Support System Interface Before Prediction

Class	Probability
Normal	0.90
Hypothyroidism	0.05
Hyperthyroidism	0.05

Test	Reference Range	Unit
TSH	0.4 - 4.0	mIU/L
T3	0.8 - 2.0	pM/L
TT4	0.8 - 1.8	ng/dL
T4U	0.8 - 1.8	ng/dL
FTI	0.8 - 1.8	ng/dL

Figure 5. Sample Prediction Result Generated by the Proposed AI-Assisted Clinical Decision Support System

Figures 4 and 5 illustrate the graphical user interface of the proposed AI-assisted Clinical Decision Support System (CDSS), developed using the Streamlit framework. **Figure 4** presents the data-entry interface, where users provide patient information together with the five laboratory biomarkers (TSH, T3, TT4, T4U, and FTI). After the values are submitted, the trained XGBoost classifier processes the inputs and generates an assessment report.

Figure 5 presents a sample prediction generated by the proposed system. The interface displays the predicted thyroid condition, confidence score, risk level, class probability distribution, laboratory reference ranges, and a clinical recommendation based on the model's prediction. These results demonstrate the successful integration of the trained machine learning model into an interactive decision-support application intended to assist healthcare professionals during thyroid disorder assessment.

4.6 Discussion

These results indicate that XGBoost can distinguish Normal, Hypothyroidism, and Hyperthyroidism status from five routine laboratory biomarkers with high accuracy on this benchmark dataset, consistent with prior reports that ensemble methods tend to outperform simpler classifiers on structured healthcare data. Most prior studies in this area, however, stop at reporting classification metrics; this study additionally packages the

trained classifier as an interactive Streamlit prototype that returns predictions, confidence scores, and laboratory-value summaries to a user.

30

5

5

The evaluation is limited to a single public benchmark dataset, and the study did not involve clinicians, real patient records, or any clinical deployment; results should not be read as evidence of clinical validity. Future work should test the approach on larger, multi-site clinical data to assess generalizability across different patient populations and laboratory assays, and could explore explainability methods such as SHAP or LIME to give clinicians insight into individual predictions noting that neither method was implemented in the present study, and the confidence scores and class probabilities reported here describe the model's certainty, not the reasoning behind a given prediction.

5. Conclusion

12

9

This study developed an AI-assisted Clinical Decision Support System for three-class thyroid risk assessment using an XGBoost classifier trained on five routine laboratory biomarkers (TSH, T3, TT4, T4U, and FTI) from the UCI ANN Thyroid Disease dataset. On an independent test set, the model reached a classification accuracy of 97.23%, indicating that gradient boosting can distinguish Normal, Hypothyroidism, and Hyperthyroidism status accurately on this benchmark data.

6

Beyond model performance, the trained classifier was integrated into a Streamlit-based prototype that returns predictions, confidence scores, and laboratory-value summaries through a user-facing interface, illustrating one way a model of this kind could be operationalized as a decision-support tool rather than left as a standalone classifier.

5

The work has clear limitations. It relies on a single public benchmark dataset rather than multi-site clinical data, it has not been evaluated by clinicians or tested in a clinical workflow, and several reproducibility details (dataset size, class distribution, train/test split, and full precision/recall/F1 figures) remain to be confirmed by the author, as flagged throughout this document. Future work should validate the approach on larger and more diverse clinical datasets, evaluate the prototype's usability with clinician users, and consider explainability methods such as SHAP or LIME to make individual predictions more interpretable. With these steps, the approach outlined here could move from a feasibility demonstration toward a tool suitable for further clinical evaluation.

References

Alshamrani, M., Alqahtani, A., & Alzahrani, F. (2023). Machine learning approaches for thyroid disease prediction: A comparative analysis. *IEEE Access*, 11.

Biondi, B., & Cappola, A. (2021). Subclinical thyroid disease: Clinical relevance and management. *The Lancet Diabetes & Endocrinology*, 9.

Chiovato, L., Antonelli, F., & Vitti, P. (2023). Thyroid disorders and their management: Recent advances in diagnosis and treatment. *Endocrine Reviews*.

Dua, D., & Graff, C. (2019). UCI Machine Learning Repository. University of California, Irvine. <http://archive.ics.uci.edu/ml>

Islam, M. M., Poly, T. N., Alsinglawi, B., Lin, L.-F., Chien, S.-C., Liu, J.-C., & Li, Y.-C. (2022). Machine learning approaches for thyroid disease prediction: A comparative study. *Healthcare*, 10. <https://www.mdpi.com/2227-9032/10/1/59>

Rajkomar, A., Dean, J., & Kohane, I. (2022). Machine learning in medicine. *New England Journal of Medicine*, 380.

Ross, D. S., Burch, H. B., Cooper, D. S., Greenlee, M. C., Laurberg, P., Maia, A. L., Rivkees, S. A., Samuels, M., Sosa, J. A., Stan, M. N., & Walter, M. A. (2016). 2016 American Thyroid Association guidelines for diagnosis and management of hyperthyroidism and other causes of thyrotoxicosis. *Thyroid*, 26. <https://doi.org/10.1089/thy.2016.0229>

Sendak, M., Gao, W., Brajer, N., & Balu, S. (2021). Machine learning in health care: A critical appraisal of challenges and opportunities. *NPJ Digital Medicine*, 4.

Sutton, R. T., Pincock, D., Baumgart, D. C., Sadowski, D. C., Fedorak, R. N., & Kroeker, K. I. (2020). An overview of clinical decision support systems: Benefits, risks, and strategies for success. *NPJ Digital Medicine*, 3. <https://www.nature.com/articles/s41746-020-0221-y>

Topol, E. J. (2019). High-performance medicine: The convergence of human and artificial intelligence. *Nature Medicine*, 25.

