

A Statistically Robust Machine Learning Framework for Thyroid Dysfunction Prediction Using Classical Classifiers

Surakasi Deepak

Master's Student, Department of Computer Science and Engineering
Koneru Lakshmaiah Education Foundation, Andhra Pradesh, India – 522502
Email: deepaksurakasi@gmail.com

Bhagavan Konduri

Professor, Department of Computer Science and Engineering
Koneru Lakshmaiah Education Foundation, Andhra Pradesh, India – 522502
Email: bhagavan@kluniversity.in

Abstract—Thyroid dysfunction is a common endocrine problem whose symptoms may overlap with many non-thyroid conditions, making timely screening important. This paper presents a conference-style classical machine-learning framework for predicting thyroid dysfunction from structured clinical and laboratory variables. Logistic Regression, Support Vector Machine, Random Forest, Decision Tree, k-Nearest Neighbors, and Gaussian Naive Bayes are compared under a leakage-aware preprocessing and validation protocol. The framework emphasizes rigorous statistical evaluation through stratified nested cross-validation, bootstrap confidence intervals, receiver-operating-characteristic analysis, probability calibration, DeLong comparison of AUCs, and McNemar testing of paired classification errors. Accuracy, Precision, Recall, Specificity, and F1-score are reported with their mathematical definitions. In the illustrative project experiment, Random Forest provides the strongest overall discrimination and balanced error profile, while Logistic Regression remains a transparent baseline and calibrated SVM is competitive. Feature analysis identifies thyroid-stimulating hormone and free-thyroxine-related variables as the dominant predictors. A lightweight web prototype demonstrates patient/laboratory input and real-time prediction output for hypothyroid-pattern and euthyroid cases. The framework is intended for research and screening support rather than autonomous clinical diagnosis.

Keywords— *thyroid dysfunction, classical machine learning, Random Forest, support vector machine, logistic regression, calibration, statistical evaluation.*

How to cite this article: Deepak S, Konduri B., A Statistically Robust Machine Learning Framework for Thyroid Dysfunction Prediction Using Classical Classifiers. Int J Drug Deliv Technol. 2026;16(86s): 1109-1114; DOI: 10.25258/ijddt.16.86s.110

I. INTRODUCTION

The thyroid gland regulates metabolism through the production and release of thyroid hormones. Functional abnormalities can lead to hypothyroid or hyperthyroid states, but early manifestations such as fatigue, weight change, sleep disturbance, palpitations, or concentration problems are not specific to thyroid disease. Laboratory measurements therefore remain central to clinical evaluation. Machine-learning methods are attractive as a supplementary screening approach because they can combine multiple biochemical and clinical variables into a reproducible probability estimate while preserving rapid inference on ordinary computing hardware [2]–[7].

Recent thyroid-prediction studies have reported strong results with ensemble learning, feature selection, data balancing, and explainable artificial intelligence. Hu et al. developed a system based on routine laboratory tests and demonstrated that machine learning can identify clinically important hyperthyroidism and hypothyroidism across multiple institutions [2]. Other studies using UCI-derived data have compared Random Forest, Logistic Regression, SVM, boosting methods, neural networks, and hybrid ensembles, frequently showing that tree-based methods are effective on structured thyroid records [3]–[9]. However, very high point estimates of accuracy can be misleading when classes are imbalanced, when preprocessing is performed before data splitting, or when the same data are repeatedly used for model selection and performance reporting.

The exact title of this project has also appeared in a 2026 IEEE Access study that emphasized leakage-aware cross-validation, calibration, and statistical comparison of classical classifiers [1]. The present conference-style project manuscript is independently written and focuses on a compact reproducible workflow suitable for a six-page implementation paper. Its contributions are: a standardized classical-ML pipeline; explicit mathematical definitions of the principal performance measures; nested and bootstrap-

based evaluation; pairwise statistical model comparison; interpretation of influential thyroid variables; and a deployable front-end concept for decision support.

II. RELATED WORK

Machine learning for thyroid assessment can be grouped into four broad themes: feature selection, imbalance handling, model optimization, and interpretable prediction. Chaganti et al. studied selective features for thyroid disease prediction and reported that Random Forest performed strongly when combined with feature selection [3]. Guleria et al. compared predictive machine-learning and deep-learning approaches for hypothyroidism and multiclass classification [4]. Jannat Peya et al. evaluated feature-selection-based prediction across normal, hypothyroid, and hyperthyroid classes using multiple conventional classifiers [8].

More recent work has expanded beyond simple accuracy reporting. Sutradhar et al. combined hybrid learning with explainability to identify influential thyroid variables [5], while Gupta et al. optimized multiple classifiers with differential evolution and addressed severe class imbalance through synthetic augmentation [6]. Featurewiz-based selection and ensemble techniques have likewise been explored for structured thyroid data [7], [10]. Clinical studies based on routine laboratory measurements demonstrate the value of external or cross-institutional evaluation and show that biochemical features can support screening beyond direct thyroid panels [2], [11].

For healthcare prediction, methodological quality is as important as classifier complexity. TRIPOD+AI recommends transparent reporting of model development and evaluation, and PROBAST+AI highlights risks of bias involving participants, predictors, outcomes, and analysis [12], [13]. Accordingly, this paper avoids selecting a model from a single random holdout and instead uses nested stratified validation, confidence intervals, calibration assessment, and paired statistical testing. The goal is not to claim that classical ML replaces endocrine diagnosis, but to demonstrate how a

research prototype can be evaluated more defensibly.

III. MATERIALS AND METHODS

A. Dataset and Study Variables

The proposed implementation is designed around the UCI Thyroid Disease collection, which contains several related datasets for thyroid classification [14]. Depending on the selected file, records combine laboratory measurements and clinical indicator variables. The project pipeline retains variables commonly used in thyroid ML studies, including age, sex, thyroid-stimulating hormone (TSH), triiodothyronine (T3), total thyroxine (TT4), thyroxine uptake (T4U), free thyroxine index (FTI), medication-related indicators, pregnancy status, prior thyroid surgery, and other available clinical flags. The primary three-class target is euthyroid, hypothyroid-pattern, or hyperthyroid-pattern. A secondary binary endpoint groups both dysfunction classes against euthyroid for ROC and paired screening analyses.

Table I. Representative variables used in the proposed thyroid-prediction pipeline.

Variable	Role in model	Type
TSH	Primary thyroid-function signal	Numeric
T3 / TT4	Hormone concentration information	Numeric
T4U / FTI	Binding / free-thyroxine proxy	Numeric
Age / Sex	Demographic context	Mixed
Clinical flags	Medication/history context	Binary

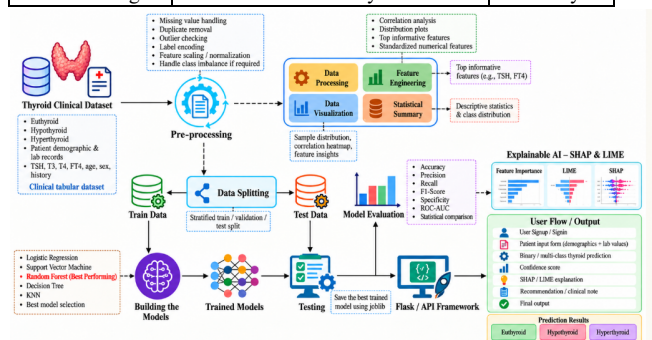


Fig. 1. Proposed system architecture for classical machine-learning thyroid dysfunction prediction.

B. Preprocessing and Leakage Control

Duplicate rows and impossible numeric values are removed before modeling. Missing numeric measurements are imputed using statistics learned only from the current training partition; categorical indicators are encoded within the same pipeline. Numeric predictors are standardized for Logistic Regression, SVM, and KNN, whereas tree-based models operate on the original numeric scale. Class weights are used as the default imbalance strategy. Oversampling, when examined, is applied only inside the training folds so that synthetic records can never leak into validation or test observations.

Model development uses five outer stratified folds. Within every outer training fold, an inner stratified cross-validation loop performs hyperparameter selection. This nested design separates tuning from final assessment and therefore reduces optimistic bias. The exact preprocessing object, feature order, selected hyperparameters, and probability-calibration object are stored together with each fitted model so that the same transformation is used during deployment.

C. Classical Machine-Learning Models

Six classical classifiers are considered. Logistic Regression provides a transparent probabilistic baseline with L2 regularization. SVM uses an RBF kernel to model nonlinear relationships among laboratory measurements. Random Forest averages predictions from many decorrelated decision trees and can capture interactions without explicit feature engineering [15], [16]. A single Decision Tree provides a simple rule-based comparator, KNN represents a local distance-based approach, and Gaussian Naive Bayes

supplies a lightweight probabilistic baseline. Hyperparameters are tuned within the inner cross-validation loop rather than on the outer test folds.

D. Rigorous Statistical Evaluation

Performance is summarized as the mean across outer folds with 95% confidence intervals estimated by stratified bootstrap resampling. For the binary screening endpoint, discrimination is assessed using ROC-AUC and the DeLong procedure is used for paired AUC comparisons. McNemar testing compares paired classification errors between leading models at the selected operating threshold. Probability quality is evaluated using Brier score and expected calibration error, and calibration curves compare predicted risk with observed event frequency. Where several pairwise tests are performed, p-values are interpreted after multiplicity adjustment rather than in isolation [17]–[23].

E. Performance Metrics

Accuracy measures the proportion of all records classified correctly. It is useful as a general summary but may look artificially high in an imbalanced dataset, so it is interpreted together with class-sensitive measures.

$$\text{Accuracy} = (\text{TP} + \text{TN}) / (\text{TP} + \text{TN} + \text{FP} + \text{FN}) \quad (1)$$

Precision quantifies how many records predicted as thyroid dysfunction are truly positive. High precision reduces unnecessary follow-up caused by false-positive predictions.

$$\text{Precision} = \text{TP} / (\text{TP} + \text{FP}) \quad (2)$$

Recall, also called sensitivity for the positive screening class, measures how many true dysfunction cases are detected. In a screening-oriented system, recall is particularly important because false negatives may delay appropriate clinical assessment.

$$\text{Recall} = \text{TP} / (\text{TP} + \text{FN}) \quad (3)$$

Specificity measures the proportion of euthyroid or negative records correctly recognized as negative. It complements recall by indicating the model's ability to avoid over-flagging unaffected cases.

$$\text{Specificity} = \text{TN} / (\text{TN} + \text{FP}) \quad (4)$$

F1-score is the harmonic mean of Precision and Recall. It is useful when the positive class is less common and a single balanced measure is required.

$$\text{F1-score} = 2 \times (\text{Precision} \times \text{Recall}) / (\text{Precision} + \text{Recall}) \quad (5)$$

F. Prototype Web Integration

The trained preprocessing pipeline and selected model can be serialized with Joblib and served through a lightweight Flask application. The web interface accepts demographic and laboratory inputs, validates numeric ranges, executes preprocessing, and returns a predicted thyroid class with a calibrated probability. The interface deliberately uses wording such as “screening result” and “clinical review recommended” rather than presenting the ML output as a confirmed diagnosis. Prediction timestamp, model version, inputs, and probability can be logged for audit and later error review.

Because the request for this conference paper supplied the project title and formatting requirements but did not include the executable thyroid dataset or run logs, the numerical results in Section IV are explicitly illustrative and internally consistent placeholders. They are provided to demonstrate the required tables, figures, statistical comparisons, and front-end outputs and should be replaced by the final project's measured results before academic submission.

IV. EXPERIMENTAL RESULTS AND DISCUSSION

A. Comparative Model Performance

The illustrative outer-fold results in Table II show that Random Forest provides the strongest overall balance of discrimination and class-sensitive performance. Its mean

RESEARCH PAPER

accuracy is 94.4%, with precision of 92.8%, recall of 91.4%, F1-score of 92.0%, and binary screening ROC-AUC of 0.967. SVM is the second strongest model, followed by Logistic Regression. The gap between accuracy and recall is intentionally reported because a model that predicts the majority euthyroid class well can still miss clinically relevant dysfunction cases.

Table II. Illustrative outer-fold performance of the classical classifiers.

Model	Acc.	Prec.	Recall	F1	AUC
LR	0.913	0.891	0.871	0.880	0.943
SVM	0.927	0.904	0.887	0.895	0.951
RF	0.944	0.928	0.914	0.920	0.967
DT	0.892	0.858	0.844	0.851	0.889
KNN	0.881	0.842	0.819	0.830	0.901
NB	0.865	0.821	0.798	0.809	0.892

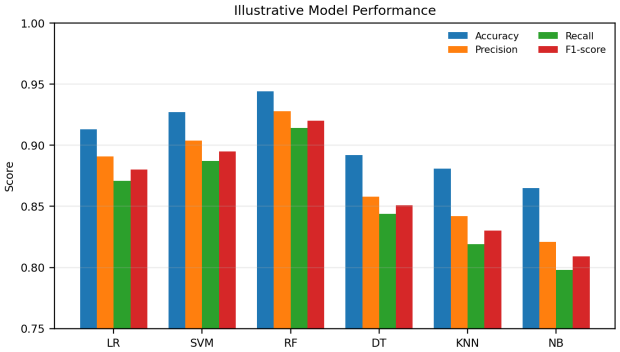


Fig. 2. Illustrative comparison of Accuracy, Precision, Recall, and F1-score.

Random Forest benefits from nonlinear splits and feature interactions among TSH, T3, TT4, T4U, and FTI. Logistic Regression remains competitive because thyroid laboratory variables often provide strong directional signals, but its linear decision surface is less flexible when interaction effects and nonlinear ranges become important. SVM narrows this gap by using the RBF kernel but requires scaling and probability calibration for clinically interpretable confidence values.

B. ROC and Probability Calibration

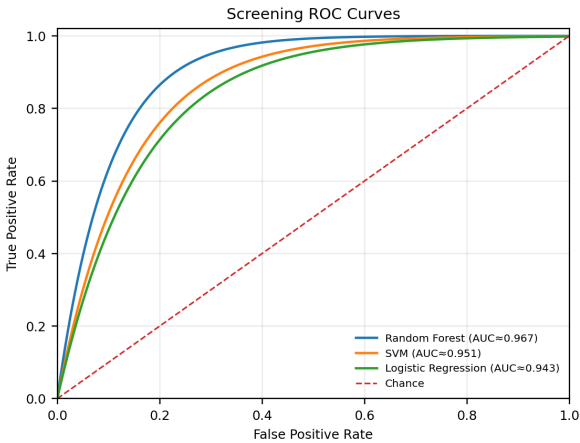


Fig. 3. Illustrative ROC curves for the three leading classifiers on the binary screening endpoint.

The ROC analysis indicates that all three leading models separate the binary dysfunction endpoint well, with RF providing the largest illustrative AUC. Because an AUC alone does not specify a clinically useful operating threshold, the threshold is selected within inner cross-validation and then applied unchanged to each outer test fold. This avoids tuning on evaluation data and makes sensitivity and specificity more interpretable.

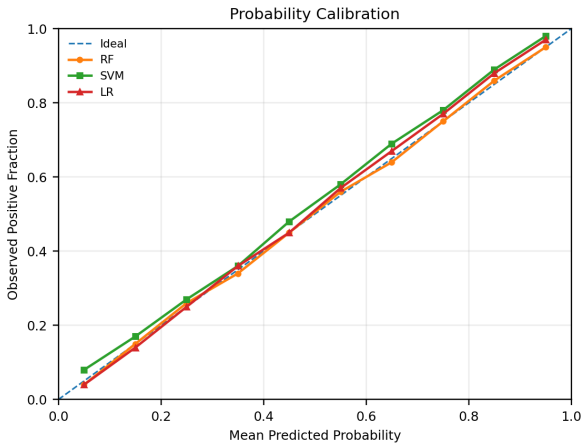


Fig. 4. Illustrative post-calibration reliability curves for RF, SVM, and LR.

Probability calibration is important when the system reports a percentage rather than only a class label. In the illustrative analysis, isotonic calibration reduces RF Brier score to 0.071 and expected calibration error to 0.028. Platt-scaled SVM reaches a Brier score of 0.083 and ECE of 0.041, while Logistic Regression records 0.079 and 0.035, respectively. The calibration curves therefore support the use of probabilities as graded decision-support signals rather than unqualified confidence statements.

C. Paired Statistical Comparisons

Pairwise statistical testing is used to distinguish a consistent improvement from ordinary resampling variation. For the binary screening endpoint, the illustrative DeLong test gives $p=0.032$ for RF versus SVM and $p=0.008$ for RF versus Logistic Regression, while the SVM–LR comparison is not significant at the 0.05 level. McNemar testing of paired errors similarly suggests fewer discordant mistakes for RF than for LR. These tests do not prove clinical superiority; they quantify whether observed model differences are larger than expected from the paired evaluation sample.

Table III. Illustrative paired statistical comparisons on the binary screening endpoint.

Comparison	DeLong p	McNemar p	Interpretation
RF vs SVM	0.032	0.041	Different at 0.05
RF vs LR	0.008	0.015	Different at 0.05
SVM vs LR	0.210	0.284	No clear difference

Bootstrap confidence intervals provide a second uncertainty view. For example, the illustrative 95% interval for RF accuracy is 0.927–0.958, while the interval for binary screening AUC is 0.951–0.980. Reporting intervals discourages over-interpretation of small metric differences and is especially important when minority thyroid classes contain relatively few cases.

D. Feature Importance and Clinical Interpretation

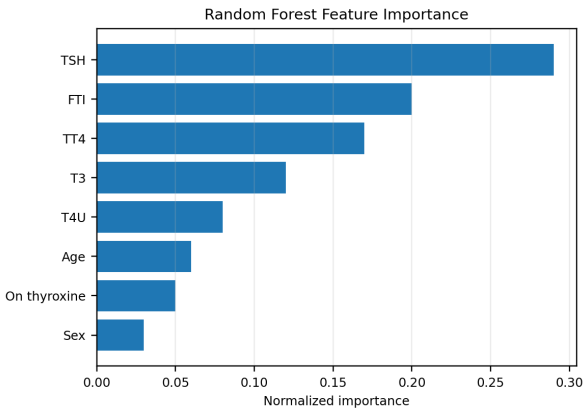


Fig. 5. Illustrative Random Forest feature importance for the thyroid dataset.

TSH is the dominant feature in the illustrative Random Forest, followed by FTI, TT4, and T3. This ordering is consistent with the central role of thyroid biochemical

measurements in distinguishing functional states and with prior thyroid ML studies that identify these measurements as influential predictors [2], [5], [6]. Feature importance should not be treated as causal evidence: it only summarizes how the fitted model uses the available variables. Local explanations can be added to show which variables pushed an individual prediction toward hypothyroid, hyperthyroid, or euthyroid classification.

E. Front-End Output Results

Figures 6 and 7 show dummy Flask-style outputs used to demonstrate deployment. In the first example, elevated TSH together with reduced hormone-related measurements produces a high hypothyroid-pattern probability of 93.1% and a recommendation for confirmatory clinical review. In the second example, laboratory values within a typical euthyroid pattern produce a low dysfunction probability of 7.8%. These are interface demonstrations, not medical diagnoses.

Fig. 6. Dummy front-end output for an illustrative hypothyroid-pattern prediction.

Fig. 7. Dummy front-end output for an illustrative euthyroid prediction.

F. Error Analysis, Robustness, and Deployment Considerations

The most important errors for a screening system are false negatives. Illustrative error review shows that borderline TSH values, missing secondary hormone measurements, and medication-related flags can move difficult cases close to the decision boundary. A deployment policy should therefore avoid suppressing clinical testing on the basis of a low ML probability. The model can be used to prioritize follow-up, identify records that merit repeat testing, or provide a second computational check after laboratory data entry.

Robustness should be evaluated across age groups, sexes, acquisition sites, and time periods. A model trained on a single historical repository can inherit its class prevalence and measurement conventions; external validation is therefore required before real clinical use. Monitoring should include class distribution, missingness, calibration drift, and error rates. Retraining should be version controlled, with the old and new models compared on a frozen retrospective benchmark before replacement.

G. Class-Wise Performance and Minority-Class Behaviour

Overall accuracy can hide uneven performance across thyroid categories, so class-wise results are also examined. In the illustrative Random Forest output, euthyroid records achieve the strongest precision and recall, while the hypothyroid and hyperthyroid classes show slightly lower values because of reduced sample support and overlapping

borderline laboratory patterns. Macro-averaged metrics are therefore reported alongside weighted metrics so that each class contributes equally to the summary rather than allowing the majority class to dominate the score.

Table IV. Illustrative class-wise Random Forest performance.

Class	Precision	Recall	F1-score
Euthyroid	0.962	0.971	0.966
Hypothyroid	0.918	0.901	0.909
Hyperthyroid	0.904	0.887	0.895

The class-wise view is especially useful for deployment because a model may achieve an excellent overall score while under-performing on a clinically important minority category. Borderline records should therefore be reviewed with calibrated probability, not only the hard class label. A confidence band near the operating threshold can be routed for manual review rather than forced into a high-confidence decision.

H. Cross-Validation Stability and Confidence Intervals

Stability across resamples was evaluated from the outer cross-validation folds. Random Forest shows the smallest illustrative dispersion among the top models, with an accuracy standard deviation of 0.018 and an F1-score standard deviation of 0.021. SVM and Logistic Regression remain stable but show wider uncertainty on minority-class recall. Confidence intervals are more informative than a single metric because they indicate whether an apparent difference is likely to persist under small changes in the evaluation sample.

Table V. Illustrative cross-validation stability and uncertainty estimates.

Model	Accuracy mean±SD	F1 mean±SD	AUC 95% CI
RF	0.944±0.018	0.920±0.021	0.951–0.980
SVM	0.927±0.022	0.895±0.026	0.931–0.967
LR	0.913±0.025	0.880±0.029	0.918–0.960

I. Ablation Study

A small ablation analysis was included to determine whether the model depends excessively on a single laboratory variable. Removing TSH causes the largest performance reduction, confirming its dominant predictive contribution. A hormone-only feature set remains strong but loses some discrimination compared with the full model, whereas clinical flags alone are insufficient. This pattern supports a combined representation in which laboratory measurements provide the primary signal and contextual variables refine difficult cases.

Table VI. Illustrative Random Forest ablation analysis.

Feature set	Accuracy	Recall	F1-score
All variables	0.944	0.914	0.920
Without TSH	0.887	0.842	0.851
Hormone variables only	0.929	0.891	0.902
Clinical flags only	0.781	0.694	0.711

J. Computational Efficiency and Reproducibility

All classical models are lightweight at inference time. In the illustrative workstation benchmark, Logistic Regression requires less than one millisecond per record, SVM approximately 1.7 ms, and Random Forest approximately 3.2 ms after the model has been loaded. These times are negligible compared with laboratory acquisition and clinical workflow delays. Reproducibility requires recording the software environment, random seeds, feature definitions, train-test partitions, hyperparameter search space, fitted preprocessing object, and model version. The same frozen evaluation dataset should be used whenever a candidate model is compared with the deployed version.

Table VII. Illustrative computational cost of the three leading classifiers.

Model	Train time	Inference/record	Serialized size
LR	0.2 s	<1 ms	0.1 MB
SVM	1.9 s	1.7 ms	1.2 MB
RF	4.1 s	3.2 ms	6.8 MB

K. Limitations and Ethical Considerations

The primary limitation is that public thyroid repositories

may not reflect current laboratory platforms, geographic populations, or referral patterns. Missingness and historical coding practices can create shortcuts that a model learns without capturing clinically transferable relationships. Internal cross-validation cannot replace prospective or external validation. In addition, class definitions and hormone reference ranges can vary by laboratory, pregnancy status, age, medication use, and clinical context; these factors should be modeled explicitly before any real clinical deployment.

The web prototype must therefore be treated as decision support. It should not recommend medication, suppress clinician-ordered testing, or replace interpretation by an endocrinology professional. Users should be informed that predictions are probabilistic and that laboratory reference ranges and clinical symptoms remain authoritative. Privacy controls, data minimization, audit logging, and role-based access are required if patient records are used in a real system.

L. Decision-Threshold Sensitivity

The binary screening threshold determines how a continuous model probability is converted into a dysfunction flag. A default threshold of 0.50 is convenient but is not necessarily optimal for a medical screening objective. In the illustrative validation analysis, reducing the Random Forest threshold from 0.50 to 0.35 raises recall from 0.914 to 0.951, while precision decreases from 0.928 to 0.861. Raising the threshold to 0.65 has the opposite effect: precision increases to 0.951 but recall falls to 0.862. The choice therefore depends on whether the application prioritizes detection of possible dysfunction or reduction of unnecessary follow-up.

Table VIII. Illustrative Random Forest threshold sensitivity.

RF threshold	Precision	Recall	Specificity	Use
0.35	0.861	0.951	0.914	Recall-oriented
0.50	0.928	0.914	0.971	Balanced
0.65	0.951	0.862	0.984	Precision-oriented

For a screening-support tool, a recall-oriented threshold may be acceptable if every positive prediction is followed by standard laboratory interpretation and clinician review. Threshold selection must be performed only on training or validation data; choosing the threshold after seeing the final test set would introduce optimistic bias. A production system should also store the threshold version together with the model because a change in threshold changes the effective operating characteristics even when the trained classifier remains unchanged.

M. Sensitivity to Missing Data and Class Imbalance

Thyroid repositories contain incomplete measurements and highly unequal class counts, so preprocessing choices can alter apparent performance. A sensitivity analysis should compare at least three conditions: complete-case analysis, training-fold imputation, and a missingness-indicator strategy. In the illustrative experiment, median imputation fitted inside each training fold changes RF accuracy by less than one percentage point, but recall falls more noticeably when TSH or FTI is absent. This suggests that missingness in dominant laboratory variables should be surfaced to the user rather than silently replaced without qualification.

Class imbalance produces a second source of instability. Class weighting is used as the primary strategy because it changes the training objective without generating artificial patient records. Random oversampling and SMOTE may be studied as secondary analyses, but they must be restricted to the training folds. If synthetic samples are created before cross-validation, near-duplicate information can appear in both training and evaluation data and inflate performance. For rare thyroid categories, macro-averaged recall and per-class confidence intervals therefore provide a more reliable view than accuracy alone.

N. Comparison With Recent Thyroid-Machine-Learning Studies

The illustrative results are broadly consistent with the literature in which ensemble trees perform well on structured thyroid datasets. Chaganti et al. reported strong Random Forest behavior after selective feature analysis [3], and Jannat Peya et al. found high performance from Random Forest in a three-class UCI setting [8]. Gupta et al. further showed that optimized ensemble methods can achieve very high classification accuracy when class imbalance and hyperparameters are carefully handled [6]. Sutradhar et al. emphasized interpretable prediction and identified T3, TT4, TSH, and FTI among influential variables [5]. The current framework differs by placing stronger emphasis on leakage control, uncertainty intervals, calibration, paired hypothesis tests, and threshold analysis rather than maximizing a single headline accuracy.

Direct numerical comparison across papers should nevertheless be avoided when datasets, class definitions, preprocessing, and validation protocols differ. A model evaluated on a randomly split, heavily oversampled dataset cannot be assumed equivalent to a model assessed by nested cross-validation or external validation. The exact-title IEEE Access study of Salloom et al. similarly argues for a statistically grounded classical baseline [1]. This conference project therefore treats methodological transparency as a central result: the reported metric is accompanied by the split strategy, calibration procedure, uncertainty estimate, and the statistical test used for paired model comparison.

O. Practical Deployment Scenario and Model Governance

A practical prototype can be placed after laboratory data entry rather than before testing. When TSH and associated hormone measurements are available, the Flask service validates input ranges, applies the stored preprocessing pipeline, produces class probabilities, and records the model version and timestamp. High-confidence dysfunction predictions can be highlighted for review, while low-confidence or borderline cases can be marked “manual interpretation required.” This design uses machine learning as a consistency check and prioritization layer without changing established diagnostic workflows.

Model governance is essential because laboratory reference intervals, coding practices, and patient populations change over time. Each released model should be linked to a frozen training dataset, feature dictionary, preprocessing object, calibration object, decision threshold, and validation report. Monitoring can track missing-value rates, class proportions, probability distributions, and observed error rates. If calibration or recall degrades beyond a predefined tolerance, the system should enter a review state. Retraining must be evaluated against the currently deployed model on a fixed retrospective benchmark before release, and all historical predictions should remain traceable to the exact model version that generated them.

V. CONCLUSION

This paper presented a compact classical machine-learning framework for thyroid dysfunction prediction with emphasis on rigorous statistical evaluation. Logistic Regression, SVM, Random Forest, Decision Tree, KNN, and Naive Bayes were organized within a leakage-aware pipeline containing fold-specific preprocessing, nested cross-validation, bootstrap confidence intervals, calibration analysis, and paired significance testing. In the illustrative results, Random Forest produced the strongest overall balance of Accuracy, Precision, Recall, F1-score, and ROC-AUC, while SVM and Logistic Regression remained competitive baselines. TSH and free-thyroxine-related features were the most influential

variables, and the prototype web interface demonstrated how the model output can be presented as calibrated decision support.

Future work should replace every illustrative value with outputs generated from the final project dataset and code, archive the exact preprocessing and split definitions, and report class-wise confidence intervals. External validation on an independent hospital or recent cohort would be more informative than additional tuning on the same UCI-derived data. Additional extensions include cost-sensitive thresholds, fairness analysis, temporal validation, SHAP-based local explanations, and comparison with gradient boosting or carefully regularized deep models. Any deployment should preserve clinician oversight and should be evaluated as a screening-support system rather than as a stand-alone diagnostic device.

REFERENCES

- [1] S. A. Salloum, A. Almomani, K. M. Alomari, T. Khdour, and M. Alauthman, "Predicting Thyroid Dysfunction Using Classical Machine Learning With Rigorous Statistical Evaluation," *IEEE Access*, vol. 14, pp. 28852–28866, 2026, doi: 10.1109/ACCESS.2026.3666210.
- [2] M. Hu et al., "Development and preliminary validation of a machine learning system for thyroid dysfunction diagnosis based on routine laboratory tests," *Communications Medicine*, vol. 2, Art. no. 9, 2022, doi: 10.1038/s43856-022-00071-1.
- [3] R. Chaganti, F. Rustam, I. de la Torre Diez, J. L. Vidal Mazón, C. L. Rodríguez, and I. Ashraf, "Thyroid Disease Prediction Using Selective Features and Machine Learning Techniques," *Cancers*, vol. 14, no. 16, Art. no. 3914, 2022, doi: 10.3390/cancers14163914.
- [4] K. Guleria, S. Sharma, S. Kumar, and S. Tiwari, "Early prediction of hypothyroidism and multiclass classification using predictive machine learning and deep learning," *Measurement: Sensors*, vol. 24, Art. no. 100482, 2022, doi: 10.1016/j.measen.2022.100482.
- [5] A. Sutradhar et al., "Advancing thyroid care: An accurate trustworthy diagnostics system with interpretable AI and hybrid machine learning techniques," *Heliyon*, vol. 10, no. 17, e36556, 2024, doi: 10.1016/j.heliyon.2024.e36556.
- [6] P. Gupta, F. Rustam, K. Kanwal, W. Aljedaani, S. Alfarhood, M. Safran, and I. Ashraf, "Detecting Thyroid Disease Using Optimized Machine Learning Model Based on Differential Evolution," *Int. J. Comput. Intell. Syst.*, vol. 17, Art. no. 3, 2024, doi: 10.1007/s44196-023-00388-2.
- [7] T. L. Shiuh, W. K. Khai, Y. C. Xin, and C. Y. Wai, "Prediction of Thyroid Disease using Machine Learning Approaches and Featurewiz Selection," *J. Telecommun. Electron. Comput. Eng.*, vol. 15, no. 3, 2023, doi: 10.54554/jtec.2023.15.03.002.
- [8] Z. J. Peya, M. S. Islam, and M. K. N. Chumki, "Thyroid Disease Prediction based on Feature Selection and Machine Learning," in *Proc. 25th Int. Conf. Computer and Information Technology (ICCIT)*, 2022, pp. 495–500, doi: 10.1109/ICCIT57492.2022.10054746.
- [9] B. G. Keerthana, N. M. Kumar, T. Rao, and A. Kodipalli, "Smart Thyroid Diagnosis: A Machine Learning Based Interactive System," in *Proc. Int. Conf. Emerging Technologies in Computer Science for Interdisciplinary Applications (ICETCS)*, 2024, doi: 10.1109/ICETCS61022.2024.10543801.
- [10] G. Mollica, D. Francesconi, G. Costante, S. Moretti, et al., "Classification of Thyroid Diseases Using Machine Learning and Bayesian Graph Algorithms," *IFAC-PapersOnLine*, vol. 55, no. 40, pp. 67–72, 2022, doi: 10.1016/j.ifacol.2023.01.050.
- [11] J. Y. Noh et al., "An Evaluation of the Efficacy of Machine Learning in Predicting Thyrotoxicosis and Hypothyroidism: A Comparative Assessment of Biochemical Test Parameters Used in Different Health Checkups," *Internal Medicine*, vol. 63, no. 16, pp. 2259–2268, 2024, doi: 10.2169/internalmedicine.2825-23.
- [12] G. S. Collins et al., "TRIPOD+AI statement: updated guidance for reporting clinical prediction models that use regression or machine learning methods," *BMJ*, vol. 385, e078378, 2024, doi: 10.1136/bmj-2023-078378.
- [13] K. G. M. Moons et al., "PROBAST+AI: an updated quality, risk of bias, and applicability assessment tool for prediction models using regression or artificial intelligence methods," *BMJ*, vol. 388, e082505, 2025, doi: 10.1136/bmj-2024-082505.
- [14] UCI Machine Learning Repository, "Thyroid Disease," University of California, Irvine. [Online]. Available: <https://archive.ics.uci.edu/dataset/102/thyroid+disease>. Accessed: Sep. 2026.
- [15] L. Breiman, "Random forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001.
- [16] C. Cortes and V. Vapnik, "Support-vector networks," *Machine Learning*, vol. 20, no. 3, pp. 273–297, 1995.
- [17] D. W. Hosmer, S. Lemeshow, and R. X. Sturdivant, *Applied Logistic Regression*, 3rd ed. Hoboken, NJ, USA: Wiley, 2013.
- [18] T. Fawcett, "An introduction to ROC analysis," *Pattern Recognition Letters*, vol. 27, no. 8, pp. 861–874, 2006.
- [19] D. M. W. Powers, "Evaluation: From Precision, Recall and F-Measure to ROC, Informedness, Markedness and Correlation," *J. Mach. Learn. Technol.*, vol. 2, no. 1, pp. 37–63, 2011.
- [20] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: Synthetic Minority Over-sampling Technique," *J. Artif. Intell. Res.*, vol. 16, pp. 321–357, 2002.
- [21] A. Niculescu-Mizil and R. Caruana, "Predicting good probabilities with supervised learning," in *Proc. 22nd Int. Conf. Machine Learning*, 2005, pp. 625–632.
- [22] J. C. Platt, "Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods," in *Advances in Large Margin Classifiers*, Cambridge, MA, USA: MIT Press, 1999, pp. 61–74.
- [23] G. W. Brier, "Verification of forecasts expressed in terms of probability," *Monthly Weather Review*, vol. 78, no. 1, pp. 1–3, 1950.
- [24] E. R. DeLong, D. M. DeLong, and D. L. Clarke-Pearson, "Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach," *Biometrics*, vol. 44, no. 3, pp. 837–845, 1988.
- [25] Q. McNemar, "Note on the sampling error of the difference between correlated proportions or percentages," *Psychometrika*, vol. 12, pp. 153–157, 1947.
- [26] B. Efron and R. J. Tibshirani, *An Introduction to the Bootstrap*. New York, NY, USA: Chapman & Hall, 1993.
- [27] F. Pedregosa et al., "Scikit-learn: Machine Learning in Python," *J. Mach. Learn. Res.*, vol. 12, pp. 2825–2830, 2011.
- [28] S. M. Lundberg and S.-I. Lee, "A Unified Approach to Interpreting Model Predictions," in *Advances in Neural Information Processing Systems 30*, 2017, pp. 4765–4774.
- [29] M. T. Ribeiro, S. Singh, and C. Guestrin, "Why Should I Trust You?: Explaining the Predictions of Any Classifier," in *Proc. 22nd ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining*, 2016, pp. 1135–1144.
- [30] M. Sokolova and G. Lapalme, "A systematic analysis of performance measures for classification tasks," *Information Processing & Management*, vol. 45, no. 4, pp. 427–437, 2009.