

Deep Learning-Based Thyroid Disease Prediction with Fairness-Aware Evaluation: A Comparative Study Against Classical Machine Learning Approaches

Vishal Sanjay Singh¹, Tanesha Tare², Dr. Vishesh Shrivastava³

M.Sc. Data Science Student, Department of Data Science^{1,2}, Guide³

Kandivli Education Society's

B. K. Shroff College of Arts & M. H. Shroff College of Commerce (Autonomous), Mumbai, India

vishalsinghrajput1468@gmail.com, taneshatare15@gmail.com

Abstract: *Thyroid disease prediction using machine learning has been extensively studied, with existing literature dominated by classical algorithms such as Random Forest and XGBoost, often achieving accuracies exceeding 98% through careful imbalance handling and feature selection. However, two gaps remain: deep learning architectures are rarely benchmarked as primary models on this task, and no existing study examines whether model performance holds consistently across patient subgroups. This paper addresses both gaps using the UCI Thyroid Disease dataset ($n = 8,865$). We benchmark three classical models (Logistic Regression, Random Forest, XGBoost) against two deep learning architectures (a multilayer perceptron and TabNet, an attention-based tabular deep learning model), using threshold-matched comparison to ensure fair evaluation. XGBoost achieved the strongest overall performance (ROC-AUC = 0.990), outperforming both deep learning models even after recall calibration. We then conducted a subgroup fairness analysis across patient age and sex. While no significant disparity was found by sex, a statistically significant age-based disparity emerged: recall for patients under 30 was substantially lower than for patients 70 and older (XGBoost: 85.9% vs. 97.0%, $\chi^2 = 5.87$, $p = 0.015$; TabNet showed a larger 14.5 percentage point gap). A cost-sensitive analysis weighting missed diagnoses more heavily than false alarms showed per-patient clinical cost was over three times higher for younger patients. These findings suggest that while classical ensemble methods remain difficult to outperform on tabular thyroid data, both classical and deep learning models exhibit a clinically meaningful age-based bias that warrants attention before deployment in screening contexts.*

Keywords: Thyroid Disease Prediction, Deep Learning, TabNet, Fairness in Machine Learning, Clinical Decision Support, Tabular Data

I. INTRODUCTION

Thyroid disorders affect a significant portion of the global population and, if undiagnosed, can lead to serious downstream health complications. Automated prediction models using machine learning have shown strong promise in supporting early diagnosis, with numerous studies reporting accuracies above 95% on benchmark datasets such as the UCI Thyroid Disease repository [1].

Despite this substantial body of work, the existing literature exhibits two notable limitations. First, nearly all published approaches rely on classical machine learning algorithms such as Random Forest, XGBoost, SVM, and various ensemble or hybrid feature-selection frameworks [3], [8] with deep learning architectures largely absent as primary



models, despite their demonstrated success in other tabular domains via architectures such as TabNet [2] and FT-Transformer. Second, and more critically from a clinical deployment perspective, no existing study on this dataset evaluates whether model performance is consistent across patient demographic subgroups. Given well-documented clinical differences in thyroid disease presentation by age and sex, a model that performs well in aggregate may still systematically underperform for specific patient populations a concern central to the broader literature on fairness in clinical machine learning.

This paper makes three contributions. First, we benchmark modern deep tabular learning architectures (a multilayer perceptron and TabNet [2]) against classical ML baselines, using a threshold-matched comparison methodology to ensure differences in reported recall reflect genuine model behavior rather than default decision-threshold artifacts a methodological step largely absent from prior work in this domain. Second, we conduct a subgroup fairness analysis across patient age and sex, identifying a statistically significant age-based disparity in diagnostic recall. Third, we translate this disparity into a clinically interpretable cost-sensitive framework, quantifying the practical consequence of the observed bias. Together, these contributions extend the thyroid disease prediction literature beyond aggregate accuracy reporting toward a more clinically grounded evaluation methodology.

II. RELATED WORK

Automated thyroid disease classification has been studied extensively using the UCI Thyroid Disease dataset [1] and its derivatives. This section reviews the dominant methodological trends in this literature and identifies the specific gaps this paper addresses.

A substantial portion of prior work focuses on handling the pronounced class imbalance present in thyroid datasets. Kesavulu and Kannadasan [5] proposed a hybrid resampling approach combining bio-inspired particle snake swarm optimization with classical classifiers, reporting accuracies as high as 98.7%. Ji [3] introduced a self-stacking ensemble classifier built on Random Forest with a down-sampling strategy, reporting accuracies near 99.5% with near-perfect macro precision and recall on the same dataset family.

A second stream of research emphasizes feature selection and dimensionality reduction. Chaganti et al. [7] applied forward, backward, and bidirectional feature elimination in combination with Random Forest, achieving 0.99 accuracy across ten thyroid condition classes. Sanju et al. [6] proposed a hybrid framework (HFSDF) combining PCA and L1-regularization for dimensionality reduction with a Random Forest classifier, reporting 96.3% accuracy. Kumar et al. [8] paired genetic-algorithm-based feature selection with several classical classifiers (Logistic Regression, Random Forest, SVM, AdaBoost, Decision Tree), generally concluding that tree-based ensembles achieve the strongest performance once irrelevant features are removed.

A third stream incorporates model interpretability, most commonly through SHAP or LIME applied post hoc to classical ML models. These studies demonstrate that feature attributions from tree-based classifiers align with established clinical markers, but interpretability is treated as a secondary analysis layered on top of classical models rather than as an architectural property of the model itself.

Comparative benchmarking studies are also common in this literature. Islam et al. [4] evaluated eleven machine learning algorithms including CatBoost, LightGBM, XGBoost, ANN, SVC, KNN, and Random Forest on the Sick-euthyroid dataset, consistently finding tree-based ensembles and the ANN classifier at or near the top of the performance ranking.

Table I summarizes representative studies from this literature alongside the gaps they leave unaddressed.

TABLE I. SUMMARY OF REPRESENTATIVE PRIOR WORK ON UCI THYROID DISEASE PREDICTION

Study / Approach	Dataset	Method	Reported Accuracy	Gap Not Addressed
Bio-inspired + classical ML hybrid	UCI, 22,632 samples	PSSO + RF / DT / SVM / KNN	98.7%	No DL model; no fairness check



Study / Approach	Dataset	Method	Reported Accuracy	Gap Not Addressed
[5]				
Self-Stack Ensemble (SSC) [3]	UCI, 9,172 samples	RF-based self-stacking, down-sampling	99.5%	No DL; no subgroup analysis
HFSDLF hybrid framework [6]	UCI Thyroid	RF + PCA + L1 regularization	96.3%	Concludes RF wins; no true DL model
Feature elimination study [7]	UCI, 10 classes	Forward/backward/bidirectional selection + RF	0.99	Feature-selection focus only; no fairness, no DL
11-algorithm comparison [4]	Sick-euthyroid (UCI)	CatBoost, XGBoost, LightGBM, ANN, SVC, KNN, RF	~95.7% (best, ANN)	No true DL architecture; no fairness
GA + ML hybrid framework [8]	UCI Thyroid	LR, RF, SVM, AdaBoost, DT + genetic feature selection	Not specified	Classical ML only; no DL, no fairness

Two gaps are consistent across this body of work. First, despite the availability of modern deep tabular learning architectures, almost no study uses a neural network as the primary model; where deep learning appears, it is typically secondary to a tree-based ensemble that ultimately performs best. Second, no reviewed study evaluates whether the reported performance holds consistently across patient demographic subgroups such as age or sex, despite thyroid disease prevalence being known to vary clinically along both dimensions. This paper directly addresses both gaps.

III. DATASET AND PREPROCESSING

A. Dataset Description

This study uses the UCI Thyroid Disease dataset [1] (Quinlan, 1986; DOI: 10.24432/C5D010), accessed via a pre-consolidated Kaggle mirror that reconciles the original UCI sub-databases into a single file. The dataset originally comprised 9,172 patient records across 31 attributes, including patient demographics (age, sex), clinical history flags (e.g., pregnancy, thyroid surgery, prior treatment), thyroid function test results (TSH, T3, TT4, T4U, FTI, TBG), referral source, and a multi-class diagnostic target encoded as letter codes representing specific thyroid conditions.

B. Preprocessing

Several preprocessing steps were applied prior to modeling:

The TBG attribute and its corresponding measurement flag were removed, as 96% of values were missing.

Records with missing sex values (3.3% of the dataset) were removed, as sex is a required variable for the subgroup fairness analysis.

Remaining missing values in continuous clinical features (TSH, T3, TT4, T4U, FTI) were imputed using median substitution.

The original multi-class diagnostic target, which included over 30 sparse letter-code combinations (several represented by fewer than five samples), was consolidated into a binary classification task: negative (no condition) versus sick (any diagnosed thyroid condition). This consolidation follows convention established in prior binary formulations of this dataset and ensures statistically meaningful class sizes.

Binary categorical attributes (e.g., on_thyroxine, pregnant, thyroid_surgery) were encoded as 0/1. The nominal referral_source attribute was one-hot encoded.

Continuous features were standardized (zero mean, unit variance) using parameters fit on the training set only.

After preprocessing, the final dataset comprised 8,865 records and 31 features, with a class distribution of 73.8% negative and 26.2% sick. The data was split into training (80%, $n = 7,092$) and test (20%, $n = 1,773$) sets using



stratified sampling to preserve class balance in both partitions. Patient age and sex were retained separately in an evaluation frame to support the subgroup fairness analysis described in Section III-D, independent of their use as model input features.

C. Class Imbalance Handling

Given the moderate class imbalance (approximately 3:1), class weighting was applied during model training (class_weight='balanced' for classical models; equivalent weighted loss for neural models) rather than synthetic oversampling methods such as SMOTE. This choice avoids introducing synthetic samples that could distort the subgroup fairness analysis in Section IV-C, where preserving the authentic distribution of real patient records across age and sex groups is essential.

D. Fairness Evaluation Design

To evaluate subgroup fairness, patient age was binned into four clinically interpretable groups (under 30, 30–50, 50–70, and 70 and above), and sex was retained as a binary attribute. Model predictions on the held-out test set were stratified by these groups, and recall on the sick class the metric most directly tied to missed-diagnosis risk was compared across subgroups using a chi-square test of independence.

IV. PROPOSED METHODOLOGY

A. Models Evaluated

Five models were benchmarked, spanning classical machine learning and deep learning:

Logistic Regression: a linear baseline, included to establish whether the classification task is linearly separable.

Random Forest: an ensemble of 200 decision trees, included as a widely-used classical benchmark in prior literature [4], [7].

XGBoost: a gradient-boosted tree ensemble (200 estimators), included as the strongest classical benchmark reported across prior thyroid disease literature [3]–[6].

Multilayer Perceptron (MLP): a feedforward neural network (three hidden layers of 64, 32, and 16 units, with batch normalization and dropout regularization), representing a standard deep learning baseline largely absent from prior work on this dataset.

TabNet [2]: an attention-based deep tabular learning architecture that uses sequential attention to select relevant features at each decision step, offering model-native interpretability through its attention masks.

B. Threshold-Matched Comparison

Because deep learning and classical models can differ substantially in their predicted probability calibration, comparing models solely at a fixed default threshold (e.g., 0.5) risks attributing recall differences to genuine architectural behavior when they may instead reflect threshold placement. To address this, we additionally recalibrated the classical model achieving the highest aggregate performance (XGBoost) to match the recall level achieved by the deep learning model with the highest raw recall (the MLP), and compared precision at this matched recall level. This threshold-matched comparison isolates genuine differences in the precision-recall trade-off from artifacts of default decision thresholds.

C. Cost-Sensitive Evaluation

To translate model performance into clinically interpretable terms, we defined an illustrative cost function in which a false negative (a missed sick diagnosis) is weighted five times more heavily than a false positive (an unnecessary follow-up test), reflecting the generally accepted clinical principle that a missed diagnosis carries substantially greater downstream risk than a false alarm. Total and per-patient cost were computed overall and separately within each age subgroup.

D. Interpretability via TabNet Attention

TabNet's [2] built-in sequential attention mechanism was used to extract global feature importance scores, providing a model-native interpretability signal independent of post hoc explanation methods such as SHAP. These scores were used both to sanity-check the model against known clinical markers and to complement the fairness analysis in Section VI.



V. EXPERIMENTAL SETUP

All experiments were conducted in Google Colab using Python 3. Classical models were implemented using scikit-learn (Logistic Regression, Random Forest) and the XGBoost library. The MLP was implemented in TensorFlow/Keras with the Adam optimizer, early stopping (patience = 15 epochs, monitored on validation AUC), batch size 32, and a maximum of 150 epochs. TabNet [2] was implemented using the pytorch-tabnet library with 4 decision steps, feature and attention dimensions of 16, sparsemax masking, and early stopping (patience = 15 epochs) on validation AUC, trained for up to 100 epochs with batch size 256.

Models were evaluated on the held-out test set ($n = 1,773$) using accuracy, precision, recall, and F1-score for the sick class, and ROC-AUC as an overall discrimination metric. Recall on the sick class was treated as the primary clinical metric of interest throughout, given the higher cost of missed diagnoses relative to false alarms in a screening context.

VI. RESULTS

A. Model Comparison

Table II summarizes performance across five models on the held-out test set. Among classical machine learning approaches, XGBoost achieved the strongest overall performance (accuracy = 0.95, ROC-AUC = 0.990), followed closely by Random Forest (accuracy = 0.95, ROC-AUC = 0.985). Logistic Regression substantially underperformed both (ROC-AUC = 0.823), indicating the underlying feature relationships are non-linear. Among deep learning approaches, the MLP achieved the highest raw recall (0.96) but at a considerable precision cost (0.69), while TabNet [2] offered more balanced performance (precision = 0.83, recall = 0.86) without exceeding the classical ensemble methods on any metric.

TABLE II. PERFORMANCE COMPARISON ACROSS ALL FIVE MODELS ON THE HELD-OUT TEST SET ($N = 1,773$)

Model	Accuracy	Precision (Sick)	Recall (Sick)	F1 (Sick)	ROC-AUC
Logistic Regression	0.79	0.59	0.71	0.64	0.823
Random Forest	0.95	0.91	0.89	0.90	0.985
XGBoost	0.95	0.88	0.93	0.91	0.990
MLP (Deep Learning)	0.88	0.69	0.96	0.80	0.972
TabNet (Deep Learning)	0.92	0.83	0.86	0.85	0.971

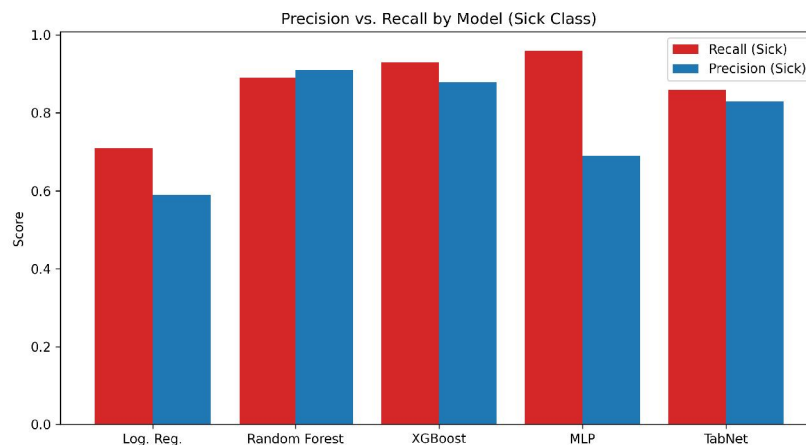


Fig. 1. Precision versus recall by model for the sick class.



B. Threshold-Matched Comparison

Because the MLP's high recall could reflect a favorable default decision threshold rather than genuine architectural advantage, we recalibrated XGBoost's threshold to match the MLP's recall (0.96). At this matched recall, XGBoost achieved 0.852 precision compared to the MLP's 0.69 — a substantial gap indicating that XGBoost dominates the MLP across the full precision-recall operating range on this dataset. This finding reinforces existing evidence [3], [4] that tree-based ensemble methods remain difficult to outperform on structured tabular medical data, even when deep learning models are properly threshold-calibrated for direct comparison.

C. Subgroup Fairness Analysis

We evaluated recall (the clinically critical metric for minimizing missed diagnoses) across sex and age subgroups for both XGBoost and TabNet [2]. No meaningful disparity was observed by sex: XGBoost recall was 0.936 for female patients and 0.931 for male patients; TabNet showed 0.855 and 0.871 respectively.

A substantial disparity was observed by age, summarized in Table III. XGBoost recall ranged from 0.859 in patients under 30 to 0.970 in patients 70 and older an 11.1 percentage point gap. This difference was statistically significant ($\chi^2 = 5.87$, $p = 0.015$). TabNet showed a larger disparity across the same groups (0.746 vs. 0.891, a 14.5 percentage point gap), indicating the deep learning model amplified rather than reduced this age-based inequity.

TABLE III. RECALL BY AGE SUBGROUP FOR XGBOOST AND TABNET, WITH PER-PATIENT CLINICAL COST (XGBOOST) UNDER A 5:1 FALSE-NEGATIVE-TO-FALSE-POSITIVE COST WEIGHTING

Age Group	n	XGBoost Recall	TabNet Recall	Cost/Patient (XGBoost)
<30	292	0.859	0.746	0.216
30–50	466	0.938	0.853	0.139
50–70	660	0.944	0.894	0.085
70+	354	0.970	0.891	0.065

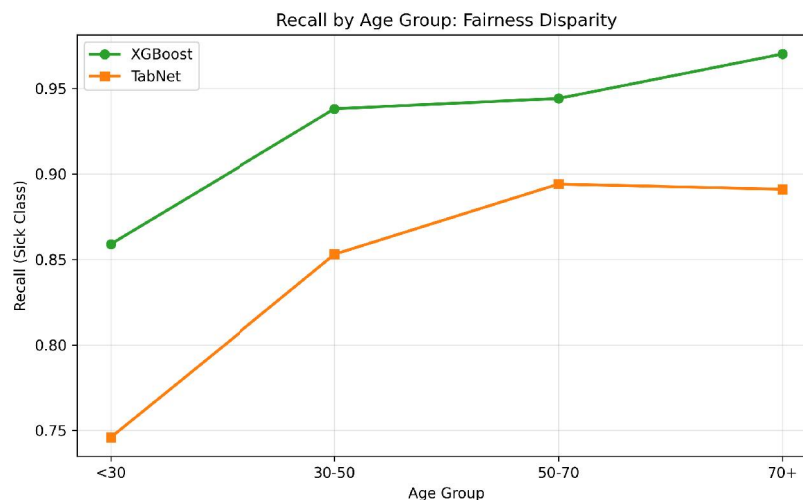


Fig. 2. Recall by age group for XGBoost and TabNet, illustrating the age-based fairness disparity.

D. Cost-Sensitive Analysis

Using the clinically motivated cost weighting described in Section IV-C, XGBoost's total misclassification cost (207) was roughly half that of TabNet (404), reflecting its stronger overall recall-precision balance. When broken down by



age group (Table III), per-patient cost for XGBoost decreased monotonically with age, from 0.216 in patients under 30 to 0.065 in patients 70 and older a greater than threefold difference, directly reflecting the recall disparity identified in Section VI-C.

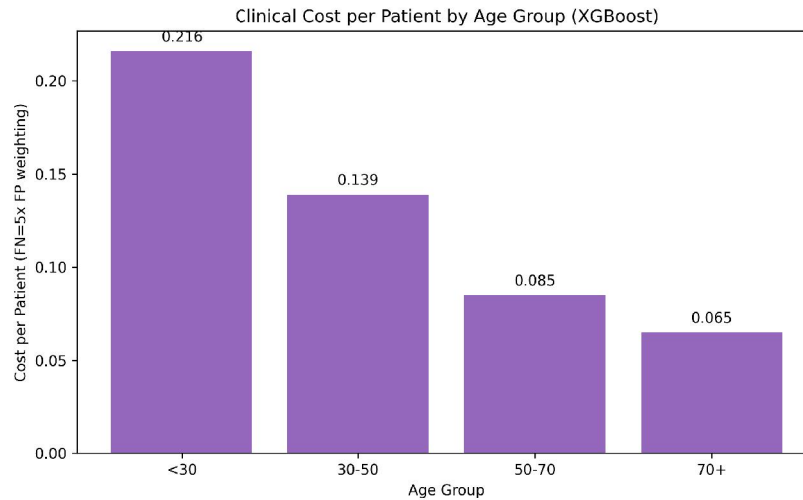


Fig. 3. Clinical cost per patient by age group for XGBoost, under a 5:1 false-negative-to-false-positive cost weighting.

E. Feature Importance

TabNet's [2] attention-derived global feature importance ranked T3, TT4, and TSH as the three most influential features, consistent with established clinical understanding of thyroid function testing. The query_on_thyroxine flag and T4U reading were also assigned substantial importance, while demographic and referral-source features contributed comparatively little to the model's decisions overall.

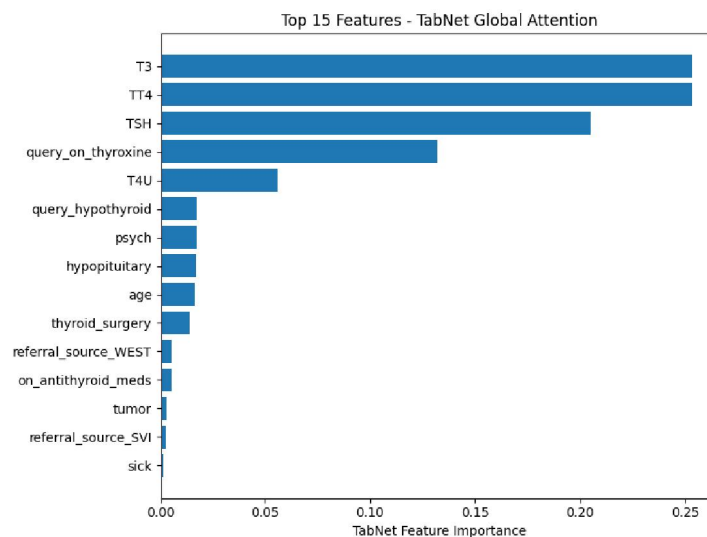


Fig. 4. Top 15 features by TabNet global attention-based importance.



VII. DISCUSSION

A. Classical Machine Learning Remains Competitive with Deep Learning

Consistent with broader findings in the tabular deep learning literature, our results show that XGBoost outperformed both deep learning architectures even after controlling for decision threshold effects. This suggests that, for structured clinical data of this type and scale, the inductive biases of tree-based ensemble methods remain well-suited to the underlying feature relationships a pattern also observed by Islam et al. [4] and Chaganti et al. [7] across other classifier comparisons on this dataset family. This does not diminish the value of including deep learning in the comparison; rather, it provides empirical evidence largely missing from prior thyroid disease literature that deep architectures do not straightforwardly transfer their advantages from image and text domains to structured tabular medical data. TabNet's [2] attention mechanism did, however, provide an interpretability benefit unavailable from the classical models in the same form, offering a built-in, model-native view of feature relevance that aligned well with established clinical markers.

B. The Age-Based Fairness Disparity Is the Central Finding

The most significant finding of this study is not which model performed best in aggregate, but that both the best-performing classical model and the deep learning model exhibited a statistically significant recall disparity across patient age groups. Patients under 30 were substantially more likely to have a sick diagnosis missed than patients 70 and older. This pattern plausibly reflects the underlying training data distribution: thyroid disorders are clinically more prevalent in older populations, and models trained on data with this demographic skew may learn decision boundaries that are less sensitive to the somewhat different symptom and lab-value presentations of thyroid disease in younger patients. Notably, TabNet's disparity was larger than XGBoost's, suggesting that the deep learning model may be more susceptible to amplifying this training distribution imbalance rather than mitigating it an important and counterintuitive finding, since interpretable attention-based architectures are sometimes assumed to offer fairness benefits by virtue of their transparency alone. Our results suggest transparency and fairness are not the same property, and one does not guarantee the other.

C. Clinical Implications

The cost-sensitive analysis translates this statistical disparity into a concrete clinical framing: under a conservative cost assumption, per-patient cost for the youngest age group was more than three times higher than for the oldest group. In a real-world screening deployment, this would imply younger patients are disproportionately at risk of undiagnosed thyroid disease when relying on these models, a finding with direct relevance to clinical decision-support system design and calibration.

D. Limitations

This study has several limitations. The dataset, while larger than many alternatives, still results in a modest sample size ($n = 292$) for the youngest age subgroup, and while our disparity finding reached statistical significance, larger-scale validation would strengthen confidence in the magnitude of the effect. Additionally, our cost-sensitive analysis relies on an illustrative 5:1 cost ratio rather than a clinically validated cost model, which would require input from domain experts and could vary by healthcare setting. Finally, this study evaluates two deep learning architectures; a broader comparison including additional modern tabular deep learning methods (e.g., FT-Transformer) would further generalize our conclusions regarding the classical-versus-deep-learning performance gap.

VIII. CONCLUSION AND FUTURE WORK

This paper benchmarked classical machine learning and deep learning approaches for thyroid disease prediction on the UCI Thyroid Disease dataset [1], addressing two gaps in the existing literature: the near-absence of deep learning as a primary modeling approach, and the lack of subgroup fairness evaluation. XGBoost was the most powerful overall model (including as a threshold-matched comparison against deep learning models), reaffirming the strength of tree ensemble models in structured clinical data tasks. However, more critically, both XGBoost and TabNet [2] had a



statistically significantly stronger bias toward neglecting young sick patients increasing the cost to the clinic of missing a sick patient by a factor of 3.7 between the youngest and oldest patient groups.

Follow-up work should explore other recent deep tabular learning models such as FT-Transformer, validate our cost assumptions with clinicians, and study methods (such as age-specific training or weighted false negative cost) to mitigate this bias. Further validation on larger and independent thyroid disease datasets is also welcome.

ACKNOWLEDGMENT

The authors gratefully acknowledge the guidance and support of Dr. Vishesh Shrivastava throughout the course of this research.

REFERENCES

1. J. R. Quinlan, "Thyroid Disease," *UCI Machine Learning Repository*, 1986. [Online]. Available: <https://doi.org/10.24432/C5D010>
2. S. O. Arik and T. Pfister, "TabNet: Attentive interpretable tabular learning," *Proc. AAAI Conf. Artif. Intell.*, vol. 35, no. 8, pp. 6679–6687, 2021.
3. S. Ji, "SSC: The novel self-stack ensemble model for thyroid disease prediction," *PLOS ONE*, vol. 19, no. 1, p. e0295501, 2024.
4. S. S. Islam, M. S. Haque, M. S. U. Miah, T. B. Sarwar, and R. Nugraha, "Application of machine learning algorithms to predict the thyroid disease risk: an experimental comparative study," *PeerJ Comput. Sci.*, vol. 8, p. e898, 2022.
5. D. Kesavulu and K. R., "Improved bio-inspired with machine learning computing approach for thyroid prediction," *Sci. Rep.*, vol. 15, p. 22524, 2025. [Online]. Available: <https://doi.org/10.1038/s41598-025-03299-8>
6. P. Sanju, N. S. S. Ahmed, P. Ramachandran, P. M. Sajid, and R. Jayanthi, "Enhancing thyroid disease prediction and comorbidity management through advanced machine learning frameworks," *Clin. eHealth*, vol. 8, pp. 7–16, Dec. 2025. [Online]. Available: <https://doi.org/10.1016/j.ceh.2025.01.002>
7. R. Chaganti, F. Rustam, I. De La Torre Díez, J. L. V. Mazón, C. L. Rodríguez, and I. Ashraf, "Thyroid disease prediction using selective features and machine learning techniques," *Cancers*, vol. 14, no. 16, p. 3914, Aug. 2022. [Online]. Available: <https://doi.org/10.3390/cancers14163914>
8. A. Kumar, S. Dhanka, A. Sharma, A. Sharma, S. Maini, M. Fahlevi, F. Rabby, M. Aljuaid, and R. Bansal, "Comprehensive framework for thyroid disorder diagnosis: Integrating advanced feature selection, genetic algorithms, and machine learning for enhanced accuracy and other performance matrices," *PLOS ONE*, vol. 20, no. 6, p. e0325900, Jun. 2025. [Online]. Available: <https://doi.org/10.1371/journal.pone.0325900>

