ISSN : 2455-135X

An Intelligent Computational Framework for Predicting Thyroid Malignancy Recurrence via Bayesian Hyperparameter Tuning and Hybrid SMOTE-Tomek Resampling

Mohit Kharbanda[1], Arshi Husain[2], Virendra P. Vishwakarma[3]

[1]*USIC&T, GGSIPU, DELHI, INDIA.* [2]*USIC&T, GGSIPU, DELHI, INDIA.* [3]*USIC&T, GGSIPU, DELHI, INDIA.*

[1]mohit93kharbanda@gmail.com

Abstract – *Thyroid carcinoma stands as a widespread endocrine malignancy, making the accurate forecast of disease return essential for improving post-treatment strategy and patient prognosis. Nevertheless, severe skewness in clinical class distributions and complex feature interactions make predicting recurrence particularly demanding. To address these hurdles, this investigation outlines a robust machine learning architecture incorporating structured preprocessing alongside Bayesian parameter search. The proposed system features duplicate record filtering, categorical label mapping, stratified 80:20 partitioning, SMOTE-Tomek dynamic balancing, and standardization engineered to eliminate data leakage across evaluation folds. Four distinct supervised models—K-Nearest Neighbors (KNN), Support Vector Machines (SVM), Random Forests (RF), and Extreme Gradient Boosting (XGBoost)—were systematically optimized using five-fold stratified cross-validation. Comprehensive benchmarking via Accuracy, Precision, Sensitivity, F1-Score, Receiver Operating Characteristic curves, and confusion matrix analysis revealed that XGBoost attained superior predictive capacity. Specifically, XGBoost achieved an overall accuracy of* ***95.89%****, precision of* ***91.30%****, sensitivity of* ***95.45%****, F1-score of* ***93.33%****, and ROC-AUC of* ***97.95%****, markedly outperforming the comparative algorithms. These outcomes confirm the efficacy of the proposed pipeline in mitigating severe class skew and improving forecasting reliability, establishing its value for clinical decision support and patient follow-up planning.*

Keywords – *Thyroid cancer recurrence, Machine learning, Bayesian optimization, Synthetic resampling, XGBoost, Ensembles, Decision support systems*

1. Introduction

The thyroid gland functions as an essential organ of the endocrine system of the human body, controlling metabolic rate via T3 and T4 hormone release responsible for regulating oxygen consumption at the cellular level along with carbohydrate and fat metabolism [1],[2]. The morphology and function of the thyroid gland largely differ in different age groups, genders, and physiological states including pregnancy and lactation. The recent literature on this topic has greatly increased the approaches to diagnosing thyroid diseases through synthetic oversampling techniques such as SMOTE-NC together with gradient boosting [5], generative adversarial networks combined with Kolmogorov-Arnold structures [6], multi-level stacking ensemble methods [7], cluster-based sampling approach [8], genetic algorithm feature selection [9], and bio-inspired deep learning algorithms including ant lion-optimized LSTM and quantum SVM [10],[11]. Even though these current models consistently prove high levels of accuracy (with values over 95% [7–10] and sometimes even higher than 98%), the critical evaluation of those models reveals several serious flaws in their methodology. In many cases, current literature is dominated by single train-test holdout splits without using nested or cross-validation techniques, which also often lacks a probability calibration and hypothesis testing in the process of comparing models [5]. This problem is critical, considering the problem of severe class imbalance existing in clinical thyroid databases, where healthy samples (euthyroid) significantly outnumber diseased samples [8], thus leading to artificial accuracy values if the problem is not solved properly. Moreover, another problem of these models lies in the need for interpretation: while there is some progress in adding the SHAP or LIME explainability layers to modern models, still, many research papers use black-box models, which have no clinical proof of their validity [10]. To circumvent these limitations, this research proposes a complete machine learning pipeline specifically designed for thyroid disease diagnosis, which goes beyond conventional measures of accuracy. The new approach ensures stringent validation criteria, considers probability calibration of models, performs statistical tests on classifiers, and deals with class imbalance without affecting the interpretability of the model. The physiological mechanism behind the thyroid function involves various parameters such as age, gender, and changing biological condition [12]. Malfunctioning of this process normally results in hyperthyroidism and hypothyroidism, resulting in varied systemic impacts on health. Though it is essential to have timely diagnostics, it is a challenging task due to the complexity involved in

ISSN : 2455-135X

interpreting laboratory parameters [14]. Incorporating supervised machine learning techniques can be considered an effective way to deal with the above-mentioned diagnostic issues. Using information from patients' medical history in combination with laboratory data, algorithmic tools can facilitate the process and help in diagnostics. Earlier computer systems based on artificial immune recognition algorithms [17] gradually evolved into more sophisticated systems such as ensemble systems [18], neural networks [19], and hybrid intelligence [21]. Expanding on these approaches, this research offers a solid contribution to the development of machine learning-based thyroid diagnostics.

2. Literature Review

The current body of knowledge in this field mainly focuses on datasets generated from institutional cohorts or the UCI database having varying sample size within the range of 500-22,632 with 31 clinical variables. Researchers have applied various techniques including SMOTE-NC with gradient boosting [5], combination of generative adversarial network and Kolmogorov-Arnold network [6], stacked classifier ensembles [7], cluster-guided sampling using SMOTE-ENN, and SMOTE-KNN [8] along with evolutionary algorithms for feature selection using random forest and adaboost estimator [9]. There is extensive research done in deep learning model using ant lion optimized LSTM network and quantum support vector machine [10]. Complementary studies by [11] provide additional evidence for the effectiveness of machine learning algorithms in detecting thyroid disorders, and studies performed by [7] prove the effectiveness of stacked multi-models. Although accuracy measurements exceeding 95% are frequently reported, almost all papers rely only on the simple train-test split, failing to offer the calibration curves or any kind of statistical comparison at all. Class imbalance becomes an important issue, since the majority of euthyroid patients can dominate rare hypothyroid or hyperthyroid ones, making the performance statistics inaccurate. Moreover, there is a lack of interpretability, since although some of the current literature provides SHAP or LIME interpretability layers, most of the models are unexplainable black-box systems without any medical supervision. Common patterns of methodological approaches in existing studies point to prevalent structural deficiencies. Studies by [5] and [7] combined SMOTE-based resampling techniques with ensembles of classifiers applied to UCI thyroid data variations, yielding accuracy results of 95–97%; nonetheless, both of the works made use of non-repeated 80/20 data splits and did not conduct any significance testing or calibration. The approach in [6] and [8] was focused on tackling the issue of class imbalance using GAN networks and clustering-based oversampling, thus reaching high validation accuracy of isolated holdout datasets. The same is true for studies by [9] and [10], which used bio-inspired optimization methods, such as swarm intelligence, genetic algorithms, quantum SVM, and deep learning approaches, to optimize feature subsets and model parameters, thus receiving maximum accuracy results above 97%.

3. Materials and Methods

3.1. Dataset Description

Our proposed computational framework was benchmarked using the Thyroid Disease Recurrence Dataset. which unified clinical and pathological metrics from patients evaluated for differentiated thyroid cancer. The original dataset comprises 383 patient records with 17 demographic, pathological, therapeutic, and outcome features. Following systematic cleaning and missing value filtering with a final cohort of 364 complete patient records was retained for experiment.

Predictor features like Age, Biological Sex, Smoking Status, Smoking History, Prior Radiotherapy, Thyroid Function Metrics, Physical Examination Findings, Adenopathy, Pathology Class, Focality, Risk Stratification, T Stage, N Stage, M Stage, Overall Cancer Stage, and Treatment Response. The target binary outcome variable is *Recurred* (Yes/No), pointing whether a patient experienced cancer recurrence after treatment. Supervised machine learning algorithms for binary classification were deployed to model this outcome.

3.2. Data Preprocessing Pipeline

To guarantee optimal data integrity and preserve methodological biasness, raw features were processed by a sequential data pipeline:

1. Duplicated patient entries were identified and eliminate data points to reduce data redundancy.
2. Categorical features were changed into discrete numerical representations by using *Label Encoding*.
3. Dataset division was performed using an 80% training and 20% testing split by using stratified sampling to maintain uniform spam of data points for each class proportions.
4. To address severe class skewness, dynamic *SMOTE-Tomek* resampling method [31] was applied only to training subsets within individual cross validation iterations.
5. Standard feature scaling was incorporated via *StandardScaler* for distance sensitive baseline models. Like KNN and SVM.

All data transformation procedures were nested within an *imblearn Pipeline* structure [33]. strictly separating synthetic oversampling with feature normalization within training folds to prevent any target or data leakage into validation data subsets.

3.3. Bayesian Hyperparameter Optimization

To develop optimal architectural parameters for each model, the Bayesian Optimization was used for utilizing the

ISSN : 2455-135X

BayesSearchCV module from Scikit-Optimize [37]. Parameter exploration was determined by a 5-fold Stratified Cross-Validation scheme to make sure consistent class distributions across all validation folds.

The hyperparameter search space structured for each evaluated classifier is detailed in Table 1. Parameter selection was guided by maximizing the Area Under the Curve (AUC).

Table 1. Classifier Hyperparameters Tuned via Bayesian Optimization

Classifier	Optimized Hyperparameter Space
KNN	Neighbor Count (k), Distance Weighting Function
SVM	Regularization Parameter (C), Kernel Coefficient (γ)
Random Forest	Estimator Count, Maximum Tree Depth, Split Threshold
XGBoost	Estimator Count, Learning Rate (η), Max Tree Depth

3.4. Evaluated Machine Learning Models

Four supervised classification algorithms were implemented to evaluate predictive performance for thyroid cancer recurrence.

3.4.1. K-Nearest Neighbors (KNN). K-Nearest Neighbors (KNN) is an instance-based non-parametric classifier that assigns target labels via majority voting across the k nearest vector points in feature space [34]. Euclidean distance metrics were utilized to determine instance proximity.

3.4.2. Support Vector Machine (SVM). Support Vector Machines (SVM) construct optimal decision hyperplanes by maximizing margin distances between target classes [35]. In this experiment, a Radial Basis Function (RBF) kernel was leveraged to model complex nonlinear attribute interactions.

3.4.3. Random Forest (RF). Random Forest (RF) is an ensemble decision tree framework that constructs multiple decision trees via bootstrap aggregation (bagging) and randomized feature subsets, obtaining final class assignments via majority voting to minimize variance and overfitting [4].

3.4.4. Extreme Gradient Boosting (XGBoost). XGBoost is an optimized gradient boosting architecture that constructs sequential decision trees aimed at minimizing a regularized loss function [36]. Combining second-order gradient approximations, tree pruning with parallel processing. XGBoost be adept at modeling intricate non-linear relationships in complex clinical data.

3.5. Experimental Workflow

The detailed working pipeline of the proposed predictive framework is shown in Fig. 1.

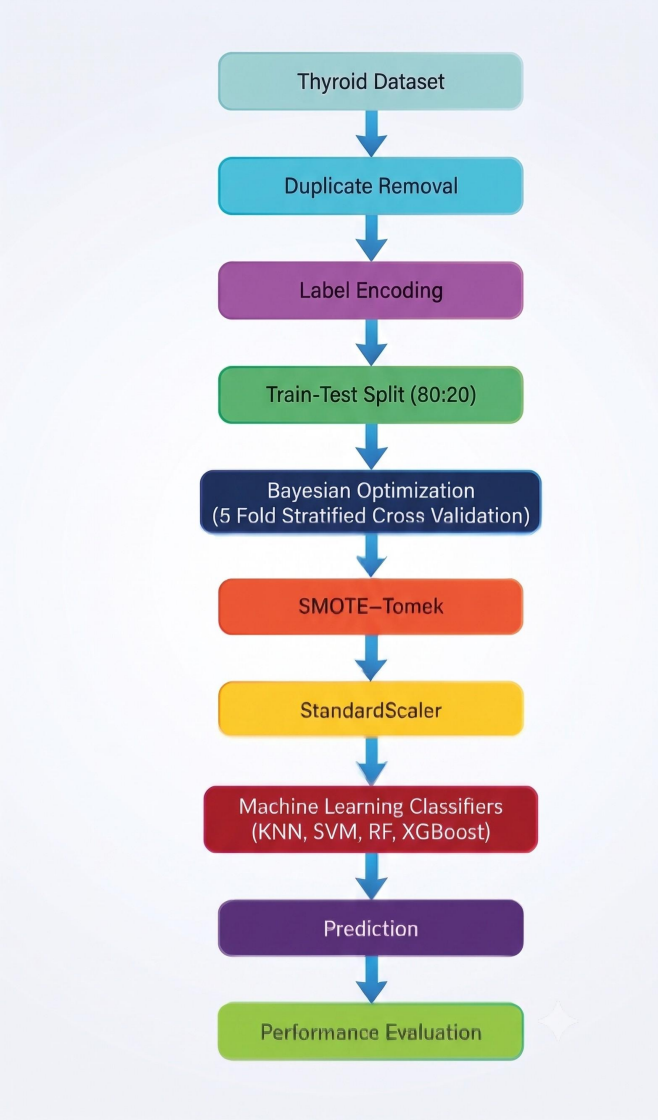

Figure 1. Systematic workflow layout of the proposed thyroid recurrence prediction framework.

3.6. Performance Metrics

Model evaluation was carried out across standard quantitative classification metrics : Accuracy, Precision, Recall (Sensitivity), F1-Score, and ROC-AUC [38].

3.6.1. Accuracy. Accuracy quantifies the overall share of rightly predicted outcomes across all data ponts:

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

3.6.2. Precision. Precision evaluates the share of true positive predictions among all positive predictions:

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

ISSN : 2455-135X

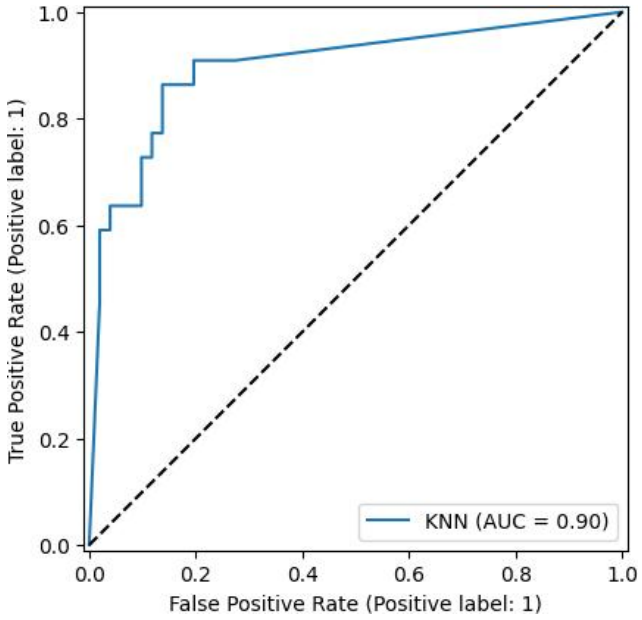

Figure 2. ROC-AUC Performance Curve for the KNN Classifier.

3.6.3. Recall. Recall (Sensitivity) evaluates the model's ability to correctly capture positive disease instances:

$$\text{Recall} = \frac{TP}{TP + FN} \tag{3}$$

3.6.4. F1-Score. The F1 score shows the harmonic mean by balancing Precision and Recall:

$$F_1 = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{4}$$

3.7. Experimental Results and Discussion

The Comparative performance metrics measured on the all four optimized classifiers are presented in Table 2.

Table 2. Performance Comparison Across Optimized Classification Models

Metric	KNN	SVM	RF	XGBoost
Accuracy (%)	82.19	91.78	91.78	**95.89**
Precision (%)	73.68	83.33	80.77	**91.30**
Recall (%)	63.64	90.91	95.45	**95.45**
F1-score (%)	68.29	86.96	87.50	**93.33**
ROC-AUC (%)	89.75	95.54	97.06	**97.95**

Empirical evaluations confirm that the XGBoost classifier achieved top performance across all benchmarked criteria. Specifically, XGBoost attained an accuracy of **95.89%**, precision of **91.30%**, sensitivity of **95.45%**, F1-score of **93.33%**, and ROC-AUC of **97.95%**. The superior accuracy of XGBoost highlights its capacity to learn nonlinear feature correlations while preventing model overfitting.

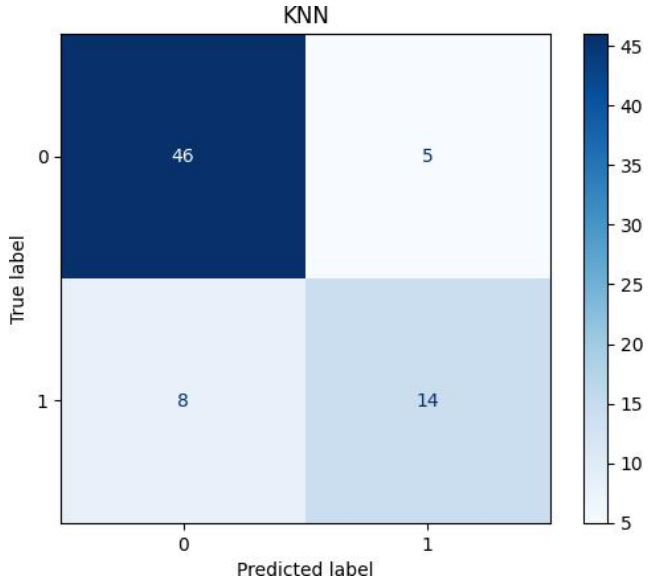

Figure 3. Confusion Matrix Displaying KNN Predictions.

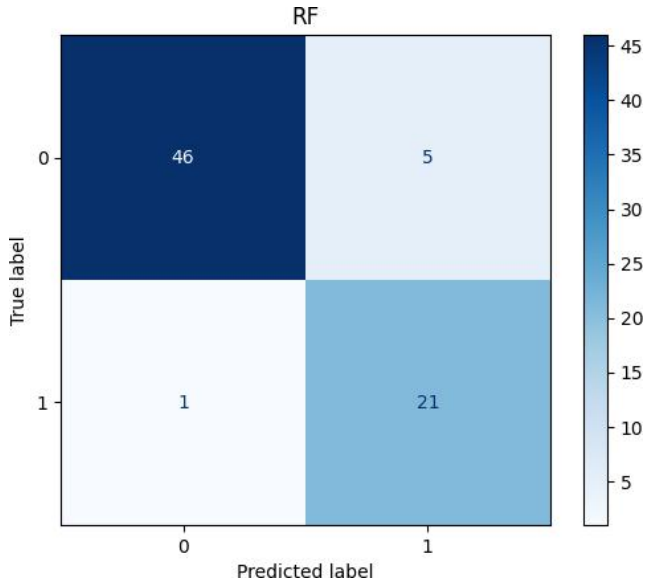

Figure 4. Confusion Matrix Displaying Random Forest Predictions.

4. Conclusion

Our study presented a machine learning methodology for predicting thyroid cancer recurrence from clinical and pathological patient records. Among the evaluated models, XGBoost achieved superior diagnostic accuracy, reaching 95.89% accuracy, 91.30% precision, 95.45% recall, 93.33% F1-score, and 97.95% ROC-AUC, outperforming KNN, SVM, and Random Forest.

ISSN : 2455-135X

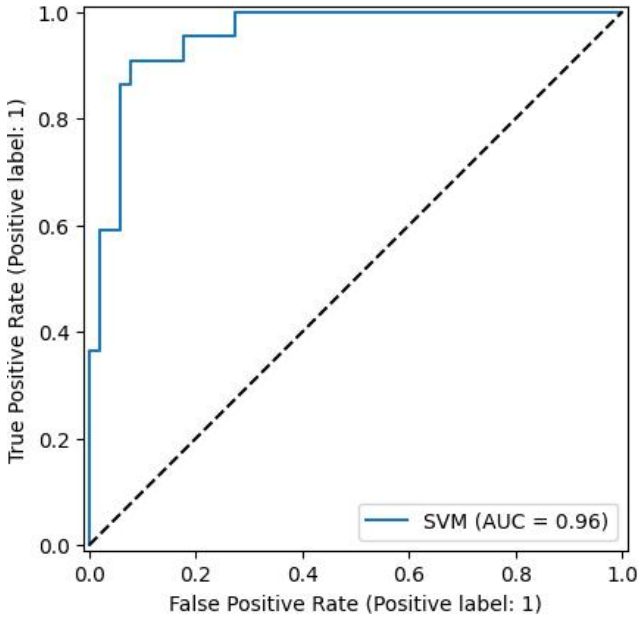

Figure 5. ROC-AUC Performance Curve for the SVM Classifier.

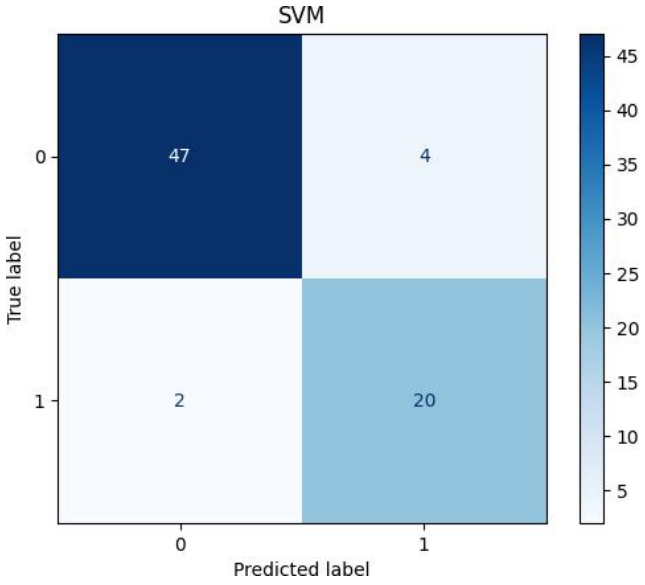

Figure 6. Confusion Matrix Displaying SVM Predictions.

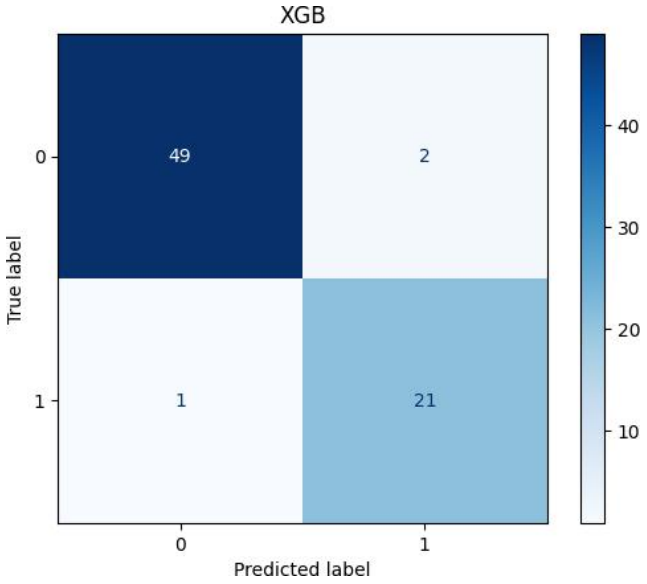

Figure 7. Confusion Matrix Displaying XGBoost Predictions.

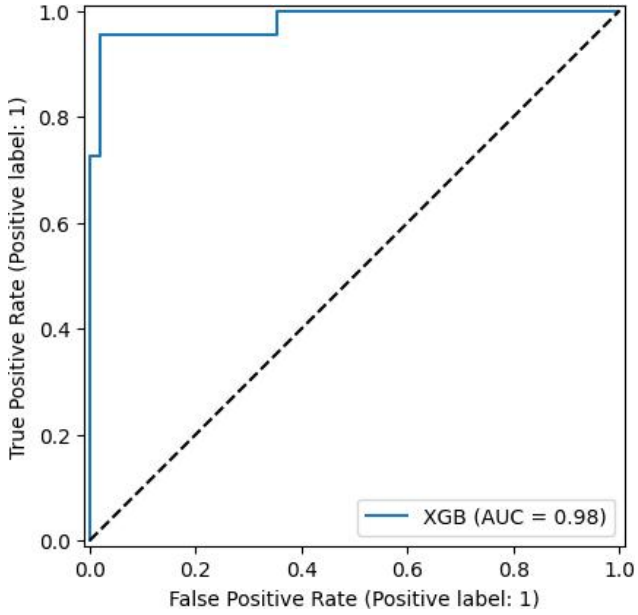

Figure 8. ROC-AUC Performance Curve for the XGBoost Classifier.

5. Future Work

Subsequent research will focus on validating this pipeline using multi-center institutional repositories to ensure broad clinical generalizability. Furthermore, future iterations will integrate explainable AI frameworks (XAI), such as SHAP and LIME, providing medical practitioners with transparent and actionable model interpretability.

Conflict of Interest

We, the authors declare that we have no conflict of interest.

Acknowledgment

We would like tp express our sincere gratitude to the Visvesvaraya PhD Scheme for providing fellowship for this research work.

ISSN : 2455-135X

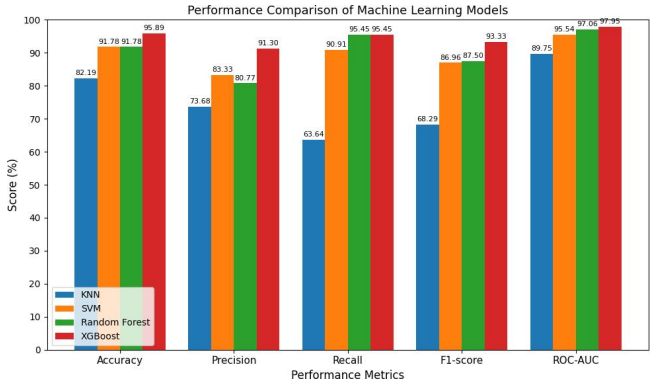

Figure 9. Comparative Overview of Model Metrics across Evaluation Benchmarks.

References

[1] M. Vanderpump, "The epidemiology of thyroid disease," *British Medical Bulletin*, vol. 99, no. 1, pp. 39–51, 2011.

[2] J. G. Hollowell, "Serum TSH, T4, and thyroid antibodies in the United States population (1988 to 1994): National Health and Nutrition Examination Survey (NHANES III)," *Journal of Clinical Endocrinology & Metabolism*, vol. 87, no. 2, pp. 489–499, Feb. 2002.

[3] F. Pedregosa *et al.*, "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, Nov. 2011.

[4] L. Breiman, "Random forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001.

[5] A. Raza, F. Eid, E. C. Montero, I. D. Noya, and I. Ashraf, "Enhanced interpretable thyroid disease diagnosis by leveraging synthetic oversampling and machine learning models," *BMC Medical Informatics and Decision Making*, vol. 24, no. 1, p. 364, Nov. 2024.

[6] A. Tops¸ir, F. Güler, E. Çetin, M. F. Burak, and M. Ag˘raz, "Thyroid disease classification using generative adversarial networks and Kolmogorov–Arnold network for three-class classification," *BMC Medical Informatics and Decision Making*, vol. 25, no. 1, p. 284, Jul. 2025.

[7] A. Hassan *et al.*, "Improving thyroid disorder diagnosis via innovative stacking ensemble learning model," *Digital Health*, vol. 11, Art. no. 20552076251341430, May 2025.

[8] S. Akter and H. A. Mustafa, "Analysis and interpretability of machine learning models to classify thyroid disease," *PLoS ONE*, vol. 19, no. 5, Art. no. e0300670, May 2024.

[9] D. Kumar, D. Raj, A. Singh, T. Rahman, and G. Sharma, "A comprehensive framework for thyroid disorder diagnosis: Integrating advanced feature selection, genetic algorithms and machine learning," *PLoS ONE*, vol. 20, no. 6, Art. no. e0325900, 2025.

[10] D. Kesavulu and R. Kannadasan, "Improved bio-inspired with machine learning computing approach for thyroid prediction," *Scientific Reports*, vol. 15, no. 1, p. 22524, Jul. 2025.

[11] G. Mollica *et al.*, "Classification of thyroid diseases using machine learning and Bayesian graph algorithms," *IFAC-PapersOnLine*, vol. 55, no. 40, pp. 67–72, 2022.

[12] V. Kumar *et al.*, "Automated segmentation of thyroid nodule, gland, and cystic components from ultrasound images using deep learning," *IEEE Access*, vol. 8, pp. 63482–63496, 2020.

[13] V. D. Leitch, J. D. Bassett, and G. R. Williams, "Role of thyroid hormones in craniofacial development," *Nature Reviews Endocrinology*, vol. 16, no. 3, pp. 147–164, 2020.

[14] K. Paczkowska *et al.*, "Laboratory interference in the thyroid function test," *Endokrynologia Polska*, vol. 71, no. 6, pp. 551–560, 2020.

[15] D. Dua and C. Graff, "UCI Machine Learning Repository," 2017. [Online]. Available: http://archive.ics.uci.edu/ml.

[16] S. Razia, P. S. Kumar, and A. S. Rao, "Machine learning techniques for thyroid disease diagnosis: A systematic review," in *Modern Approaches in Machine Learning and Cognitive Science*, vol. 885, pp. 203–212, 2020.

[17] K. Polat, S. S¸ ahan, and S. Günes¸, "A novel hybrid method based on artificial immune recognition system (AIRS) with fuzzy weighted pre-processing for thyroid disease diagnosis," *Expert Systems with Applications*, vol. 32, no. 4, pp. 1141–1147, 2007.

[18] A. Akbas, U. Turhal, S. Babur, and C. Avci, "Performance improvement with combining multiple approaches to diagnosis of thyroid cancer," *Engineering*, vol. 5, no. 10, pp. 264–267, 2013.

[19] W. Ahmad *et al.*, "Thyroid diseases forecasting using a hybrid decision support system based on ANFIS, k-NN and information gain method," *Journal of Applied Environmental and Biological Sciences*, vol. 7, no. 10, pp. 78–85, 2017.

[20] K. S. Parikh *et al.*, "Diagnosing common skin diseases using soft computing techniques," *International Journal of Bio-Science and Bio-Technology*, vol. 7, no. 6, pp. 275–286, 2015.

[21] L. Ozyilmaz and T. Yildirim, "Diagnosis of thyroid disease using artificial neural network methods," in *Proc. 9th Int. Conf. Neural Information Processing (ICONIP)*, vol. 4, 2002, pp. 2033–2036.

[22] I. Ioni¸ta˘ and L. Ioni¸ta˘, "Prediction of thyroid disease using data mining techniques," *BRAIN: Broad Research in Artificial Intelligence and Neuroscience*, vol. 7, no. 3, pp. 115–124, 2016.

[23] V. Talasila *et al.*, "The prediction of diseases using rough set theory with recurrent neural network in big data analytics," *International Journal of Intelligent Engineering and Systems*, vol. 13, no. 5, pp. 10–18, 2020.

[24] A. Keles and A. Keles¸, "ESTDD: Expert system for thyroid diseases diagnosis," *Expert Systems with Applications*, vol. 34, no. 1, pp. 242–246, 2008.

[25] A. Bastias *et al.*, "Predictive model based on neural networks to assist the diagnosis of malignancy of thyroid nodules," in *Proc. 41st Int. Conf. Computers & Industrial Engineering*, 2011, pp. 92–97.

ISSN : 2455-135X

[26] A. T. Azar, S. A. El-Said, and A. E. Hassanien, "Fuzzy and hard clustering analysis for thyroid disease," *Computer Methods and Programs in Biomedicine*, vol. 111, no. 1, pp. 1–16, 2013.

[27] D. Chen *et al.*, "A deep-learning based ultrasound text classifier for predicting benign and malignant thyroid nodules," in *Proc. Int. Conf. Green Informatics (ICGI)*, 2017, pp. 199–204.

[28] V. Chaurasia, S. Pal, and B. Tiwari, "Prediction of benign and malignant breast cancer using data mining techniques," *Journal of Algorithms & Computational Technology*, vol. 12, no. 2, pp. 119–126, 2018.

[29] L. Aversano *et al.*, "Thyroid disease treatment prediction with machine learning approaches," *Procedia Computer Science*, vol. 192, pp. 1031–1040, 2021.

[30] H. H. Kumar, "A novel approach of SVM based classification on thyroid disease stage detection," in *Proc. 3rd Int. Conf. Smart Systems and Inventive Technology (ICSSIT)*, 2020, pp. 836–841.

[31] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: Synthetic Minority Over-sampling Technique," *Journal of Artificial Intelligence Research*, vol. 16, pp. 321–357, 2002.

[32] I. Tomek, "Two Modifications of CNN," *IEEE Transactions on Systems, Man, and Cybernetics*, vol. 6, no. 11, pp. 769–772, Nov. 1976.

[33] G. Lemaître, F. Nogueira, and C. K. Aridas, "Imbalanced-learn: A Python Toolbox to Tackle the Curse of Imbalanced Datasets in Machine Learning," *Journal of Machine Learning Research*, vol. 18, no. 17, pp. 1–5, 2017.

[34] T. M. Cover and P. E. Hart, "Nearest Neighbor Pattern Classification," *IEEE Transactions on Information Theory*, vol. 13, no. 1, pp. 21–27, Jan. 1967.

[35] C. Cortes and V. Vapnik, "Support-Vector Networks," *Machine Learning*, vol. 20, no. 3, pp. 273–297, 1995.

[36] T. Chen and C. Guestrin, "XGBoost: A Scalable Tree Boosting System," in *Proc. 22nd ACM SIGKDD (KDD)*, San Francisco, CA, USA, 2016, pp. 785–794.

[37] J. Snoek, H. Larochelle, and R. P. Adams, "Practical Bayesian Optimization of Machine Learning Algorithms," in *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 25, 2012.

[38] D. M. W. Powers, "Evaluation: From Precision, Recall and F-Measure to ROC, Informedness, Markedness and Correlation," *Journal of Machine Learning Technologies*, vol. 2, no. 1, pp. 37–63, 2011.