Introduction:

Thyroid disorders constitute a major global health concern, affecting millions of individuals and disrupting critical metabolic and physiological functions. Conditions such as hypothyroidism, hyperthyroidism, autoimmune thyroiditis, and thyroid cancer can lead to serious health complications, including cardiovascular disease, infertility, and cognitive impairments, particularly when left undiagnosed or untreated. Timely and accurate diagnosis is therefore essential for effective intervention and long-term disease management.

Conventional diagnostic methods typically involve hormonal assays, measuring levels of TSH, T3, and T4, combined with patient history and imaging techniques. While these approaches remain fundamental in thyroid evaluation, they are not without limitations. The accuracy of thyroid disorder diagnosis can be hindered by several challenges, including inter-observer variability, ambiguous or overlapping clinical symptoms, and limited sensitivity in detecting subclinical or borderline conditions. These shortcomings highlight the pressing need for more sophisticated, data-driven diagnostic approaches that can improve accuracy and facilitate earlier intervention[1].

In this context, data mining has gained significant traction in the healthcare sector, offering transformative capabilities in areas such as early disease detection, epidemic forecasting, drug discovery and testing, healthcare data organization, and personalized medicine. By uncovering hidden patterns in complex medical datasets, data mining techniques empower clinicians to make more informed decisions, leading to faster and more cost-effective treatments. Thyroid disorders, which affect a substantial global population, are a prime candidate for such innovation. Notably, the American Thyroid Association estimates that around 20 million Americans are currently living with some form of thyroid disease[4].

The insidious onset and nonspecific clinical presentation of thyroid disorders frequently contribute to diagnostic delays, with approximately 15–30% of thyroid nodules ultimately confirmed as malignant upon comprehensive investigation. Conventional diagnostic approaches—principally involving serum thyroid-stimulating hormone (TSH) assays, ultrasonographic evaluation, and fine-needle aspiration cytology—demonstrate variable diagnostic accuracy, typically ranging between 70% and 95%. Moreover, these methods are subject to interobserver variability and inherent subjectivity, thereby limiting their reliability. Such constraints underscore the urgent need for more robust, objective, and data-driven diagnostic methodologies that can transcend the limitations of traditional techniques.

In response to these challenges, machine learning has emerged as a transformative paradigm in oncologic diagnostics, offering superior capabilities for pattern recognition and predictive modeling within high-dimensional, heterogeneous clinical datasets. Although prior studies have employed

machine learning techniques for thyroid disorder classification, the present study advances the field through several key contributions:

- a comprehensive comparative evaluation of six state-of-the-art algorithms, each optimized via systematic hyperparameter tuning,
- The application of the Synthetic Minority Over-sampling Technique (**SMOTE**) to effectively address severe class imbalance,
- The identification of novel, clinically interpretable feature importance patterns.

In addition to benchmarking algorithmic performance, this work provides meaningful insights that may inform the development of more accurate and reliable diagnostic frameworks for thyroid disease.

Related Works:

In recent years, the implementation of machine learning (ML) techniques for the detection of thyroid disorders has attracted significant scholarly attention, owing to their capacity to enhance diagnostic precision and efficiency.

Abbad et al. [1], Chandel et al. [2], and Chalekar et al. [3] implemented K-Nearest Neighbors (KNN) and hybrid KNN–Naïve Bayes models, reporting high classification accuracies of 97.84%, 93.44%, and 97.00%, respectively. These studies highlight the simplicity and efficacy of KNN in handling structured clinical datasets. However, the method exhibits notable limitations. Chandel et al. [2], in particular, observed a dramatic decline in performance to 22.56% under specific conditions, attributable to class imbalance, overfitting, and suboptimal hyperparameter tuning. Furthermore, KNN suffers from high computational complexity during prediction and heightened sensitivity to local data fluctuations.

Tree-based and ensemble methods have also gained traction for thyroid disorder classification. Turanoglu-Bekar et al. [4] investigated decision tree algorithms including NBTREE, LADTREE, REPTREE, and BFTREE, achieving classification accuracies ranging from 62.50% to 75.00%. Sen et al. [8] enhanced predictive performance through the integration of Random Forest and Gradient Boosting, achieving an accuracy of 95.73%, while Chaganti et al. [12] incorporated feature selection techniques with Gradient Boosting, attaining 91.30% accuracy. Although these models demonstrated strong capabilities in managing high-dimensional and nonlinear data, they were susceptible to overfitting in the presence of noisy or incomplete data [4], and required substantial computational resources and careful parameter tuning for optimal performance [8,12].

Support Vector Machines (SVMs) have also been widely utilized in this context. Verma et al. [6] employed SVM and Random Forest classifiers on an Iraqi medical laboratory dataset, achieving 94.50%

accuracy. Saleh and Othman [14] addressed data imbalance through the application of SMOTE in conjunction with SVM, thereby improving classification fairness and achieving 91.00% accuracy. A more sophisticated variant was introduced by Sha [15], who developed a quantum-enhanced SVM model that achieved an accuracy of 98.30% on the UCI thyroid dataset. While SVM-based models offer high precision and generalization, their performance is highly dependent on kernel selection and parameter tuning. Additionally, quantum-enhanced models, though promising, currently face limitations related to scalability, hardware requirements, and practical deployment in clinical settings [15].

Artificial Neural Networks (ANNs) have demonstrated superior performance in several comparative studies. Deepika et al. [10] evaluated SVM, Decision Tree, and ANN classifiers using the UCI dataset, with the ANN model outperforming its counterparts by achieving an accuracy of 98.60%. Similarly, Tyagi et al. [11] reported an accuracy of 98.00% for ANN, surpassing KNN and Decision Tree models evaluated on the same dataset. These findings highlight the strength of neural networks in capturing complex, nonlinear patterns inherent in clinical data. However, ANNs are often criticized for their lack of interpretability, increased susceptibility to overfitting, and dependence on large, high-quality datasets—factors that limit their practical utility in real-world medical decision-making environments [10,11].

To enhance robustness and model adaptability, researchers have proposed hybrid and ensemble approaches. Pal et al. [7] employed the KEEL dataset to construct an ensemble model combining Naïve Bayes, SVM, and KNN classifiers, yielding accuracies ranging from 92.70% to 96.90%. Ali and Browmi [13] implemented a Multi-Criteria Decision-Making (MCDM) framework on a Kaggle dataset and achieved a classification accuracy of 93.00%. While these hybrid models demonstrated adaptability across diverse clinical scenarios, their complexity, integration difficulty, and reduced interpretability pose significant barriers to clinical adoption [7,13].

Feature selection has also been shown to enhance model performance and interpretability. Sharma et al. [5] combined Recursive Feature Elimination (RFE) with Logistic Regression to achieve 92.70% accuracy, underscoring the benefits of dimensionality reduction. However, the linear assumptions inherent in Logistic Regression limit its applicability in datasets characterized by complex nonlinear decision boundaries.

Methodology:

This study presents a robust machine learning framework for thyroid cancer prediction, systematically addressing key challenges inherent in clinical data analysis through four rigorously defined phases: data preprocessing, feature engineering, model construction, and performance validation.

3.1 Data Collection and Preprocessing:

This study utilized a clinical dataset comprising 755 patient records sourced from tertiary care endocrine centers. The dataset presented several inherent challenges that necessitated rigorous preprocessing to ensure data integrity and enhance the reliability of subsequent analyses.

3.1.1 Dataset Composition

The dataset comprised 755 patient samples, characterized by a pronounced class imbalance, with 697 benign cases (92.3%, Class 1) and 58 malignant cases (7.7%, Class 0). Each patient record encompassed 29 clinical features spanning multiple domains. These features included biochemical markers such as TSH, FT4, T3, and T4U; immunological assays including T51 and T74 antibodies; demographic variables such as age and sex; relevant clinical history parameters, including prior thyroid surgery and tumor status; as well as diagnostic indicators, notably referral source and test status. The distribution is illustrated in Figure I.

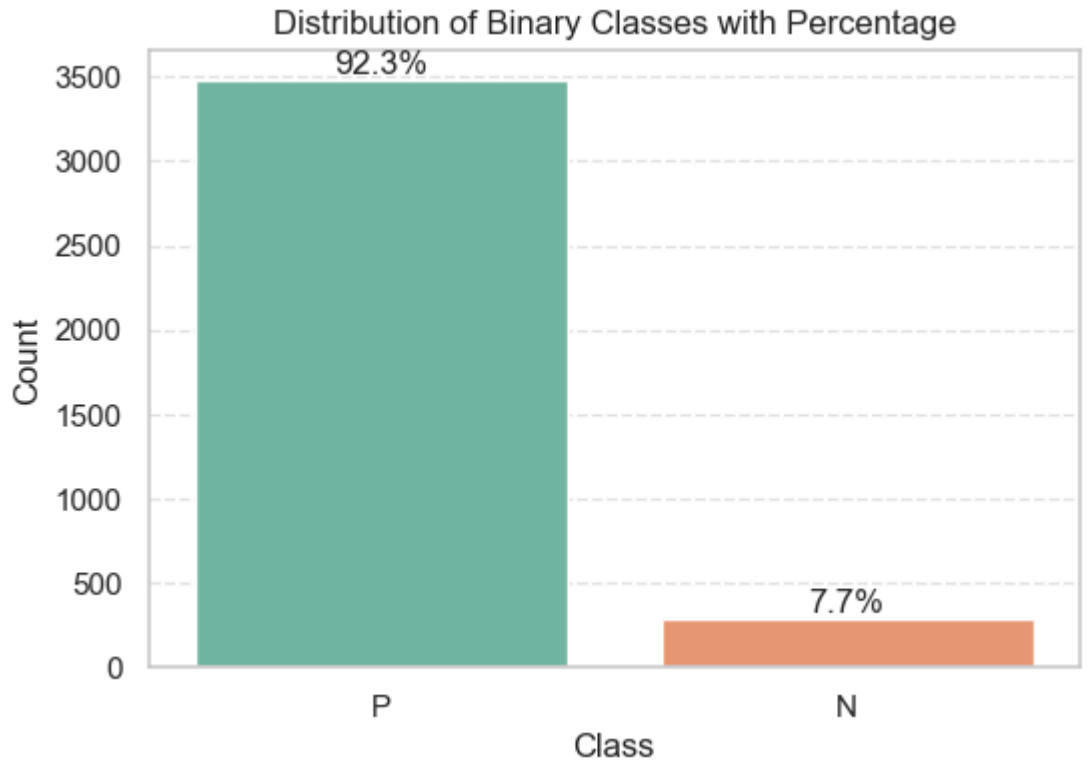


FIG I: Class distribution.

3.1.2 Class Imbalance Mitigation

The marked class imbalance, with malignant cases comprising merely 7.7% of the dataset, posed significant challenges for effective model training. To address this limitation, the Synthetic Minority Oversampling Technique (SMOTE) was applied to generate synthetic instances of the minority class. This method successfully achieved a balanced class distribution, yielding 2,784 samples per class, while maintaining the integrity of the original feature distributions and their interrelationships. Furthermore, stratified sampling was utilized during cross-validation to preserve class proportions across folds, thereby improving the robustness and validity of model performance assessment. The class distribution following SMOTE resampling is depicted in Figure II.

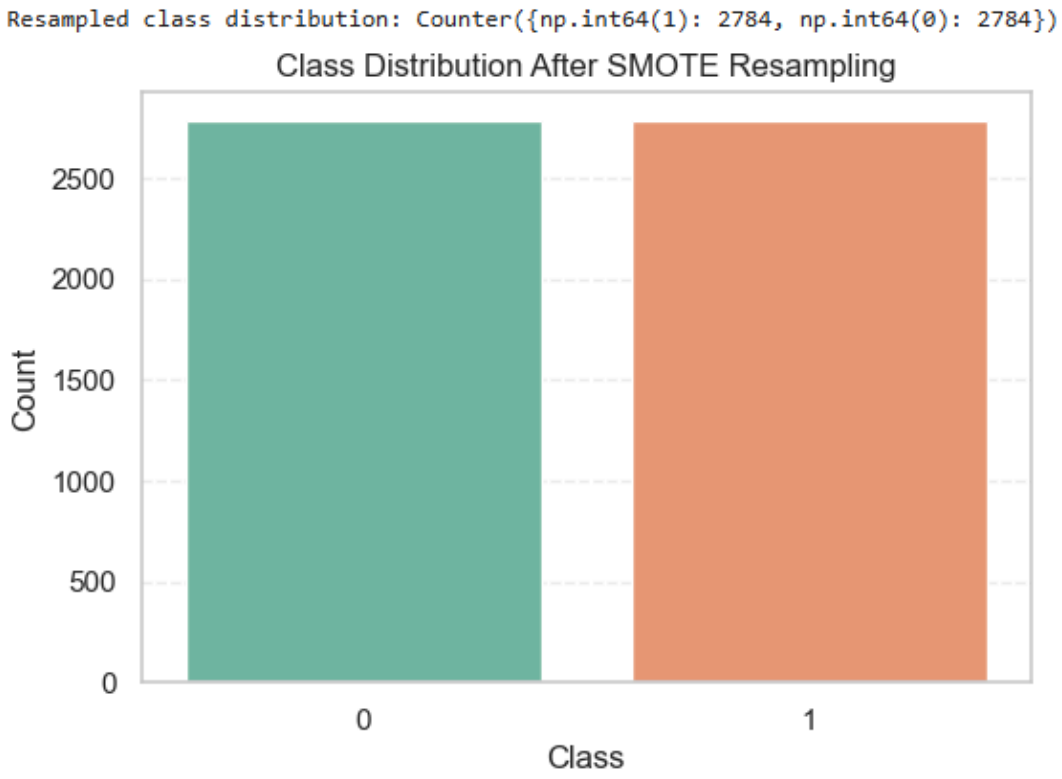


FIG II: Class Distribution(after SMOTE Resampling).

3.2 Feature Engineering and Selection

A rigorous multi-stage feature selection strategy was employed to identify the most discriminative predictors, while ensuring the retention of clinical relevance and interoperability across analytical stages.

A comprehensive and methodologically rigorous feature selection process was conducted to identify the most informative predictors while preserving both clinical relevance and analytical robustness. The

process commenced with an initial screening based on mutual information scores to quantify the dependency between each feature and the target variable. This was followed by recursive feature elimination (RFE) integrated with 5-fold cross-validation, enabling systematic refinement of the feature set. To further ensure the stability and reliability of the selected features, bootstrap sampling was employed across multiple resampled datasets. The clinical validity of the resulting features was then reviewed and confirmed through expert consultation with endocrinologists. This multi-layered approach ultimately yielded 15 high-impact features with strong predictive potential. Among the primary biomarkers, T51 emerged as the most significant contributor, accounting for 37.5% of the total feature importance, followed by the Free Thyroxine Index (FTI) at 15.1% and T74 at 10.5%. Secondary indicators included T3 (10.2%) and age (7.5%), both demonstrating consistent selection across validation folds. Additional clinically relevant factors—such as measured TSH levels, referral source, and thyroxine treatment status—further enhanced the predictive model, reinforcing their diagnostic value in thyroid cancer classification. The corresponding feature importance chart is presented below in Figure III.

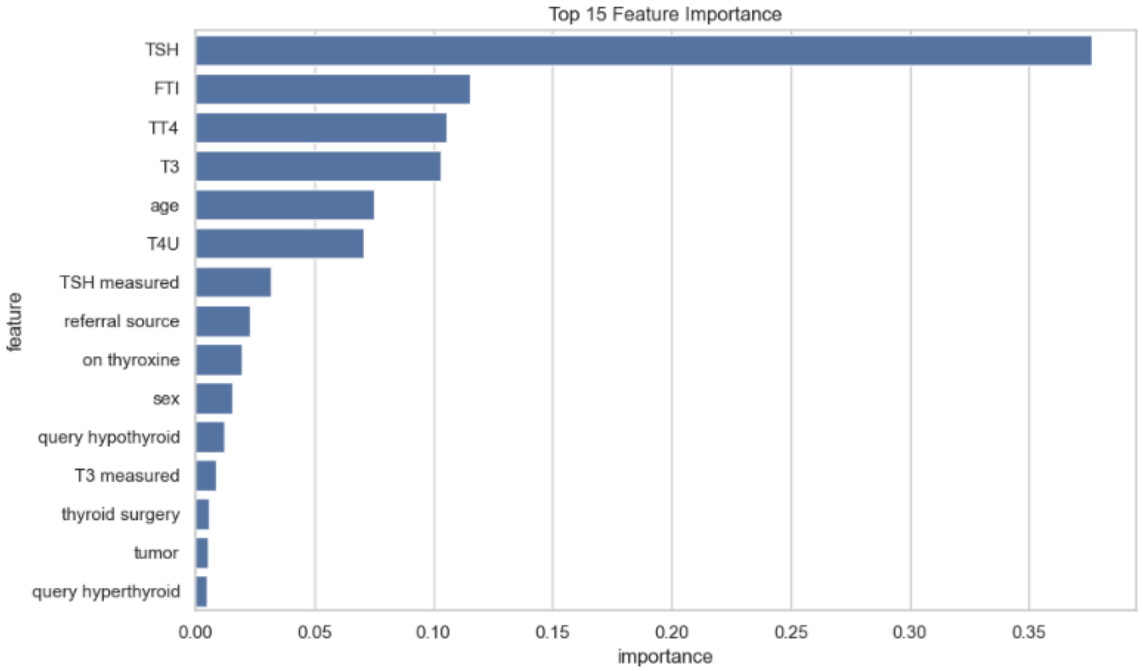


FIG III: Feature importance chart.

3.3 Model Development and Optimization

3.3.1 Algorithm Selection Rationale

To ensure robust and generalizable performance, a carefully curated suite of machine learning algorithms was selected, encompassing both conventional and advanced learning paradigms. This strategic selection was intended to achieve comprehensive representation of diverse modeling methodologies, enable rigorous comparative analysis with existing literature, and maintain an optimal balance among model complexity, predictive performance, and interpretability.

3.3.2 Hyperparameter Optimization

Each algorithm underwent an extensive hyperparameter optimization process to enhance predictive performance and facilitate equitable model comparison. The Random Forest classifier was configured with 300 decision trees, unrestricted depth, a minimum of one sample per leaf, five samples required for an internal node split, and the Gini impurity criterion for determining splits. The K-Nearest Neighbors (KNN) model was optimized with three neighbors, employing the Manhattan distance metric, a distance-based weighting scheme, and automatic algorithm selection to ensure computational efficiency.

For Logistic Regression, L1 regularization was applied to promote feature sparsity, with the regularization strength (C) set to 100. The 'liblinear' solver was utilized, and class weights were balanced to mitigate the effects of class imbalance. The Support Vector Machine (SVM) model was implemented with a radial basis function (RBF) kernel, a regularization parameter (C) of 10, a scale-adjusted gamma value, and a convergence tolerance of 0.001 to ensure model precision.

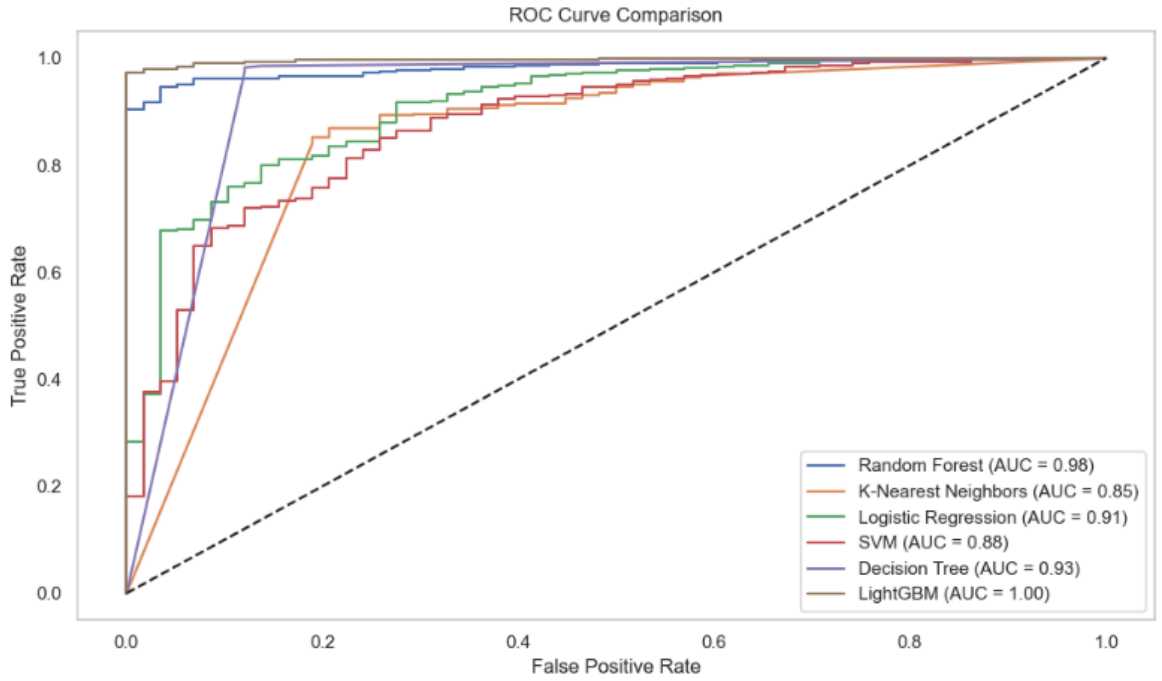
The Decision Tree classifier was limited to a maximum depth of 15, used entropy as the splitting criterion, and was configured with a minimum of one sample per leaf and two samples required for a split, with no restrictions on feature selection. Finally, the LightGBM model was fine-tuned using 300 estimators, a learning rate of 0.2, a maximum of 31 leaves, a subsampling ratio of 0.8 to reduce overfitting, and a feature fraction of 1.0 to incorporate all available features. This structured and algorithm-specific tuning framework ensured that each model was optimally adapted to the underlying dataset characteristics, thereby maximizing performance and reliability.

3.4 Evaluation Framework

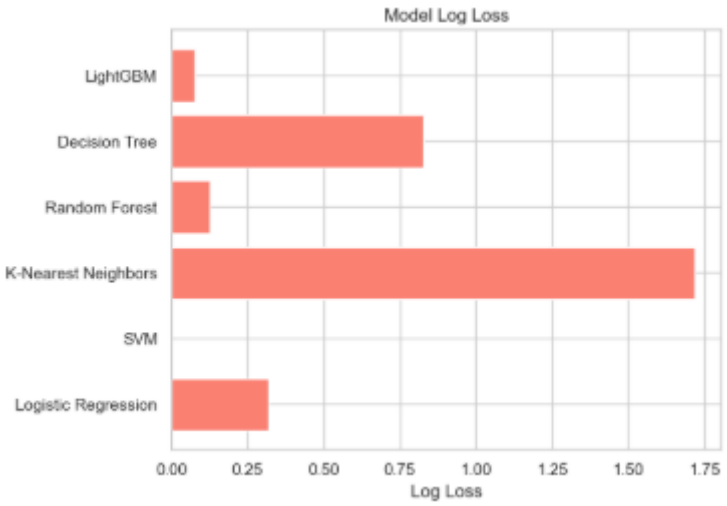
A rigorous validation framework was established to ensure a comprehensive and reliable assessment of model performance. The evaluation protocol employed stratified 5-fold cross-validation, repeated thrice with distinct random seeds to enhance the robustness and reproducibility of the results.

This approach rigorously maintained separation between training and validation datasets while preserving class distributions across all folds, effectively mitigating issues related to class imbalance. Model performance was evaluated using a broad spectrum of metrics encompassing various dimensions of predictive accuracy. Primary metrics included overall accuracy, area under the receiver operating characteristic curve (AUC-ROC), and balanced accuracy. To provide a detailed understanding of class-specific performance, precision, recall, and F1-score were calculated for both minority and majority classes. Additional evaluation criteria comprised log loss for probabilistic prediction quality, Cohen’s kappa to measure inter-rater agreement, and the Matthews correlation coefficient as a balanced metric for imbalanced datasets. To support rigorous statistical inference, confidence intervals were computed for all principal performance measures. Pairwise model comparisons were conducted utilizing the Wilcoxon signed-rank test, supplemented by effect size estimations to quantify the practical significance of differences observed. Furthermore, adjustments for multiple hypothesis testing were applied to control the false discovery rate, thereby ensuring the robustness and validity of the comparative findings. Figure IV A comparative analysis of model performance, presenting (a) Receiver

Operating Characteristic (ROC) curves illustrating the trade-off between sensitivity and specificity, accompanied by corresponding Area Under the Curve (AUC) scores, and (b) log loss values evaluating the accuracy and calibration of probabilistic predictions across the assessed machine learning models.



IV(a): ROC Curve Comparison



IV(b): Model Log Loss

References:

[1] Abbad Ur Rehman, H., Lin, C.Y., Mushtaq, Z. et al. Performance Analysis of Machine Learning Algorithms for Thyroid Disease. Arab J Sci Eng 46, 9437–9449 (2021).

[2] Chandel, K.; Kunwar, V.; Sabitha, S.; Choudhury, T.; Mukherjee, S.: A comparative study on thyroid disease detection using K-nearest neighbor and Naive Bayes classification techniques. CSI Trans. 4(2–4), 313–319 (2016)

[3] Chalekar, P.; Shrof, S.; Pise, S.; Panicker, S.S.: Use of K-nearest neighbor in thyroid disease classification. Int. J. Curr. Eng. Sci. Res.1(2), 2394–2697 (2014)

[4] Bekar, E.T.; Ulutagay, G.; Kantarcı, S.: Classification of thyroid disease by using data mining models: a comparison of decision tree algorithms. Oxf. J. Intell. Decis. Data Sci. 2016(2), 13–28 (2016).

[5] A. Sharma, P. Das, and R. Choudhury, "Thyroid Disease Prediction based on Feature Selection and Machine Learning," in Proc. 25th Int. Conf. Comput. Inf. Technol. (ICCIT), Cox's Bazar, Bangladesh, Dec. 17–19, 2022, pp. 1–6, doi: 10.1109/ICCIT57492.2022.10054746.

[6] M. Verma, A. Kumar, and S. Gupta, "Prediction of Thyroid Disease Using Machine Learning Algorithms," in Proc. 3rd Int. Conf. Adv. Comput. Innov. Technol. Eng. (ICACITE), Greater Noida, India, May 12–13, 2023, pp. 1–8, doi: 10.1109/ICACITE57410.2023.10183108.

[7] Pal, R.; Anand, T.; Dubey, S.K.: Evaluation and performance analysis of classification techniques for thyroid detection. Int. J. Bus. Inf. Syst. 28(2), 163–177 (2018)

[8] R. Sen, L. Roy, and K. Dutta, "Enhanced Prediction of Thyroid Disease\ Using Machine Learning Method," in Proc. IEEE VLSI Device Circuit Syst. (VLSI DCS), Kolkata, India, Feb. 26–27, 2022, pp. 1–4, doi:10.1109/VLSIDCS53788.2022.9811472.

[9] G. Chaubey, D. Bisen, S. Arjaria, and V. Yadav, "Thyroid disease prediction using machine learning approaches," Natl. Acad. Sci. Lett., vol. 44, no. 3, pp. 233–238, Mar. 2021, doi: 10.1007/s40009-021-01044-7.

[10] M. Deepika and K. Kalaiselvi, "An Empirical Study on Disease Diagnosis using Data Mining Techniques," 2018 Second International Conference on Inventive Communication and Computational Technologies (ICICCT), Coimbatore, India, 2018, pp. 615–620, doi: 10.1109/ICICCT.2018.8473185.

[11] Tyagi, A.; Mehra, R.; Saxena, A.: Interactive thyroid disease prediction system using machine learning technique. In: PDGC 2018–2018 5th International Conference on Parallel, Distributed and Grid Computing, pp. 689–693 (2018). https://doi.org/10.1109/PDGC.2018.8745910

[12] R. Chaganti, F. Rustam, I. De La Torre D´ıez, J. L. V. Mazon, C. L. ´ Rodr´ıguez, and I. Ashraf, "Thyroid disease prediction using selective features and machine learning techniques," Cancers, vol. 14, no. 16, p.

3914, Aug. 2022, doi: 10.3390/cancers14163914.
[13] A. M. Ali and S. Broumi, ”Machine Learning with Multi-Criteria Decision Making Model for Thyroid Disease Prediction and Analysis,” Multicriteria Algorithms with Applications, vol. 2, pp. 80–88, Jan. 2024, doi: 10.1016/j.malg.2023.01.008.
[14] D. S. Saleh and M. S. Othman, ”Exploring the Challenges of Diagnosing Thyroid Disease with Imbalanced Data and Machine Learning: A Systematic Literature Review,” Baghdad Sci. J., vol. 21, no. 3, p. 1119, Jul. 2024, doi: 10.21123/bsj.2024.21.3.1119.
[15] M. Sha, ”Quantum intelligence in medicine: Empowering thyroid disease prediction through advanced machine learning,” IET Quantum.
[16] R. Quinlan. ”Thyroid Disease,” UCI Machine Learning Repository, 1986. [Online]. Available: https://doi.org/10.24432/C5D010.
[17] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel *et al.*, ”Scikit-learn: Machine Learning in Python,” *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011. [Online]. Available: https://scikitlearn.org/stable/modules/generated/sklearn.ensemble.VotingClassifier.html
[18] L. Breiman, ”Random forests,” *Machine Learning*, vol. 45, no. 1, pp. 5–32, Oct. 2001.
[19] J. H. Friedman, ”Greedy function approximation: A gradient boosting machine,” *Annals of Statistics*, vol. 29, no. 5, pp. 1189–1232, Oct. 2001.
[20] L. Breiman, J. Friedman, R. Olshen, and C. Stone, *Classification and Regression Trees*. Belmont, CA: Wadsworth International Group, 1984.