54 Enhancing thyroid cancer diagnosis with ensemble machine learning techniques

Nomula Nagarjuna Reddy[a], *Samanvitha Maddula*[b], *Pogiri Deepika*[c], *Rakonda Sanjana*[d], *Lingadally Nipun*[e] *and Thoutireddy Shilpa*[f]

Department of Computer Science and Engineering, B V Raju Institute of Technology, Narsapur, Medak, Telangana, India

Abstract

The accurate detection of thyroid cancer (TC) remains a significant challenge in medical diagnostics. The research introduces a novel ensemble machine learning system designed to enhance the precision and reliability of thyroid cancer diagnosis. The proposed system employs a voting classifier (VC) that integrates random forest (RF), XGBoost (XGB), support vector machine (SVM) with RBF kernel, and logistic regression (LR), forming a robust predictive model. To address challenges of class imbalance and high-dimensionality, the system incorporates synthetic minority over-sampling technique (SMOTE) and principal component analysis (PCA). Additionally, StratifiedKFold (SKF) and GridSearchCV (GSCV) are utilized for optimization and validation, ensuring robust model performance. The integration of domain-specific features leveraging expert knowledge further improves model efficacy. Comprehensive data preprocessing, feature engineering, and hyperparameter tuning enhance model performance. The ensemble approach outperforms individual classifiers, achieving an accuracy of 81.1% and an F1-Score of 80.5%. These results underscore the effectiveness of advanced ensemble techniques in improving thyroid cancer diagnosis and offer a promising direction for future advancements in medical diagnostics.

Keywords: Ensemble learning, feature engineering, hyperparameter tuning, machine learning, medical diagnostics, thyroid cancer

Introduction

The thyroid gland, a small organ resembling a butterfly and located at the base of the neck, is integral to the regulation of important physiological processes like metabolism, body temperature, blood pressure, and heart rate, is the source of thyroid cancer (TC) (Figure 54.1). TC is still quite curable and has a great cure rate [1,2], despite being relatively rare when compared to other cancers. In recent decades, the cancer's prevalence has increased. The intricacy of TC, which can present as anaplastic, follicular, medullar, or papillary cancers, makes accurate diagnosis extremely difficult. With 9.6 million deaths from cancer in 2018, it is the second greatest cause of death worldwide. Cancer is defined by unchecked cell development that invades surrounding tissues and can spread to other bodily areas. People and health systems are severely strained financially, emotionally, and physically as a result of the worldwide cancer burden. While early detection and high-quality treatment have helped high-income countries improve management, many low and middle-income nations fail to provide timely and effective care, which has a negative impact on results. Given its various forms and the fact that its early stages are frequently asymptomatic, an accurate diagnosis of TC is essential. Even with improvements in diagnostic techniques, class imbalance and high-dimensional data continue to be problems in TC prediction.

The efficacy of treatment and the accuracy of forecasts are negatively impacted by the frequent shortcomings of traditional approaches in handling these complexities. This work focuses on improving TC diagnosis using sophisticated ensemble ML algorithms in order to overcome these issues. This approach seeks to enhance

[a]nagarjunareddynomula2@gmail.com, [b]maddulasamanvitha04@gmail.com, [c]pogirideepika2111@gmail.com, [d]sanju123sailaja@gmail.com, [e]nipun.poh@gmail.com, [f]shilpathoutireddy@gmail.com

DOI: 10.1201/9781003616399-54

treatment strategies and diagnostic accuracy by leveraging multiple ML models, PCA for dimensionality reduction, SMOTE for addressing class imbalance, and hyperparameter tuning. These developments could lead to improved TC treatment, improving patient outcomes and enabling a more efficient approach to this challenging illness. The ensuing parts provide an explanation of the same.

Research Objective

We aim to develop a system for accurate thyroid disease detection by addressing class imbalance and effective dimensionality reduction. It aims to maintain consistent class distribution during cross-validation, ensure systematic hyperparameter tuning to optimize model parameters, and enhance performance and generalizability. This approach is designed to significantly improve diagnostic accuracy and reliability, resulting in a more effective tool for better medical decision-making and patient outcomes.

Literature Review

Recent advancements in ML have significantly enhanced the diagnosis of TC, particularly with the use of ensemble methods and advanced feature selection techniques. RF has been especially effective, achieving an accuracy of 89% in one study [3], and improving further when combined with extra tree classifiers [4]. Other research [5] has shown that EL, coupled with detailed clinical feature analysis, can achieve high accuracy, sensitivity, and specificity. ANN and SVM have also been widely studied. ANN outperformed other algorithms in thyroid risk prediction with an F1-score of 0.957 [6], while SVM achieved 96% accuracy in a MIL-based CAD system for thyroid nodule classification [7]. Furthermore, a study [8] found XGB to outperform k-Nearest Neighbors and DT, suggesting that future research should explore LSTM for real-time analysis.

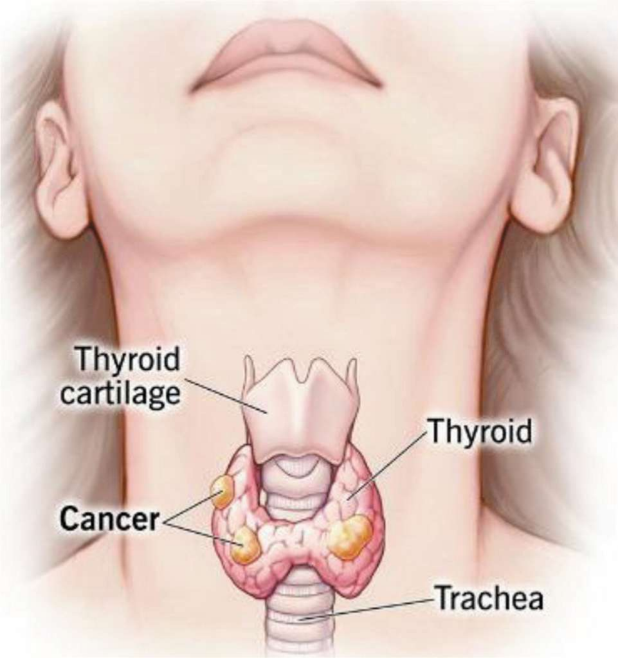


Figure 54.1 Thyroid cancer
Source: Author

Feature selection and dimensionality reduction are critical for improving model performance. Studies have demonstrated that chi-square-based selection [9] and filter-based methods [10] can lead to significant improvements in precision, recall, and F1-scores. Additionally, research [11] combining RF with LASSO identified key risk factors and achieved 99% accuracy in thyroid disease classification.

DL approaches have also been explored, with a study [12] integrating knowledge graphs and BLSTM models to leverage diverse medical data, enhancing diagnostic accuracy. Comparative studies have highlighted the strengths of specific algorithms, with MLP achieving 95.73% accuracy and an AUC of 94.23% in one case [13].

Other notable contributions include research on SVM, RF, and XGB for improving diagnosis [14], and using statistical moment-based features with RF to achieve high accuracy in HBP identification [15]. Another study [16] emphasized the importance of timely detection of TC, showing that RF, with an accuracy of 94.8%, can significantly enhance diagnostic outcomes from medical images.

Despite advancements, a gap remains in integrating advanced ensemble methods with feature engineering and real-time analysis in TC diagnosis. Current studies focus on individual models or simple ensembles, missing hybrid approaches. This research addresses this by introducing a novel ensemble framework combining RF, XGB, SVM, and LR with PCA and SMOTE. The framework optimizes performance through hyperparameter tuning and soft voting, enhancing predictive accuracy and efficiency in clinical settings.

Methodology

Dataset

This dataset comprises 1,232 thyroid nodules from 724 patients who had thyroidectomy surgery. It includes 16.23% nodules from male patients and 83.77% from female patients. Further details of this regard are explained in the Table 54.1.

System framework

The framework of this model (Figure 54.2) outlines an advanced methodology for detecting TC through sophisticated ML techniques. The process starts with the acquisition of a TC dataset, followed by a thorough data preprocessing phase. This includes data cleaning, addressing missing values, encoding categorical variables, and applying feature scaling. PCA is then utilized to retain 95% of the variance, while class imbalance is corrected using the SMOTE. A critical new feature, the FT4-to-FT3 ratio, is engineered due to its relevance in TC detection.

Subsequent to preprocessing, multiple ML algorithms namely RF, XGB, SVM, and LR are deployed on the processed data. Hyperparameter tuning is performed through grid search with cross-validation (CV) to refine model performance. An ensemble learning (EL) approach with soft voting is then employed to enhance prediction accuracy. The model's performance is assessed and the optimal model is chosen for making predictions. The process concludes with the successful detection of TC based on the input data, showcasing a robust and effective approach to leveraging ML diagnosis.

Data preprocessing

It plays a major role in preparing dataset for analysis and model building.

1) **Data cleaning:** The data preprocessing process begins with cleaning the dataset to ensure quality and consistency. This involves

Table 54.1 Description of features in dataset

S.no	Feature Name	Description
1	Age	Patient's age.
2	FT3	Result of the triiodothyronine hormone test.
3	FT4	Thyroxine hormone test result.
4	TSH	Thyroid-Stimulating hormone test result.
5	TPO	Test result for thyroid peroxidase Antibodies.
6	TGAb	Thyroglobulin antibodies test result.
7	Site	Nodule's location within the thyroid.
8	Echo pattern	Echogenicity of thyroid tissue.
9	Multifocality	Presence of multiple nodules in a single location.
10	Size	Measurement of nodule diameter.
11	Shape	Regularity of nodule's shape.
12	Margin	Clarity of the nodule's boundary.
13	Calcification	Presence of calcifications in the nodule.
14	Echo strength	Level of echogenicity in the nodule.
15	Blood flow	Blood flow status in the nodule.
16	Composition	Composition type of the nodule.
17	Multilateral	Presence of nodules in multiple locations.
18	Mal (Target)	Status of the nodule's malignancy.

Source: Nan Miles Xi, Lin Wangand Chuanjia Yang, "Improving The Diagnosis of Thyroid Cancer by Machine Learning and Clinical Data", Scientific Reports. Zenodo, Apr. 16, 2022. doi: 10.5281/zenodo.6465436.

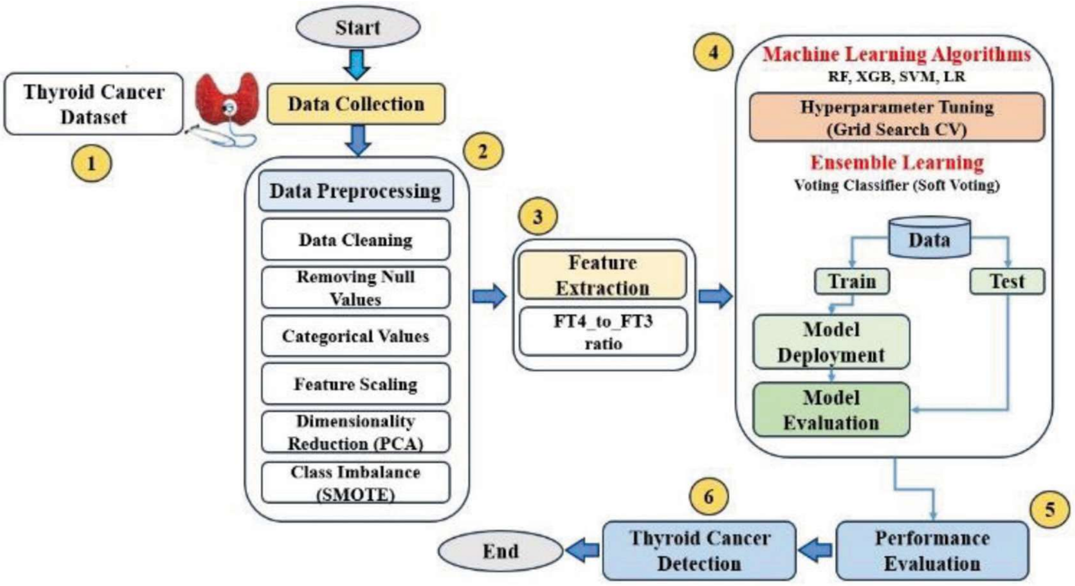


Figure 54.2 System framework
Source: Author

removing duplicate entries, which can introduce bias and distort analysis, and dropping irrelevant columns, such as unique identifiers that don't contribute to the model. After these steps, the dataset includes 19 variables, with patient ages ranging from 13 to 82 years.

2) **Handling categorical variables:** Categorical variables, such as gender and site, are modified into numerical format using one-hot encoding. This technique creates binary features for each category, allowing ML algorithms to process the data effectively. To avoid multicollinearity and manage the feature space, one category from each variable is dropped.
3) **Feature scaling***:* Continuous variables like age and hormone levels are standardized to ensure all features contribute equally to the model. Standardization transforms features to a mean of zero and a standard deviation of one, which is vital for the effective performance of algorithms like SVM and LR. It ensures no single feature overpowers the learning process and boosts the effectiveness of gradient-based algorithms.
4) **Dimensionality reduction using PCA:** To address potential issues with high-dimensional data, PCA is applied. It reduces the number of features while preserving most of the data's variance, simplifying the model and reducing noise. In this case, it retained 95% of the variance, enhancing the model's interpretability and generalization. Its uniqueness lies in its ability to simplify complex datasets by reducing noise and redundant features, making the data easier to analyze and improving the performance of machine learning models.
5) **Addressing class imbalance using SMOTE:** It balances class distribution by generating synthetic samples for the minority class, enhancing the model's ability to learn from imbalanced data and accurately predicting minority cases in new data. It is particularly valuable in medical diagnosis, where detecting rare but critical conditions is essential.

Feature engineering

It included the creation of the FT4-to-FT3 ratio, a domain- specific feature designed to enhance model performance by leveraging expert knowledge. FT4 and FT3 are critical thyroid hormones, and their ratio can offer greater diagnostic insight than considering the hormone levels independently. This ratio can reveal underlying physiological conditions that might not be evident when the hormones are analyzed separately, as certain thyroid disorders are characterized by imbalances in this ratio. By incorporating the FT4-to-FT3 ratio as a feature, the model is better equipped to understand thyroid function, potentially improving its ability to detect and classify abnormalities, while also reducing data dimensionality.

Proposed ensemble methodology

Random forest: It is an EL technique that builds multiple DTs and aggregates their predictions to improve accuracy and reduce overfitting. Each tree is trained on a randomly chosen subset of data and features, it introduces diversity among the trees, enhancing robustness. Key parameters include the number of trees, tree depth, and the minimum samples required for node splitting. RF effectively handles both categorical and numerical features, making it a strong choice for structured data problems, particularly in medical diagnosis.

Feature importance can be calculated using:

$$\mathrm{RF} = \frac{1}{N_{\mathrm{trees}}} \sum_{t=1}^{N_{\mathrm{trees}}} \Delta i_t \qquad (1)$$

where 'Δi_t' represents the decrease in impurity for feature 't' across all trees in the forest.

XGBoost: It is an efficient gradient boosting algorithm that excels at handling complex patterns, especially in imbalanced and sparse datasets. It constructs an additive model by sequentially training trees, where each tree corrects the errors of the previous one. Critical parameters encompass the number of trees, learning rate, and the depth of each tree, which help balance model complexity and prevent overfitting. XGB is highly accurate and well-suited for tasks requiring detailed pattern recognition, such as medical data analysis.

Logistic regression: It's a linear model for binary classification that predicts the probability of an outcome using given predictor variables. It is valued for its simplicity and interpretability, making it useful for understanding relationships between features and outcomes. The model's performance is influenced by regularization strength, which regulates the trade-off between fitting the training data and ensuring model simplicity. Despite being a baseline model, it often performs well in cases where there is a linear relationship between the features and the target outcome. It is calculated using sigmoid function.

$$P(z = 1|Y) = \frac{1}{1 + \exp\left(-(a_0 + \sum_{i=1}^{n} a_i y_i)\right)} \qquad (2)$$

where a_0 is the intercept and a_i are the coefficients associated with the predictor variables y_i.

Support vector machine with RBF kernel*:* It is a powerful classifier for non-linear decision boundaries. It projects input features into a higher-dimensional space, facilitating the model's ability to uncover complex data relationships. Flexibility of the decision boundary is controlled by the penalty parameter and the kernel coefficient. It is particularly effective in high-dimensional spaces and for classification tasks with overlapping classes, making it suitable for distinguishing complex patterns in medical data. It is represented by decision function.

$$f(a) = \mathrm{sign}\left(\sum_{i=1}^{n} \alpha_i z_i K(a_i, a) + b\right) \qquad (3)$$

where $K(a_i, a) = \exp(-\gamma|a_i - a|^2)$ is the RBF kernel function, α_i are the Lagrange multipliers, z_i are the class labels, and b is the bias term.

Hyperparameter tuning

GridSearchCV: Hyperparameters for each model are optimized using GridSearchCV (GSCV), a method that systematically searches through a predefined range of hyperparameters to find the combination that yields the best performance on the training data. GSCV works by exhaustively evaluating all possible combinations of hyperparameters, running each combination through a model training process to determine which set provides the highest accuracy or other desired metrics.

StratifiedKFold: To ensure robust evaluation, this process is combined with CV, specifically employing StratifiedKFold (SKF). CV entails dividing the dataset into multiple folds, where the model is trained on some folds and evaluated on others, rotating through all possible fold combinations. SKF is a variation of this technique that maintains the balance of class distributions within each fold, ensuring that each subset is representative of the overall data, especially in cases of imbalanced datasets.

Combination of GSCV and SKF: By using SKF in conjunction with GSCV, the model undergoes thorough and fair evaluation across different data splits, mitigating overfitting risks and amplifying the model's performance on unseen data. This approach is particularly valuable in tasks like medical diagnosis, where precision and reliability are paramount.

Ensemble learning

We used an ensemble approach in this research by combining four ML models: RF, XGB, SVM with an RBF kernel, and LR. Each model has strengths and limitations that the ensemble aims to balance. RF is robust for high-dimensional data but can be biased and less interpretable. XGB excels with structured data but is prone to overfitting and complexity in tuning. SVM effectively finds non-linear boundaries but is computationally intensive and sensitive to feature scaling. LR is simple and interpretable but limited to linear relationships. The ensemble leverages the advantages of various models while minimizing their individual weaknesses, resulting in a more balanced and effective predictive model.

Voting classifier: It is an EL technique integrating the predictions of various ML models elevating overall accuracy and robustness. It leverages the strengths of diverse models by combining their outputs. The voting classifier (VC) uses a Soft Voting mechanism. In soft voting, each model predicts class probabilities, and the final decision is based on the average of these probabilities. This method considers the confidence levels of each model, leading to more nuanced and accurate predictions. By averaging the predicted probabilities, soft voting effectively balances the strengths and weaknesses of individual models, reduces overfitting, and improves generalization, making it a powerful approach for creating robust predictive systems.

Model training and testing

Using an 80:20 split ratio and stratification to preserve class distribution, the dataset was allocated into training set for model development, and the testing set to evaluate the model's performance, ensuring an unbiased measure of its predictive accuracy.

Performance analysis

We use the following metrics to evaluate model performance, each providing distinct insights into the model's reliability and diagnostic capability. It delineates true negatives (TrNe), false negative (FaNe), and false positive (FaPo), true positives (TrPo).

1) **Accuracy (Acc):** It provides a general measure of model performance but can be deceptive in cases where datasets are imbalanced, where some classes are more prevalent than others.

$$\mathrm{Acc} = \frac{\mathrm{TrPo} + \mathrm{TrNe}}{\mathrm{TrPo} + \mathrm{TrNe} + \mathrm{FaPo} + \mathrm{FaNe}} \tag{4}$$

2) **Precision (Pre):** It's significant when the cost of FaPo is substantial, as it assesses how many predicted +ve cases are TrPo.

$$\mathrm{Pre} = \frac{\mathrm{TrPo}}{\mathrm{TrPo} + \mathrm{FaPo}} \tag{5}$$

3) **Recall (Re):** When the cost of FaNe is elevated, it becomes crucial as it reflects the model's capability in identifying all +ve cases.

$$\mathrm{Re} = \frac{\mathrm{TrPo}}{\mathrm{TrPo} + \mathrm{FaNe}} \tag{6}$$

4) **F1-Score (F1):** It is valuable when you need to harmonize Pre and Re, especially in the case of imbalanced datasets, providing a comprehensive measure of model performance than accuracy alone.

$$\mathrm{F1} = 2 \times \frac{\mathrm{Pre} \times \mathrm{Re}}{\mathrm{Pre} + \mathrm{Re}} \tag{7}$$

These metrics are used to evaluate the effectiveness of models in the ensemble and guide adjustments to improve their predictive model evaluation.

Results

The performance of individual ML models and the ensemble approach was evaluated as shown in Figure 54.3. RF model recorded an acc of 80.69% and an F1 of 80.08%, indicating strong effectiveness in identifying positive cases of TC. Key predictors identified by RF are shown in Figure 54.4. XGB performed robustly, with an acc of 79.67% and an F1 of 79.59%, managing the trade-off between precision and recall effectively (shown in Figure 54.5). The SVM with RBF kernel had an acc of 78.05% and an F1 of 77.59%, maintaining solid precision but slightly lower recall, as illustrated in Figure 54.6. LR demonstrated an acc of 77.64% and an F1 of 77.18%, struggling with complex patterns (as shown in Figure 54.7). The DT model attained the least, with an acc of 73.78% and an F1 of 73.18%, reflecting overfitting.

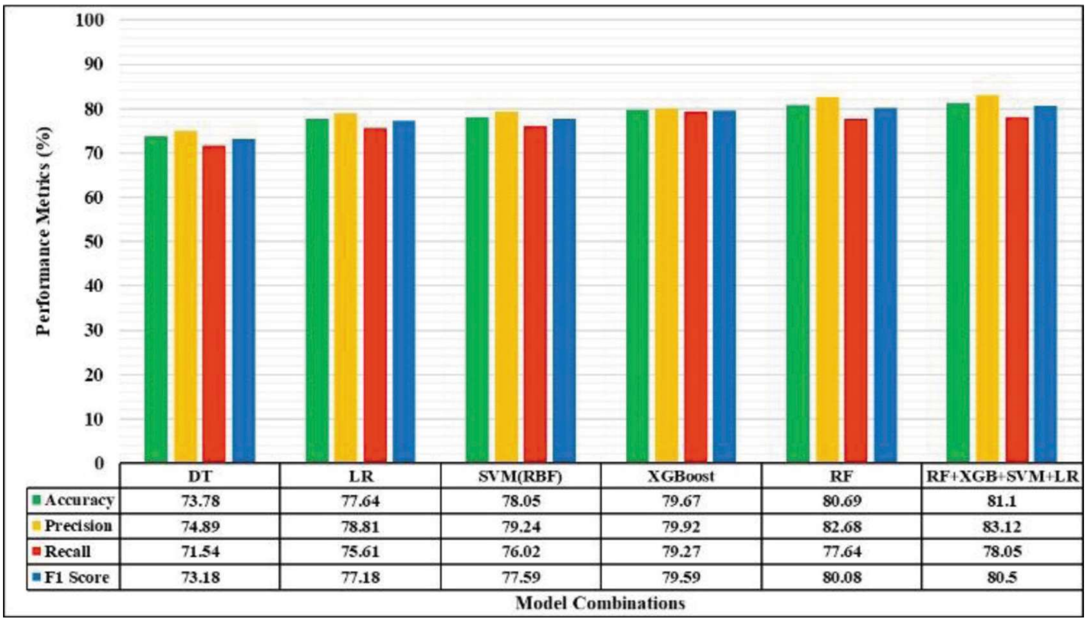
The ensemble model, combining RF, XGB, SVM, and LR through a VC, achieved the highest performance, recording an accuracy, precision, recall, F1 of 81.1%, 83.12%, 78.05%, and 80.5% respectively. These results underscore the impact of EL, as the combined strengths of the individual models led to improved overall performance, as shown in Figure 54.8. Additionally, Figure 54.9 provides a detailed analysis of the differences between the predicted outcomes of the individual models and the ensemble, offering further insights into the effectiveness of combining models.

Discussion

The results reveal the varied strengths and limitations of different ML models for TC classification. RF exhibited high precision and balanced performance, effectively managing complex data interactions, while XGB demonstrated similar robustness, balancing precision and recall effectively.

SVM with RBF kernel was effective but showed slightly lower recall, indicating room for improvement through further tuning or complex kernels. LR, despite its lower performance, provides a useful baseline and highlights the challenges linear models face with non-linear data. DT model performed the weakest, likely due to overfitting, suggesting that techniques like pruning or ensemble methods could improve performance. The ensemble model, combining RF, XGB, SVM, and LR via a VC, achieved the highest accuracy and precision. This model leverages the strengths of individual models, resulting in superior overall performance. Despite its slightly lower recall, the ensemble approach effectively mitigates the limitations of individual models.

Future research should focus on optimizing the individual models within the ensemble and exploring advanced techniques. Enhancing feature engineering and incorporating diverse datasets could further improve accuracy and robustness. Evaluating models in clinical settings and integrating expert feedback will also be crucial for practical applications.



	DT	LR	SVM(RBF)	XGBoost	RF	RF+XGB+SVM+LR
🟩 Accuracy	73.78	77.64	78.05	79.67	80.69	81.1
🟨 Precision	74.89	78.81	79.24	79.92	82.68	83.12
🟥 Recall	71.54	75.61	76.02	79.27	77.64	78.05
🟦 F1 Score	73.18	77.18	77.59	79.59	80.08	80.5

Figure 54.3 Existing and proposed model's performance comparison graph
Source: Author

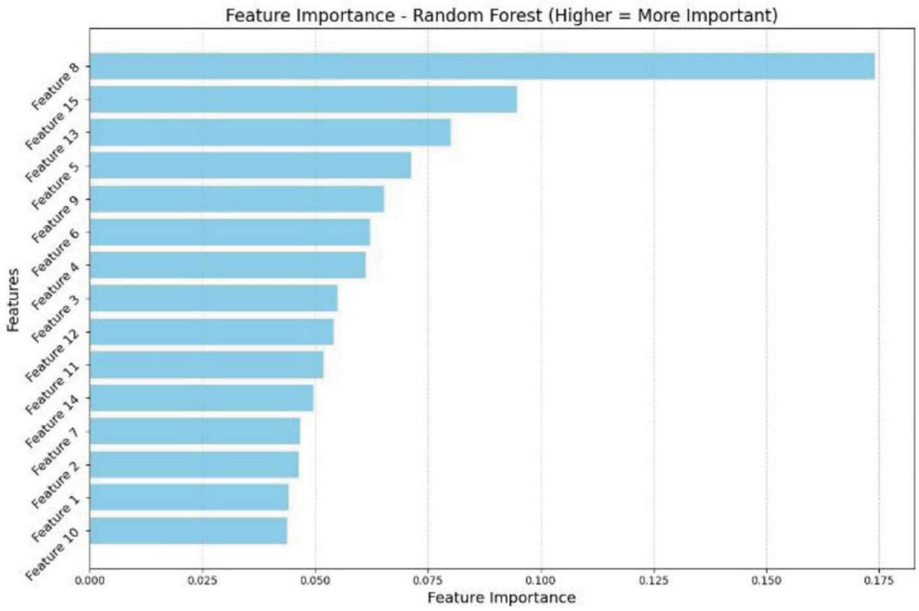


Figure 54.4 Random forest feature importance
Source: Author

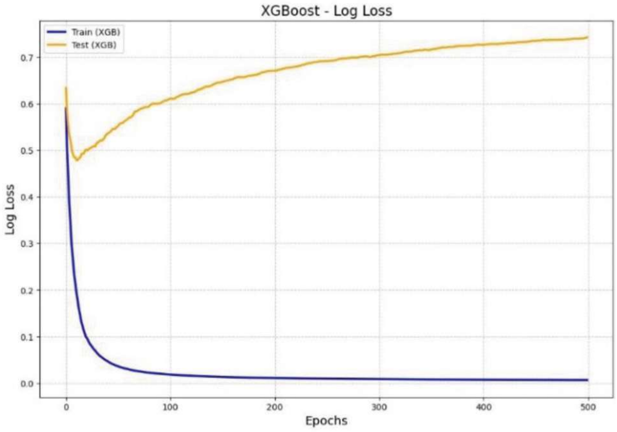


Figure 54.5 XGBoost training and loss curve
Source: Author

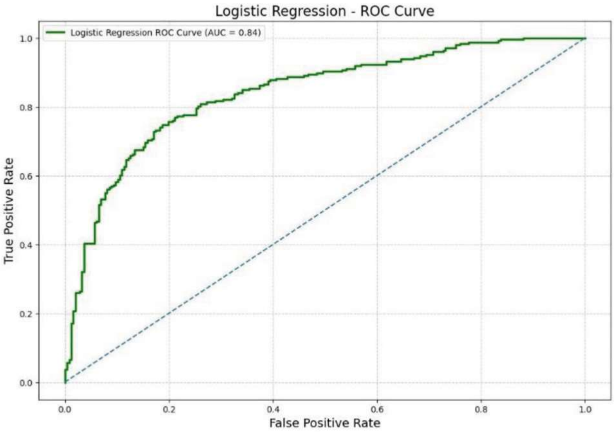


Figure 54.7 Logistic regression ROC
Source: Author

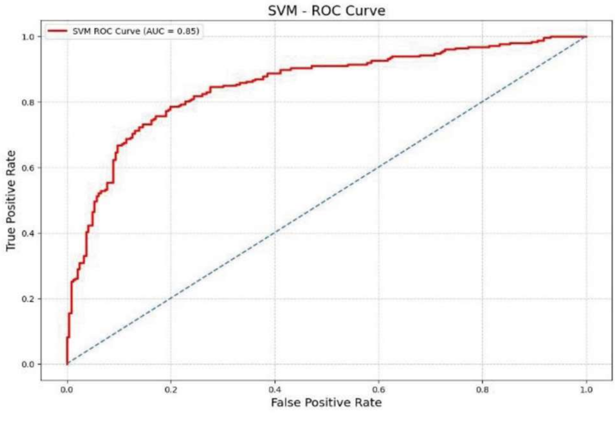


Figure 54.6 Support vector machine ROC
Source: Author

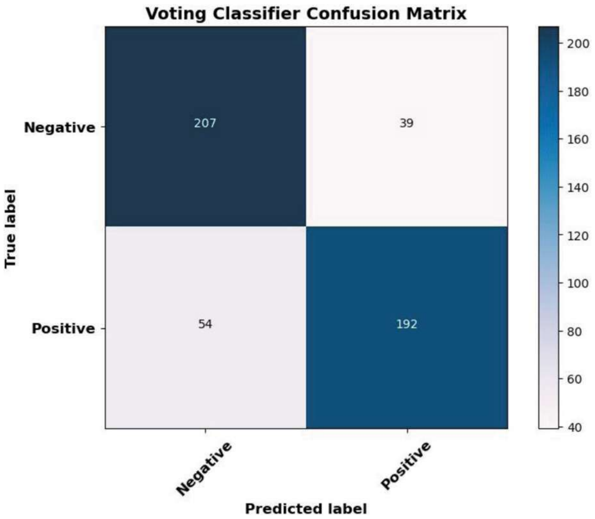


Figure 54.8 Voting classifier confusion matrix
Source: Author

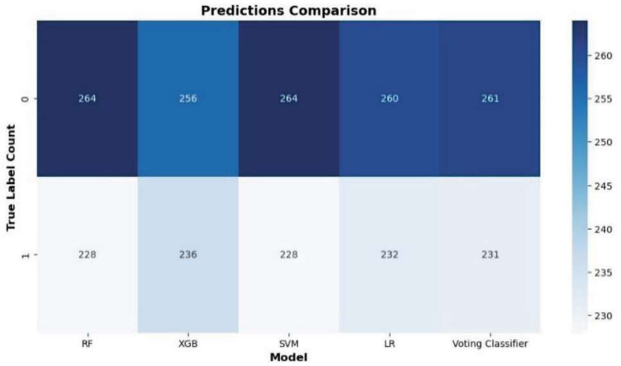


Figure 54.9 Prediction comparison matrix
Source: Author

In summary, the ensemble approach demonstrates significant improvements over individual models and holds promise for advancing medical diagnostics through enhanced accuracy and reliability.

Conclusion

This paper analyzed the efficacy of various ML models for thyroid cancer (TC) classification, including RF, XGB, SVM with RBF kernel, LR, and DT. The study utilized rigorous data preprocessing, feature engineering, SMOTE for addressing class imbalance, and PCA for dimensionality reduction, evaluating models based on evaluation metrics.

Our findings indicate that RF and XGB provided the highest performance, demonstrating strong accuracy and balanced metrics. The VC, integrating these models with SVM and LR, achieved the best overall results, underscoring the power of EL in enhancing classification accuracy.

Despite these advancements, the study faced limitations such as reliance on a single dataset, which may impact generalizability, and the increased complexity and computational cost introduced by the VC. Additionally, expanding the feature set could further improve model performance.

Future work should aim to diversify the dataset for better generalizability, explore advanced techniques like DL, and refine feature extraction methods. Further optimization of the ensemble model is also necessary to maintain a balance between performance and computational efficiency. This study provides valuable insights into to the healthcare domain by demonstrating the effectiveness of EL in improving diagnostic accuracy for thyroid disease. The success of the VC highlights its potential as a valuable tool for developing reliable diagnostic systems, setting the stage for future advancements in medical diagnostics and enhanced patient outcomes. In conclusion, this study underscores the potential of combining ML models to improve thyroid disease classification, addresses current limitations, and provides a foundation for future research aimed at advancing diagnostic technologies in medicine.

References

[1] Xi, N. M., Wang, L., and Yang, C. (2022). Author correction: improving the diagnosis of thyroid cancer by machine learning and clinical data. *Scientific Reports*, 12(1), 13252. doi: 10.1038/s41598-022-17659-1.

[2] Chaganti, R., Rustam, F., De La Torre Díez, I., Mazón, J. L. V., Rodríguez, C. L., and Ashraf, I. (2022). Thyroid disease prediction using selective features and machine learning techniques. *Cancers*, 14(16), 3914. doi: 10.3390/cancers14163914.

[3] Alshayeji, M. H. (2023). Early thyroid risk prediction by data mining and ensemble classifiers. *Machine Learning and Knowledge Extraction*, 5(3), 1195–1213. doi: 10.3390/make5030061.

[4] Islam, S. S., Haque, M. S., Miah, M. S. U., Sarwar, T. B., and Nugraha, R. (2022). Application of machine learning algorithms to predict the thyroid disease risk: an experimental comparative study. *PeerJ Computer Science*, 8, e898. doi: 10.7717/peerj-cs.898.

[5] Vadhiraj, V. V., Simpkin, A., O'Connell, J., Singh Ospina, N., Maraka, S., and O'Keeffe, D. T. (2021). Ultrasound image classification of thyroid nodules using machine learning techniques. *Medicina*, 57(6), 527. doi: 10.3390/medicina57060527.

[6] Sankar, S., Potti, A., Chandrika, G. N., and Ramasubbareddy, S. (2022). Thyroid disease prediction using XGBoost algorithms. *Journal of Micromechanics and Microengineering*, 18(3), 1–18. doi: 10.13052/jmm1550-4646.18322.

[7] Setiawan, K. E. (2024). Predicting recurrence in differentiated thyroid cancer: a comparative analysis of various machine learning models including ensemble methods with chi-squared feature selection. *Communications in Mathematical Biology and Neuroscience*. doi: 10.28919/cmbn/8506.

[8] Obaido, G., Achilonu, O., Ogbuokiri, B., Amadi, C. S., Habeebullahi, L., Ohalloran, T., et al. (2024). An improved framework for detecting thyroid disease using filter-based feature selection and stacking ensemble. *IEEE Access*, 12, 89098–89112. doi: 10.1109/ACCESS.2024.3418974.

[9] Sultana, A., and Islam, R. (2023). Machine learning framework with feature selection approaches for thyroid disease classification and associated risk factors identification. *Journal of Electrical Systems and Information Technology*, 10(1), 32. doi: 10.1186/s43067-023-00101-5.

[10] Chai, X. (2020). Diagnosis method of thyroid disease combining knowledge graph and deep learning. *IEEE Access*, 8, 149787–149795. doi: 10.1109/ACCESS.2020.3016676.

[11] Pal, M., Parija, S., and Panda, G. (2022). Enhanced prediction of thyroid disease using machine learning method. In 2022 IEEE VLSI Device Circuit and System (VLSI DCS), Kolkata, India: IEEE, (pp. 199–204). doi: 10.1109/VLSIDCS53788.2022.9811472.

[12] Kumari, P., Kaur, B., Rakhra, M., Deka, A., Byeon, H., Asenso, E., et al. (2024). Explainable artificial intelligence and machine learning algorithms for classification of thyroid disease. *Discover Applied Sciences*, 6(7), 360. doi: 10.1007/s42452-024-06068- w.

[13] Butt, A. H., Alkhalifah, T., Alturise, F., and Khan, Y. D. (2023). Ensemble learning for hormone binding protein prediction: a promising approach for early diagnosis of thyroid hormone disorders in serum. *Diagnostics*, 13(11), 1940. doi: 10.3390/diagnostics13111940.

[14] Alyas, T., Hamid, M., Alissa, K., Faiz, T., Tabassum, N., and Ahmad, A. (2022). [Retracted] empirical method for thyroid disease classification using a machine learning approach. *BioMed Research International*, 2022(1), 9809932. doi: 10.1155/2022/9809932.

[15] Shulhai, A. M., Rotondo, R., Petraroli, M., Patianna, V., Predieri, B., Iughetti, L., et al. (2024). The role of nutrition on thyroid function. *Nutrients*, 16(15), 2496. doi: 10.3390/nu16152496.

[16] Hollywood, J. B., Hutchinson, D., Feehery-Alpuerto, N., Whitfield, M., Davis, K., and Johnson, L. M. (2023). The effects of the paleo diet on autoimmune thyroid disease: a mixed methods review. *Journal of the American Nutrition Association*, 42(8), 727–736. doi: 10.1080/27697061.2022.2159570.