Jurnal Masyarakat Informatika, 16(2), 2025
DOI:10.14710/jmasif.16.2.75073
ISSN: 2777-0648

Comparative Evaluation of Machine Learning Algorithms with Data Balancing Approach and Hyperparameter Tuning in Predicting Thyroid Disorder Recurrence

Darnell Ignasius, Rhyan David Levandra, Ramadhan Rakhmat Sani*, Ika Novita Dewi
Faculty of Computer Science, Universitas Dian Nuswantoro, Semarang, Indonesia
* *Corresponding author: ramadhan_rs@dsn.dinus.ac.id*

Abstract
This research *evaluates and compares the performance of five machine learning algorithms (Logistic Regression, K-Nearest Neighbors, Decision Tree, Random Forest, and Gradient Boosting) in predicting thyroid disease recurrence using patient data. The analysis was conducted on the Thyroid Disease Dataset from the UCI Machine Learning Repository. The methodology includes data preprocessing, normalization, and class balancing with the Synthetic Minority Over-sampling Technique (SMOTE). Additionally, hyperparameter tuning was conducted using GridSearchCV to optimize model performance. The results demonstrate that ensemble-based models, specifically Random Forest and Gradient Boosting, consistently outperform the other algorithms in terms of accuracy and robustness. These models achieve 95–96% accuracy across various scenarios.A key finding is that SMOTE significantly improves recall for minority classes, highlighting its value in imbalanced medical datasets.*

Keywords: *Medical Diagnosis, Thyroid Disease, Machine Learning, SMOTE, Hyperparameter Tuning,*

1 Introduction

Thyroid dysfunction is among the most prevalent endocrine disorders worldwide, second only to diabetes. An estimated that 300 million cases of thyroid dysfunction are currently present globally [1]. A significant aspect of this disorder is the prevalence of undiagnosed cases, with approximately 50% of affected individuals remaining unaware of their condition [2]. This phenomenon underscores the critical nature of thyroid disorders as a public health concern that frequently eludes detection and awareness [3]. Early detection is of paramount importance, as the symptoms of thyroid disorders are often non-specific and may resemble those of other common conditions, such as fatigue or stress [4].

The thyroid gland, the largest endocrine gland in the human body, is located in the neck and plays an important role in regulating the body's metabolism [5]. The thyroid gland produces two main hormones, T3 (triiodothyronine) and T4 (thyroxine), which regulate the metabolic rate of cells and affect nearly all organs of the body [6]. An imbalance in the levels of these hormones can lead to two main conditions: hyperthyroidism (excess hormones) and hypothyroidism (deficiency of hormones). Each of these conditions can have a significant impact on the quality of life of the sufferer [7]. The symptoms associated with this condition can include sleep disturbances, mood swings, chronic fatigue, and serious complications in pregnancy [8]. The prevalence of thyroid disorders is notably high, particularly among the productive age group, and is more prevalent in women [9]. A multitude of factors, including hormonal, genetic, and immunological elements, have been identified as

Received: 25 June 2025; Revised: 1 October 2025; Accepted: 10 November 2025; Published: 30 November 2025

contributing to the development of these disorders [10]. However, delayed diagnosis remains a pervasive issue in numerous healthcare facilities, particularly in regions with constrained diagnostic resources [11]. It is noteworthy that many patients only become aware of their condition after the emergence of significant complications that necessitate additional treatment. Consequently, there is an imperative for initiatives aimed at enhancing access to precise and economical diagnostic techniques.

One of the most frequently employed diagnostic methods in contemporary medicine is the Fine Needle Aspiration Biopsy (FNAB), which entails the extraction of thyroid tissue samples for subsequent analysis [12]. Despite its notable efficacy, this method is subject to certain limitations, including a reported accuracy of approximately 62.2%, which necessitates validation through histopathological examination, as the prevailing standard of truth. Furthermore, the procedure is considered to be invasive and necessitates the expertise of trained medical professionals, whose availability is not guaranteed in all healthcare settings. This underscores the necessity for a more pragmatic, effective, and dependable solution, particularly in resource-limited settings [13].In recent years, the application of machine learning technology in the medical field has demonstrated considerable potential, particularly in the diagnosis of diseases that involve complex data analysis, such as thyroid disorders [14]. The utilization of patient medical data, including TSH, T3, T4, age, gender, and medical history, enables the training of machine learning algorithms to discern patterns that are imperceptible to conventional observation [15]. This model has the capacity to provide precise and automated predictions of thyroid function status, even in circumstances where data is incomplete or variable, thus making it an optimal solution for supporting the clinical decision-making process [16].

Recent advances in machine learning enable the development and assessment of that can detect thyroid dysfunction and recurrence with greater accuracy. However, medical datasets often characterized by skewed class distributions, with positive cases such as recurrence being significantly underrepresented. This imbalance can lead to models that are biased and fail to detect critical but rare cases. To address this issue, synthetic data balancing techniques such as SMOTE (Synthetic Minority Over-sampling Technique) can be employed to improve minority class recall [17].In addition to class balancing, optimising the performance of an algorithm requires the systematic tuning of its hyperparameters. Default hyperparameters are rarely optimal for datasets with different distributions and noise levels. Techniques such as GridSearchCV enable the systematic exploration of different hyperparameter combinations to identify the optimal model configuration, which can significantly impact the accuracy of predictions, especially in sensitive medical fields [18].

Due to the varying capabilities and learning mechanisms of different machine learning approaches, it is essential to conduct a comprehensive comparative analysis of multiple models to determine which algorithm exhibits the highest performance, stability, and generalization capacity when applied to medical datasets. Each model category employs a distinct strategy for learning from data [19]. Linear models, such as logistic regression, are ideal for problems where the relationship between features and outcomes is approximately linear. These models offer interpretability and computational simplicity. In contrast, distance-based methods such as KNN rely on the proximity of data points in feature space [20]. This can be effective for non-linear relationships, but the method is often sensitive to noise and computationally expensive in large datasets. Rule-based classifiers, such as decision trees, provide intuitive if-then decision logic, enabling model transparency and ease of interpretation; however, they are prone to overfitting, especially when left unpruned. Finally, ensemble

learning methods, including Random Forest and Gradient Boosting, aggregate multiple weak learners to form a robust model that can capture complex patterns and interactions among features [21]. These ensemble techniques are renowned for their ability to mitigate overfitting and enhance predictive accuracy, rendering them highly suitable for heterogeneous and high-dimensional data, common in medical diagnostics.

Recent studies have further emphasized the importance of handling class imbalance and optimization strategies in thyroid disease prediction. Clark et al [22] demonstrated that Random Forest consistently outperformed other classifiers, with its performance enhanced through SMOTE and hyperparameter tuning, highlighting the importance of these optimization techniques. Similarly, Atay et al [23] introduced a hybrid model combining association rule mining with classification algorithms, which yielded interpretable and highly accurate predictions for differentiated thyroid cancer recurrence. Reinforcing this point, another investigation by Agarwal et al [24] also applied SMOTE with conventional classifiers such as SVM, KNN, and Random Forest, reported notable gains in accuracy and recall, further reinforcing the significance of oversampling in medical datasets.

Despite these advancements, existing work often remains limited to evaluating a single algorithm or does not systematically compare model performance across different preprocessing and tuning conditions. This research contributes by conducting a comprehensive comparative evaluation of five representative algorithms, explicitly analyzing their behavior under both balanced and imbalanced data conditions, and before and after hyperparameter optimization. In doing so, the study not only consolidates prior evidence but also clarifies how different algorithms respond to balancing and tuning strategies, thereby offering methodological insights and practical implications for developing clinical decision support systems for predicting thyroid disorder recurrence.

2 Research Methods

The methodological framework employed in this research follows a structured, data-centric strategy that is aimed at effectively addressing the critical issue of data imbalance. As shown in Figure 1, the framework consists of a series of sequential stages, starting with data acquisition and ending with model evaluation. Each stage is designed to maintain statistical rigor, ensure reproducibility, and support adaptability across various classification algorithms, particularly when applied to imbalanced datasets.

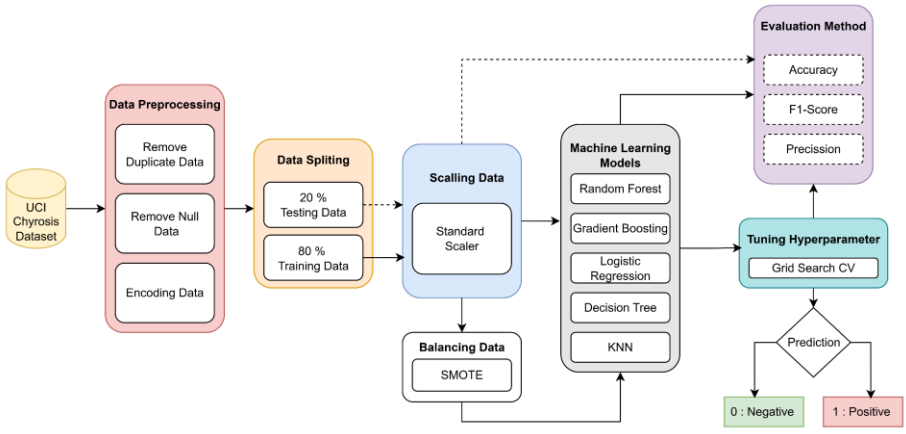

Figure 1 Methodological Workflows

2.1 Dataset

This research utilizes the Thyroid Disease Dataset from the UCI Machine Learning Repository the Thyroid Disease Dataset [25]. This dataset has been used in many previous research studies, particularly those focusing on the classification of thyroid disease. It contains the medical information of 364 patients, with 17 attributes reflecting their clinical condition related to thyroid disorders.

The dataset comprises both numerical and categorical attributes. One of the main numerical attributes is age, ranging from 15 and 82 years. The categorical attributes encompass patient demographic and medical history information, such as gender (M for male, F for female), smoking status (yes or no), and smoking history. A detailed summary of all attributes, including their respective data types and values, is presented in Table 1.

Table 1 Dataset Specification

No	Column Name	Data Type	Possible Value
1	Age	Integer	15-82
2	Gender	Categorical	M , F
3	Smoking	Categorical	No , Yes
4	Hx Smoking	Categorical	No , Yes
…	…	…	…
17	Reccured	Categorical	No , Yes

The main target variable for classification is the 'Recurred' column, which indicates whether or not the patient has experienced a recurrence of thyroid disease. This attribute is binary coded, where ‘Yes’ indicates a recurrence of thyroid disease and ‘No’ indicates no recurrence. In addition to the attributes shown in the above table, the dataset includes laboratory information relating to thyroid hormones, such as TSH, T3 and T4. As changes in these hormone levels are often early indicators of thyroid dysfunction, they are critical predictors in the classification process.

2.2 Preprocessing Data

Data preprocessing was performed to ensure data integrity prior to training the machine learning model. Duplicate data were removed using the *drop_duplicates()* function, with the objective of averting potential bias. Rows with missing values were removed using the dropna() function, thereby ensuring the dataset's integrity and completeness. Subsequently, the categorical features were converted to numeric values using the LabelEncoder from the *scikit-learn* library to ensure compatibility with the machine learning algorithms. These steps are critical to guarantee that the data utilized possesses a consistent and valid structure during the development of the classification model [26].

2.3 Data Splitting

Subsequent to the completion of the data cleaning and transformation process, the subsequent step is to divide the dataset into training data and testing data. The objective of this division is to distinguish between data utilized for the construction of models and data employed for the objective evaluation of model performance [27]. In this research, the dataset is divided into two parts using an 80/20 ratio, where 80% of the data is used for training and 20% for testing. The division was performed using the *train_test_split* function from the *scikit-learn* library. To maintain equilibrium in the

proportion of the target class distribution between the training and testing data, the data is divided through the implementation of a stratification technique based on the target label.

2.4 Feature Normalization

To ensure that all numerical features are on a comparable scale, a normalization process is performed using the standardization method. This technique is implemented using StandardScaler from the scikit-learn library, which scales each feature to have a mean of zero and a standard deviation of one [28]. Mathematically, this standardization process can be expressed as (1).

$$z = \frac{x - \mu}{\sigma} \tag{1}$$

In this context, x denotes the original value of a feature, μ signifies the mean value of the feature, and σ represents the standard deviation. The result of this transformation is the z value, which is a standardized score that represents the extent to which the x value deviates from the mean in standard deviation units.

2.5 Synthetic Minority Over-Sampling Technique (SMOTE)

The class imbalance in the target variable was addressed using SMOTE. SMOTE is an oversampling technique that works by generating new synthetic samples in the minority class through interpolation between existing data points, thereby ensuring a more balanced class distribution [29]. In this research, SMOTE was applied only to the training data, while the test set remained composed solely of original data. This strategy is essential to prevent data leakage, ensuring that the model is evaluated only on authentic, unseen data. By applying SMOTE only during training, the model learns to better recognize the minority class without inflating performance metrics on the test set.

2.6 Data Modelling

In this research, five distinct machine learning algorithms were employed to construct classification models aimed at predicting the recurrence status of thyroid disease. The selection of models was based on the diversity of algorithmic approaches, ranging from linear models to tree-based and ensemble methods. The applied algorithms include logistic regression, a basic linear model that is efficient for binary data, as well as KNN, which operates based on the proximity of the distance between samples [30]. Furthermore, a Decision Tree is employed as a tree-based model that can explicitly capture nonlinear relationships [31]. Finally, to assess the performance of more complex models known for their high accuracy and robustness to overfitting, two ensemble methods, namely Random Forest and Gradient Boosting, were implemented [21]. All models were trained using normalized and balanced data to ensure that the classification performance reflects the generalization ability of the patterns in the data.

2.7 Model Evaluation

The evaluation of model performance was conducted using four primary metrics: accuracy, precision, recall, and F1-score. They were chosen because they provide a comprehensive assessment of the performance of the classification model, particularly for imbalanced datasets [32].

Accuracy is a metric that quantifies the proportion of correct predictions to the total number of predictions made. Mathematically, accuracy is expressed by (2).

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{2}$$

TP (true positive) denotes the number of positive cases that were accurately predicted; TN (true negative) signifies the number of negative cases that were correctly predicted, FP (false positive) represents the number of negative cases that were erroneously predicted as positive, and FN (false negative) indicates the number of positive cases that were inaccurately predicted as negative.

Precision is employed to evaluate the model's accuracy in predicting positive classes, defined as the proportion of accurate positive predictions out of all positive predictions made. The precise formula is calculated by (3).

$$\text{Precision} = \frac{TP}{TP + FP} \tag{3}$$

Concurrently, recall measures the sensitivity of the model or its capacity to identify all true positive cases. This is expressed by (4).

$$\text{Recall} = \frac{TP}{TP + FN} \tag{4}$$

Given the inherent trade-off between precision and recall in information retrieval, a combined metric of F1-score is employed, representing the harmonic mean between precision and recall. In instances where the distribution of the class is found to be uneven, the F1-score provides a more balanced assessment. This calculation can be performed using (5).

$$\text{F1-score} = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \tag{5}$$

2.8 Hyperparameter Tuning

To attain optimal model performance, a hyperparameter was tuned using GridSearchCV. This approach aims to identify the optimal combination of parameters by exhaustively searching of the predefined parameter grid. The search process is executed with cross-validation, thereby ensuring that the tuning results enhance not only the accuracy of the model on the training data but also its capacity to generalize to data that has not been previously encountered [33]. The application of GridSearchCV facilitates the calibration of models to enhance efficiency and accuracy, while mitigating the risks of overfitting or underfitting.

3 Result And Discussion

This section presents the results of the data analysis and modeling, followed by an in-depth discussion of the findings. The analysis phase began with Exploratory Data Analysis (EDA) to investigate the characteristics and relationships between variables. Then, several machine learning algorithms were developed using classification models. The models are evaluated based on their performance before and after handling class imbalance using the SMOTE technique.

3.1 Feature Correlation Analysis

The analysis began with an exploratory phase to examine the dataset's fundamental characteristics, encompassing the structure, the quantity of features, and the patterns of relationships between variables. One method employed is the visualization of correlations between numeric features in the form of a correlation heatmap, as illustrated in Figure 2. This visualization offers a comprehensive overview of the degree of linear relationship between pairs of features.

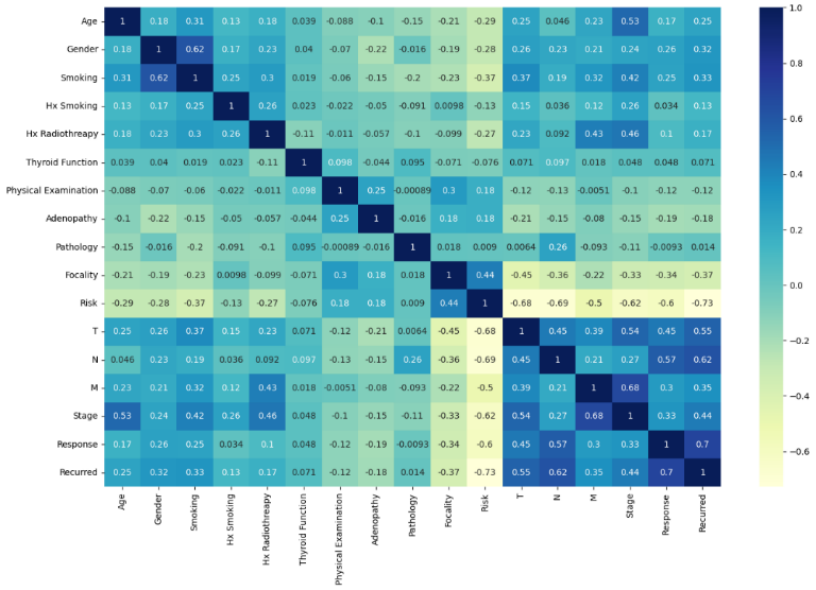

Figure 2 Correlation Feature

The heatmap revealed several highly correlated feature pairs. This finding suggests the possibility of multicollinearity, which can compromise the stability and interpretation of the model, particularly in linear models like logistic regression. Conversely, features that exhibit a strong correlation with the target variable are strong candidates for predictive modeling. Consequently, this correlation analysis is an initial basis for feature selection and understanding.

The heatmap analysis highlights potential multicollinearity, which can compromise the stability and interpretation of the model, particularly when linearity-sensitive algorithms such as logistic regression are employed. Strong positive correlations were notably observed between Stage and Age ($r = 0.53$), Stage and M ($r = 0.68$), T and N ($r = 0.45$), Response and N ($r = 0.57$), and Recurred and Response ($r = 0.70$). These pairs indicate closely related variables that may convey redundant information. In terms of their relevance to the prediction task, the 'Recurred' label showed strong correlations with 'Risk' ($r = –0.73$), 'T' ($r = 0.55$), 'N' ($r = 0.62$), and 'Response' ($r = 0.70$), suggesting that these variables play a significant role in determining recurrence. Conversely, several attributes exhibited weak or negligible correlation with the target variable, including Pathology ($r = 0.014$), Adenopathy ($r = −0.18$), and Physical Examination ($r = −0.12$). These uncorrelated features may contribute little to model performance and could potentially be excluded during feature selection to improve efficiency and reduce noise.

3.2 Result of Baseline Model

Next, the analysis focused on the class distribution of the target variable. As illustrated in Figure 3, there is a significant imbalance between the majority and minority classes. This imbalance is a

common challenge in classification problems, as models tend to adopt a cautious approach by predicting the majority class, which is statistically more dominant. Consequently, models frequently fail to identify patterns within the minority class, which is frequently the class of interest. This challenge is common in real-world applications, such as rare disease detection or fraud identification.

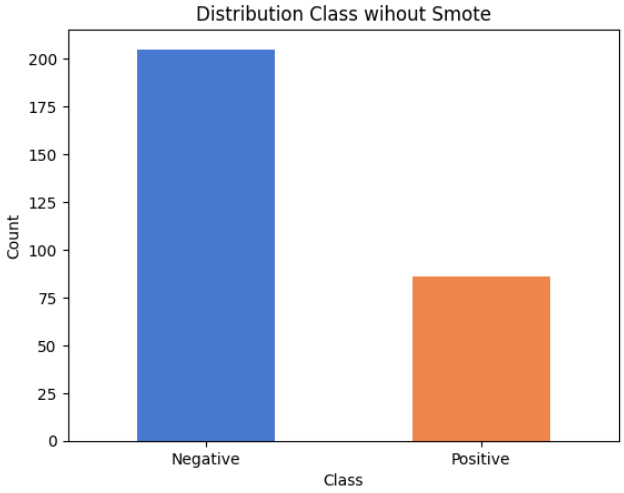

Figure 3 Class distribution of the baseline model

This imbalance can lead to misleading performance metrics. For example, a model may achieve high overall accuracy simply by predicting the majority class, while its recall for the minority class remains extremely poor. This necessitates specialized strategies to address the imbalance, such as the SMOTE oversampling technique, which is detailed in the following section.

Table 2 Baseline model performance (without SMOTE)

Model	Accuracy
Random Forest	96 %
Logistic Regression	85 %
Decision Tree	92 %
Gradient Boosting	93 %
KNN	82 %

To establish a baseline, an initial evaluation was conducted on the original, imbalanced dataset without any hyperparameter tuning. Five classification models were tested under these conditions to measure their out-of-the-box performance. The algorithms evaluated were Random Forest, Logistic Regression, Decision Tree, Gradient Boosting, and KNN. All models were executed using default parameters to establish an initial baseline before optimization.

Table 2 presents the accuracy results of each model when trained on the original data with default parameters. The Random Forest algorithm achieved the highest performance, with an accuracy of 96%. This was followed by the Gradient Boosting algorithm, which achieved an accuracy of 93%, and the Decision Tree algorithm, which achieved an accuracy of 92%. The two models under consideration, logistic regression and KNN, achieved lower accuracies of 85% and 82%, respectively. Random Forests and Gradient Boosting tend to perform better than linear or distance-based models. This phenomenon could be attributed to the ensemble's capacity to manage data complexity and variation effectively. However, it should be noted that these results remain provisional, as they have not yet accounted for factors such as parameter optimization or class imbalance management.

Table 3 Result of Gridsearch CV on the baseline model

Model	Result Parameter Tuning	Accuracy
Random Forest	'bootstrap': False, 'max_depth': None, 'min_samples_leaf': 1, 'min_samples_split': 5, 'n_estimators': 50	96 %
Logistic Regression	'C': 0.1, 'max_iter': 1000, 'solver': 'lbfgs'	84 %
Decision Tree	'criterion': 'entropy', 'max_depth': None, 'min_samples_leaf': 2, 'min_samples_split': 10	96 %
Gradient Boosting	'learning_rate': 0.1, 'max_depth': 3, 'n_estimators': 50	93 %
KNN	'metric': 'manhattan', 'n_neighbors': 15, 'p': 1, 'weights': 'distance'	90 %

To enhance performance, a hyperparameter tuning procedure was performed on all five models using GridSearchCV with cross-validation. As illustrated in Table 3, the optimal parameters for each model are enumerated, along with the accuracy achieved after tuning on the original dataset. The tuning results demonstrate substantial performance improvements across several models, particularly KNN, which increased accuracy from 82% to 90%. Decision Tree also achieved 96% accuracy, a result that aligns with Random Forest, which demonstrated consistent accuracy. Logistic regression showed a modest decline in accuracy to 84%, while gradient boosting maintained its superior performance at 93%. The table provides details on optimized parameters, such as 'n_estimators' and 'max_depth' for tree-based models, 'C' for Logistic Regression, and 'n_neighbors' and 'metric' for KNN.

The accuracy comparison after tuning is shown in Figure 4. The results clearly show that the tree-based models, Random Forest and Decision Tree, achieved the highest performance on the original, imbalanced dataset. This finding highlights the significant impact of parameter optimization, which is crucial for maximizing a model's predictive capabilities and minimizing the risks of overfitting or underfitting.

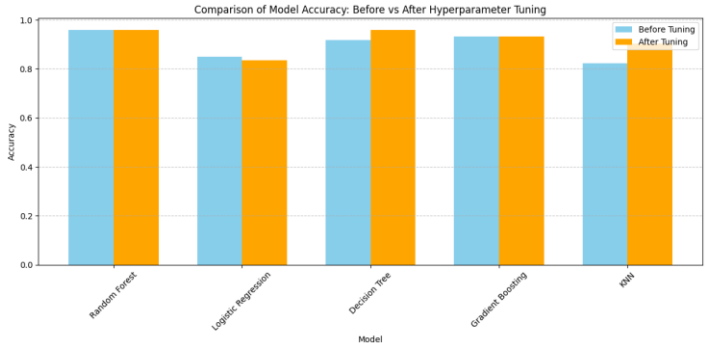

Figure 4 Comparison of the Accuracy of the baseline model before and after Hyperparameter Tuning

To better understand the model's predictive performance, the confusion matrices and classification reports for the two top-performing models, Random Forest and Decision Tree, were analyzed. Although both models achieved an identical accuracy of 96%, this metric alone can be misleading on imbalanced datasets. Therefore, it was essential to evaluate their effectiveness in identifying the minority class by examining additional metrics, namely precision, recall, and the F1-score.

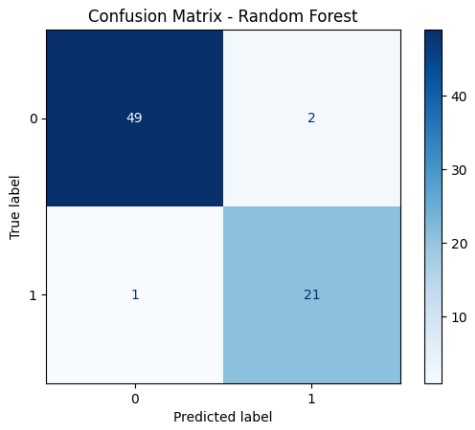

Figure 5 Confusion Matrix of Random Forest model and hyperparameter tuning on baseline model

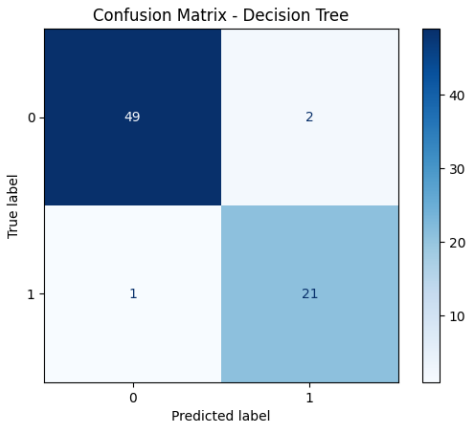

Figure 6 Confusion Matrix of the Decision Tree model and hyperparameter tuning on the baseline model

The confusion matrices for the Random Forest and Decision Tree models are presented in Figures 5 and 6, respectively. These matrices display the distribution of correct and incorrect predictions for each class. This information is crucial for identifying the types of errors that the model frequently makes, such as the tendency to misclassify minority classes as majority high False Negative Rate.

Classification Report: Random Forest

	precision	recall	f1-score	support
0	0.9800	0.9608	0.9703	51
1	0.9130	0.9545	0.9333	22
accuracy			0.9589	73
macro avg	0.9465	0.9577	0.9518	73
weighted avg	0.9598	0.9589	0.9592	73

Figure 7 Classification Report of the Random Forest model and hyperparameter tuning on the baseline model

Classification Report: Decision Tree

	precision	recall	f1-score	support
0	0.9800	0.9608	0.9703	51
1	0.9130	0.9545	0.9333	22
accuracy			0.9589	73
macro avg	0.9465	0.9577	0.9518	73
weighted avg	0.9598	0.9589	0.9592	73

Figure 8 Classification Report of the Decision Tree model and hyperparameter tuning on the baseline model

Additionally, the classification report in Figures 7 and 8 provides a more detailed evaluation. These reports break down the precision, recall, and F1-score for each class. In the context of imbalanced data, the recall for minority classes serves as a critical performance metric. The evaluation results indicate that, while the overall accuracy is high, the recall for minority classes can be further enhanced.

3.3 Result of SMOTE SMOTE-based Model

Figure 9 illustrates the effect of applying the SMOTE technique to the training data. It is evident that the proportion of minority class samples has increased, resulting in a balanced distribution between the two classes. rebalancing is expected to enhance the model's ability to learn the patterns of the previously underrepresented minority class, thereby improving its predictive performance.

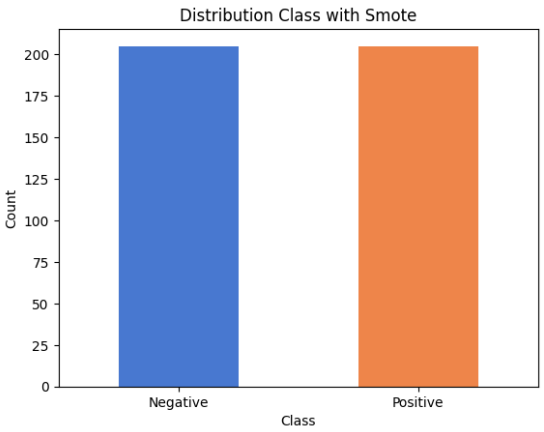

Figure 9 Distribution class of training data on SMOTE based model

After rebalancing the training data with SMOTE, the five models were re-evaluated using their original default hyperparameters. This step was performed to isolate the impact of class balancing on model performance, independent of hyperparameter tuning. The resulting accuracies for each model are presented in Table 3.

Table 4 Testing result on SMOTE based model

Model	Accuracy
Random Forest	95 %
Logistic Regression	84 %
Decision Tree	92 %
Gradient Boosting	95 %
KNN	85 %

After retraining on the SMOTE-balanced data with default parameters, the ensemble models maintained strong performance. Random Forest and Gradient Boosting both achieved 95% accuracy. In contrast, the single Decision Tree's accuracy decreased from its previous high of 96% to 92%, while Logistic Regression and KNN showed marginal improvements but still lagged behind the ensemble methods. These results suggest that while ensembles are robust, the performance of simpler models like Decision Tree can be sensitive to the synthetic data introduced by SMOTE. The application of SMOTE fundamentally alters the distribution and characteristics of the training data. Consequently, the hyperparameters that were optimal for the original, imbalanced dataset may no longer be effective for the new, balanced dataset. Therefore, to ensure each model is fully optimized for the new data structure, a final round of hyperparameter tuning is required. This re-tuning process is detailed in the subsequent section.

Table 5 presents the final performance of each model after hyperparameter tuning on the SMOTE-balanced dataset. The Random Forest and Gradient Boosting algorithms continue to demonstrate high performance, with both achieving 95% accuracy. The KNN model demonstrates enhanced performance following tuning, attaining 88% accuracy. Conversely, Decision Tree demonstrated a decline in performance, from 96% (without SMOTE) to 92%. Logistic Regression remained stable, showing only a marginal increase to 85%.

Table 5 Gridsearch CV result on SMOTE SMOTE-based model

Model	Result Parameter Tuning	Accuracy
Random Forest	'bootstrap': True, 'max_depth': None, 'min_samples_leaf': 1, 'min_samples_split': 2, 'n_estimators': 100	95 %
Logistic Regression	'C': 0.1, 'max_iter': 1000, 'solver': 'lbfgs'	85 %
Decision Tree	'criterion': gini, 'max_depth': None, 'min_samples_leaf': 1, 'min_samples_split': 2	92 %
Gradient Boosting	'learning_rate': 0.1, 'max_depth': 3, 'n_estimators': 100	95 %
KNN	'metric': 'manhattan', 'n_neighbors': 3, 'p': 1, 'weights': 'distance'	88 %

The efficacy of the SMOTE and tuning algorithms can be observed in Figure 10. Overall, this two-step optimization process successfully maintained the high accuracy of the ensemble models while significantly improving the performance of the KNN model.

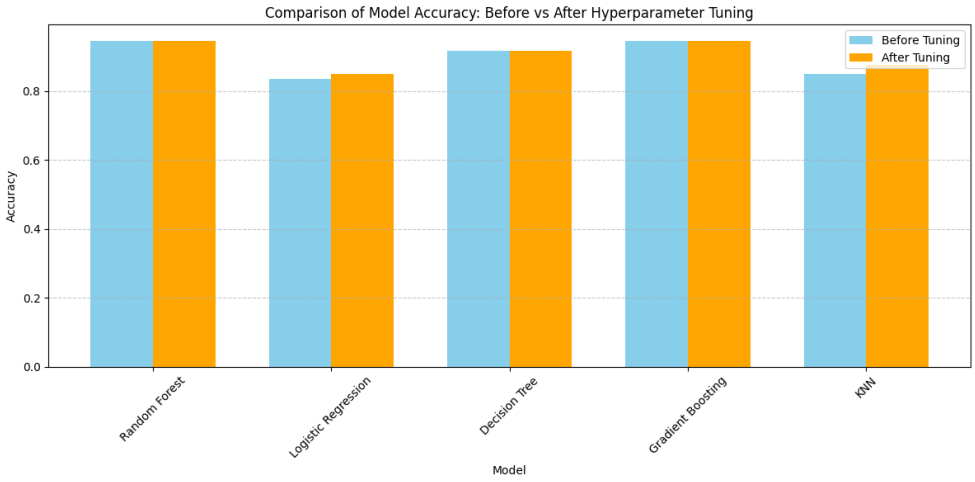

Figure 10 Comparison result of before and after tuning on SMOTE based model

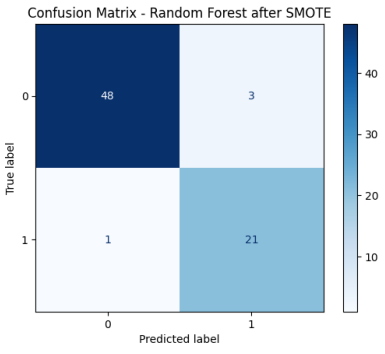

Figure 11 Confusion Matrix on Random Forest model and hyperparameter tuning on SMOTE-based model

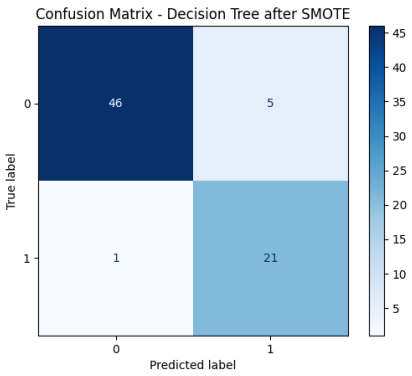

Figure 12 Confusion Matrix on the Decision Tree model and hyperparameter tuning on SMOTE-based model

The confusion matrices for Random Forest (Figure 11) and Decision Tree (Figure 12) show that both models perform well in classifying both classes. However, a key difference emerges in their misclassification patterns, particularly in their ability to correctly predict the minority class.

The classification reports (Figures 11 and 12) detail the precision, recall, and F1-score for each model. This finding suggests that SMOTE effectively enhanced the models' ability to identify the minority class while maintaining strong performance on the majority class.

Classification Report: Random Forest after SMOTE

	precision	recall	f1-score	support
0	0.9796	0.9412	0.9600	51
1	0.8750	0.9545	0.9130	22
accuracy			0.9452	73
macro avg	0.9273	0.9479	0.9365	73
weighted avg	0.9481	0.9452	0.9458	73

Figure 13 Classification Report on Random Forest model and hyperparameter tuning on SMOTE-based model

Classification Report: Decision Tree after SMOTE

	precision	recall	f1-score	support
0	0.9787	0.9020	0.9388	51
1	0.8077	0.9545	0.8750	22
accuracy			0.9178	73
macro avg	0.8932	0.9283	0.9069	73
weighted avg	0.9272	0.9178	0.9196	73

Figure 14 Classification Report on Decision Tree model and hyperparameter tuning on SMOTE-based model

The classification reports for each model in Figures 13 and 14 provide a comprehensive account of precision, recall, and F1 scores. The reports reveal a substantial increase in recall for the minority class compared to the pre-SMOTE results. This finding suggests that SMOTE effectively enhanced the models' ability to identify the minority class while maintaining strong performance on the majority class.

Despite a slight decrease in overall decision tree accuracy, the increase in minority class recall indicates that the trade-off improves the class-balanced prediction. This outcome aligns perfectly with the primary goal of SMOTE, which is to enhance the representation and detection of infrequent classes.

Table 6 The accuracy results of tuning before and after SMOTE

Model	Accuracy Tuning	Accuracy Tuning + Smote
Random Forest	**96 %**	**95 %**
Logistic Regression	84 %	85 %
Decision Tree	96 %	92 %
Gradient Boosting	93 %	95 %
KNN	90 %	88 %

To provide a comprehensive overview, a comparative analysis was conducted between the pre- and post-SMOTE conditions, utilizing both the tuned and untuned models. The summary of these results is presented in Table 6.

Figure 15 shows the changes in accuracy across all conditions. The findings indicate that while certain models may exhibit a modest decline in accuracy following SMOTE, this is an acceptable trade-off for the crucial gain in recall for the minority class. In both scenarios (with and without SMOTE), Random Forest and Gradient Boosting consistently performed best.

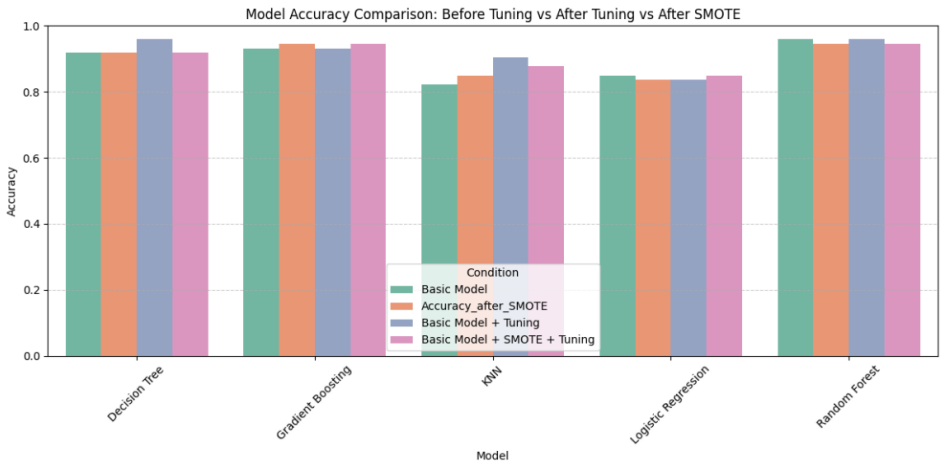

Figure 15 Overall accuracy comparison

3.4 Discussion

Although some models, such as Logistic Regression, demonstrated a slight decline in accuracy following the implementation of SMOTE, this is an acceptable trade-off in a clinical context. In clinical prediction tasks, the primary concern is recall, since missing a recurrence case (false negative) could delay intervention and lead to severe consequences. Therefore, an improvement in recall for the minority class justifies the minor reduction in overall accuracy, highlighting the importance of prioritizing sensitivity over general performance metrics in medical applications.

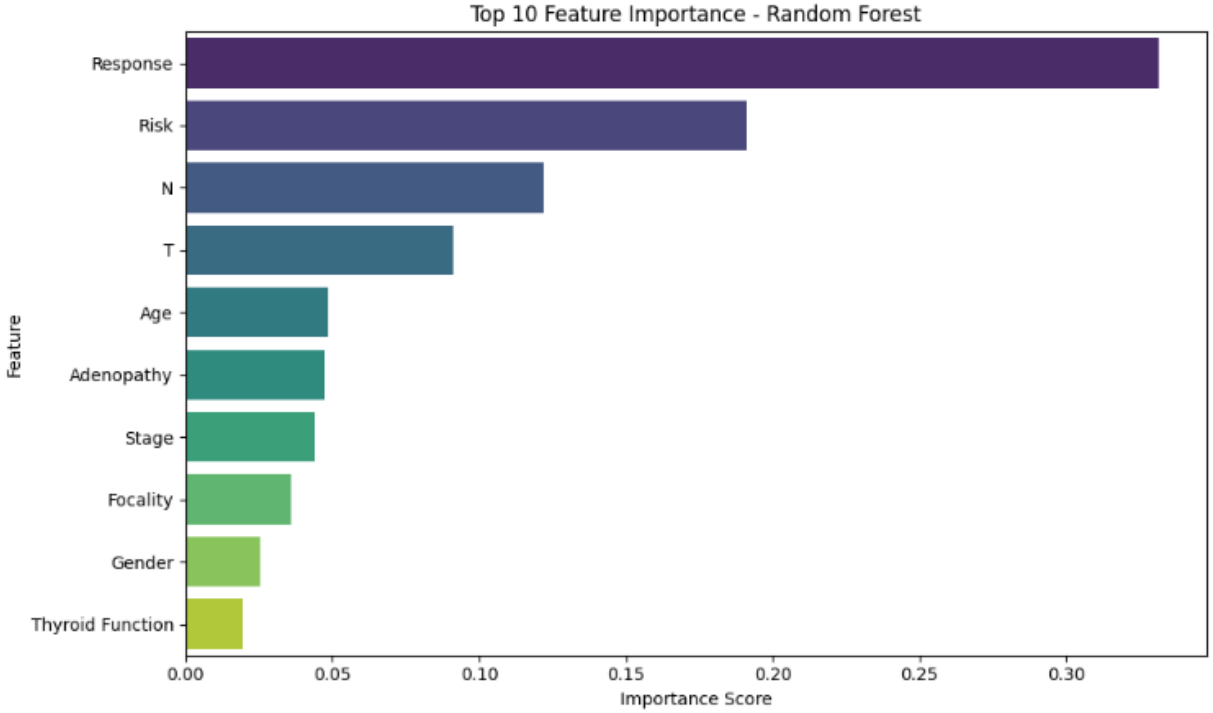

Figure 16 Top 10 Feature importance of the Random Forest model

The Random Forest and Gradient Boosting models consistently demonstrated superior performance across all scenarios. These algorithms not only achieved high accuracy on the original dataset but also remained stable after SMOTE and hyperparameter tuning, confirming their robustness and generalizability. Their ability to capture complex patterns, including subtle variations in minority class data, reinforces their suitability for supporting clinical decision-making systems.

From a practical standpoint, integrating such models into clinical workflows could facilitate early screening and risk stratification. For example, patients predicted as high risk for recurrence could be prioritized for follow-up imaging or laboratory testing, even in hospitals with limited specialist availability. Moreover, since the models provide interpretable outputs such as feature importance rankings, clinicians could better understand which variables, such as TSH, T3, and T4 levels, most influence recurrence predictions. This interpretability increases the likelihood of adoption in clinical environments, where transparency and explainability are critical for trust and confidence accountability.

To enhance model interpretability, feature importance analysis was conducted using the top-performing Random Forest algorithm. As illustrated in Figure 16, the sequence of features is determined by their impact on model decision-making.

The feature with the highest importance value is regarded as the leading indicator in classification. This analysis facilitates understanding of the model's internal processes or model transparency and provides practical insights applicable to real decision-making scenarios, such as feature selection or expert system development.

4 Conclusion

This research highlights the effectiveness of machine learning, particularly ensemble models, in predicting thyroid disorder recurrence. It demonstrates that the predictive performance of imbalanced medical datasets can be substantially enhanced by adopting a structured experimental approach that incorporates SMOTE-based resampling to address class imbalance and GridSearchCV for hyperparameter optimisation. Of the five algorithms evaluated, Logistic Regression, K-Nearest Neighbours, Decision Tree, Random Forest, and Gradient Boosting, the ensemble methods performed consistently better than the others, achieving accuracy rates above 93% across different conditions. These models also demonstrated superior sensitivity to the minority class, a critical factor in medical applications. Furthermore, feature importance analysis revealed which clinical attributes contributed most to prediction outcomes, thereby supporting both model interpretability and clinical relevance.

The findings, however, are based on a single publicly available dataset, which may limit the generalizability of the results to different patient populations or clinical settings. The retrospective nature of the dataset also means that potential confounding variables or unmeasured risk factors were not included. In addition, although SMOTE improved class balance, synthetic oversampling may introduce artifacts or patterns that are not representative of the true patient population.

To enhance the generalizability of these findings, future research should use larger, multi-center datasets that more accurately reflect diverse patient demographics and clinical conditions. Additionally, exploring deep learning or hybrid approaches that combine multimodal data, such as structured data with imaging or genomic results, could lead to further improvements in predictive performance accuracy. From a clinical perspective, the proposed models hold promise for integration into decision-support systems, enabling earlier identification of high-risk patients and supporting follow-up protocols more efficiently, especially in hospitals with limited resources.

Bibliography

[1] H. Risdianti, Y. Sofiani, and T. Muhaemin, “Sleep Quality and Quality of Life of Hyperthyroidism Patients at Bhayangkara Hospital Level I R. Said Sukanto Kramat Jati, East Jakarta,” *Jurnal Penelitian Kesehatan “SUARA FORIKES” (Journal of Health Research “Forikes Voice”)*, vol. 12, no. 0, Art. no. 0, Apr. 2021, doi: 10.33846/sf12nk212.

[2] R. C. Kusumadewi *et al.*, “Understanding Thyroid Carcinoma: Clinical Manifestations, Diagnosis, and Management,” *JBT*, vol. 24, no. 4, pp. 645–656, Oct. 2024, doi: 10.29303/jbt.v24i4.7664.

[3] N. K. M. Suryantini, L. L. Putri, B. H. Salim, A. Mawaddah, and E. Triani, “Thyroid Hormone Disorders: Hypothyroidism,” *J.MedHealth*, vol. 11, no. 6, pp. 1227–1234, June 2024, doi: 10.33024/jikk.v11i6.14697.

[4] S. Park *et al.*, “Dynamic Risk Stratification for Predicting Recurrence in Patients with Differentiated Thyroid Cancer Treated Without Radioactive Iodine Remnant Ablation Therapy,” *Thyroid*, vol. 27, no. 4, pp. 524–530, Apr. 2017, doi: 10.1089/thy.2016.0477.

[5] G. A. Arie, S. D. Santoso, and R. I. Santosa, "The Relationship Between Thyroid Function Disorders and LDL-Cholesterol Levels," *JSH*, vol. 5, no. 2, pp. 6–12, Oct. 2021, doi: 10.51804/jsh.v5i2.1018.6-12.
[6] M. A. Arosemena, N. A. Cipriani, and A. M. Dumitrescu, "Graves' disease and papillary thyroid carcinoma: case report and literature review of a single academic center," *BMC Endocr Disord*, vol. 22, no. 1, p. 199, Aug. 2022, doi: 10.1186/s12902-022-01116-1.
[7] M. Huang, S. Yang, G. Ge, H. Zhi, and L. Wang, "Effects of Thyroid Dysfunction and the Thyroid-Stimulating Hormone Levels on the Risk of Atrial Fibrillation: A Systematic Review and Dose-Response Meta-Analysis from Cohort Studies," *Endocrine Practice*, vol. 28, no. 8, pp. 822–831, Aug. 2022, doi: 10.1016/j.eprac.2022.05.008.
[8] S. Zhang *et al.*, "High level of thyroid peroxidase antibodies as a detrimental risk of pregnancy outcomes in euthyroid women undergoing ART: A meta-analysis," *Molecular Reproduction Devel*, vol. 90, no. 4, pp. 218–226, Apr. 2023, doi: 10.1002/mrd.23677.
[9] S. L. Andersen and S. Andersen, "Turning to Thyroid Disease in Pregnant Women," *Eur Thyroid J*, vol. 9, no. 5, pp. 225–233, 2020, doi: 10.1159/000506228.
[10] E. Beka and O. Gimm, "Voice Changes Without Laryngeal Nerve Alterations After Thyroidectomy: The Need For Prospective Trials - A Review Study," *Journal of Voice*, vol. 38, no. 1, pp. 231–238, Jan. 2024, doi: 10.1016/j.jvoice.2021.07.012.
[11] G. Grani and A. Fumarola, "Thyroglobulin in Lymph Node Fine-Needle Aspiration Washout: A Systematic Review and Meta-analysis of Diagnostic Accuracy," *The Journal of Clinical Endocrinology & Metabolism*, vol. 99, no. 6, pp. 1970–1982, June 2014, doi: 10.1210/jc.2014-1098.
[12] Y. Zhu, Y. Song, G. Xu, Z. Fan, and W. Ren, "Causes of misdiagnoses by thyroid fine-needle aspiration cytology (FNAC): our experience and a systematic review," *Diagn Pathol*, vol. 15, no. 1, p. 1, Dec. 2020, doi: 10.1186/s13000-019-0924-z.
[13] K. Y. Na, H.-S. Kim, J.-Y. Sung, W. S. Park, and Y. W. Kim, "Papillary Carcinoma of the Thyroid Gland with Nodular Fasciitis-like Stroma," *Korean J Pathol*, vol. 47, no. 2, p. 167, 2013, doi: 10.4132/KoreanJPathol.2013.47.2.167.
[14] M. N. Nikiforova *et al.*, "Analytical performance of the ThyroSeq v3 genomic classifier for cancer diagnosis in thyroid nodules," *Cancer*, vol. 124, no. 8, pp. 1682–1690, Apr. 2018, doi: 10.1002/cncr.31245.
[15] R. G. Wardhana, G. Wang, and F. Sibuea, "Application of Machine Learning in Predicting Disease Case Levels in Indonesia," *JOISM*, vol. 5, no. 1, pp. 40–45, July 2023, doi: 10.24076/joism.2023v5i1.1136.
[16] A. R. Pratama, F. Wabula, H. Ilmandry, M. L. Isabela, M. Raharjo, and R. Sianipar, "Literature Review The Impact of Machine Learning in Modern Industries," *NianTanaSikka*, vol. 3, no. 1, pp. 177–182, Jan. 2025, doi: 10.59603/niantanasikka.v3i1.680.
[17] T. Wongvorachan, S. He, and O. Bulut, "A Comparison of Undersampling, Oversampling, and SMOTE Methods for Dealing with Imbalanced Classification in Educational Data Mining," *Information*, vol. 14, no. 1, Art. no. 1, Jan. 2023, doi: 10.3390/info14010054.
[18] S. Shekhar, A. Bansode, and A. Salim, "A Comparative study of Hyper-Parameter Optimization Tools," in *2021 IEEE Asia-Pacific Conference on Computer Science and Data Engineering (CSDE)*, Dec. 2021, pp. 1–6. doi: 10.1109/CSDE53843.2021.9718485.
[19] I. H. Sarker, "Machine Learning: Algorithms, Real-World Applications and Research Directions," *SN COMPUT. SCI.*, vol. 2, no. 3, p. 160, May 2021, doi: 10.1007/s42979-021-00592-x.
[20] S. Uddin, I. Haque, H. Lu, M. A. Moni, and E. Gide, "Comparative performance analysis of K-nearest neighbour (KNN) algorithm and its different variants for disease prediction," *Sci Rep*, vol. 12, no. 1, p. 6256, Apr. 2022, doi: 10.1038/s41598-022-10358-x.

[21] T. Kavzoglu and A. Teke, "Predictive Performances of Ensemble Machine Learning Algorithms in Landslide Susceptibility Mapping Using Random Forest, Extreme Gradient Boosting (XGBoost) and Natural Gradient Boosting (NGBoost)," *Arab J Sci Eng*, vol. 47, no. 6, pp. 7367–7385, June 2022, doi: 10.1007/s13369-022-06560-8.
[22] E. Clark, S. Price, T. Lucena, B. Haberlein, A. Wahbeh, and R. Seetan, "Predictive Analytics for Thyroid Cancer Recurrence: A Machine Learning Approach," *Knowledge*, vol. 4, no. 4, pp. 557–570, Nov. 2024, doi: 10.3390/knowledge4040029.
[23] F. Firat Atay *et al.*, "A hybrid machine learning model combining association rule mining and classification algorithms to predict differentiated thyroid cancer recurrence," *Front. Med.*, vol. 11, p. 1461372, Oct. 2024, doi: 10.3389/fmed.2024.1461372.
[24] R. H. Agarwal, S. Degadwala, and D. Vyas, "Predictive Modeling for Thyroid Disease Diagnosis using Machine Learning," in *2024 International Conference on Inventive Computation Technologies (ICICT)*, Lalitpur, Nepal: IEEE, Apr. 2024, pp. 227–231. doi: 10.1109/ICICT60155.2024.10544462.
[25] A. T. Shiva Borzooei, "Differentiated Thyroid Cancer Recurrence." UCI Machine Learning Repository, 2023. doi: 10.24432/C5632J.
[26] H. M. Marin-Castro and E. Tello-Leal, "Event Log Preprocessing for Process Mining: A Review," *Applied Sciences*, vol. 11, no. 22, Art. no. 22, Jan. 2021, doi: 10.3390/app112210556.
[27] V. R. Joseph and A. Vakayil, "Split: An Optimal Method for Data Splitting," *Technometrics*, vol. 64, no. 2, pp. 166–176, Apr. 2022, doi: 10.1080/00401706.2021.1921037.
[28] M. M. Ahsan, M. A. P. Mahmud, P. K. Saha, K. D. Gupta, and Z. Siddique, "Effect of Data Scaling Methods on Machine Learning Algorithms and Model Performance," *Technologies*, vol. 9, no. 3, Art. no. 3, Sept. 2021, doi: 10.3390/technologies9030052.
[29] G. A. Pradipta, R. Wardoyo, A. Musdholifah, I. N. H. Sanjaya, and M. Ismail, "SMOTE for Handling Imbalanced Data Problem : A Review," in *2021 Sixth International Conference on Informatics and Computing (ICIC)*, Nov. 2021, pp. 1–8. doi: 10.1109/ICIC54025.2021.9632912.
[30] P. Cunningham and S. J. Delany, "k-Nearest Neighbour Classifiers - A Tutorial," *ACM Comput. Surv.*, vol. 54, no. 6, pp. 1–25, July 2022, doi: 10.1145/3459665.
[31] Z. Azam, Md. M. Islam, and M. N. Huda, "Comparative Analysis of Intrusion Detection Systems and Machine Learning-Based Model Analysis Through Decision Tree," *IEEE Access*, vol. 11, pp. 80348–80391, 2023, doi: 10.1109/ACCESS.2023.3296444.
[32] M. Owusu-Adjei, J. Ben Hayfron-Acquah, T. Frimpong, and G. Abdul-Salaam, "Imbalanced class distribution and performance evaluation metrics: A systematic review of prediction accuracy for determining model performance in healthcare systems," *PLOS Digit Health*, vol. 2, no. 11, p. e0000290, Nov. 2023, doi: 10.1371/journal.pdig.0000290.
[33] P. G. Brindha, R. Boobesh, and S. S. Yokanandh, "Optimization of ML Algorithms in CAD Diagnosis Using Grid Search CV," in *2025 3rd International Conference on Intelligent Data Communication Technologies and Internet of Things (IDCIoT)*, Feb. 2025, pp. 2229–2234. doi: 10.1109/IDCIOT64235.2025.10914959.