**Jurnal Elektronika dan Telekomunikasi (JET)**, Vol. 25, No. 2, December 2025, pp. 93-101
Accredited by RISTEKDIKTI, Decree No: **158/E/KPT/2021**
doi: 10.55981/jet.758

# Feature Selection and Class Imbalance Machine Learning for Early Detection of Thyroid Cancer Recurrence: A Performance-Based Analysis

**Agus Wantoro[a*], Wahyu Caesarendra[b], Admi Syarif[c], Hari Soetanto[d]**

[a]*Department of Technology and Informatics,*
*Universitas Aisyah Pringsewu*
*Street A Yani Number 1 A Tambak Rejo*
*Pringsewu, Indonesia*
[b]*Department of Mechanical and Mechatronics Engineering, Malaysia*
*Curtin University Malaysia*
*Lot 13149, Land District, Block 5, CDT 250, Kuala Baram*
*Sarawak, Malaysia*
[c]*Department of Computer Sciences,*
*Universitas Lampung*
*Street Soemantri Brojonegoro Number 1 A Rajabasa*
*Bandar Lampung, Indonesia*
[d]*Department of Technology Information,*
*Universitas Budi Luhur*
*Street Ciledug Raya, RT.10, RW.2 Petukangan Utara*
*Jakarta, Indonesia*

**Abstract**

Early detection of thyroid cancer recurrence is a crucial factor in patient survival and treatment effectiveness. Misdetection results in disease severity, high cost, recovery time, and decreased service quality. In addition, the main challenges in developing a Machine Learning (ML)-based detection decision support system are class imbalance in medical data and high feature dimensions that can affect model accuracy and efficiency. This study proposes a feature selection-based approach and class imbalance handling to improve the performance of early detection of Thyroid cancer. Several feature selection techniques, such as Information Gain (IG), Gain Ratio (GR), Gini Decrease (GD), and Chi-Square (CS), can select features based on weighted ranking. In addition, to overcome the imbalanced class distribution, we use the Synthetic Minority Over-Sampling Technique (SMOTE). ML classification models such as k-NN, Tree, SVM, Naive Bayes, AdaBoost, Neural Network (NN), and Logistic Regression (LR) are tested and evaluated based on a confusion matrix, including accuracy, precision, recall, time, and log loss. Experimental results show that the combination of imbalanced class handling strategies significantly improves the prediction performance of ML algorithms. In addition, we found that the combination of CS+NN feature selection techniques consistently showed optimal performance. This study emphasizes the importance of data pre-processing and proper algorithm selection in the development of a machine learning-based thyroid cancer detection system.

**Keywords:** Class imbalance, Feature selection, Machine Learning, Thyroid cancer.

## I. INTRODUCTION

Thyroid cancer is a serious global health problem, accounting for the majority of cancer cases and cancer-related deaths worldwide [1]. Thyroid cancer is one of the leading causes of death due to complications [2]. Recent data show that the incidence of thyroid cancer continues to increase, making it the most frequently detected type of cancer. The main causes of death in thyroid cancer are complications of the cancer itself and the presence of other accompanying diseases, especially cardiovascular disease [3]. Early detection is a fundamental key to improving survival rates and treatment success. When detected at an early stage, the patient's chances of recovery are significantly higher compared to cases detected at an advanced stage [4].

In recent years, rapid advances in the field of Machine Learning (ML) have paved the way for the development of more accurate and efficient thyroid cancer early detection systems [5]. ML algorithms can analyze medical data, including genetic data, medical images, and clinical history, to identify complex patterns that may not be visible to the human eye. The potential of ML is to improve the accuracy of detection and predict disease risk in various studies [6].

However, the application of ML in early detection of thyroid cancer faces two main challenges, namely feature selection and class imbalance. The Thyroid Cancer Recurrence (TCR) Dataset is often used as a basis for evaluating early detection models for thyroid cancer [7].

* Corresponding Author.
Email: aguswantoro@aisyahuniversity.ac.id
Received: June 27, 2025 ; Revised: September 9, 2025
Accepted: October 16, 2025 ; Published: December 31, 2025

This dataset has 30 features, where all features need to be analyzed, whether they are relevant or not to the model-accuracy. The presence of irrelevant features can cause a decrease in model performance, increase computational complexity, and overfitting [8]. Therefore, an effective feature selection technique is essential to identify the most informative feature subset that has an impact on model performance and reduces noise [9].

In addition, the TCR Dataset has a class imbalance, where the number of recurred "Yes" classes is less than the "No" class. This imbalance causes the ML model to be biased towards the majority class, making it less accurate in predicting the minority class, which is the main target of early detection [10]. Various studies have highlighted the negative impact of class imbalance on ML performance and proposed methods to address it, such as oversampling approaches [8], [10].

To overcome the problem of class imbalance, one can use oversampling techniques such as SMOTE (Synthetic Minority Over-sampling Technique). This technique increases the representation of minority classes in the dataset, so that the model can learn better [11]. The integration of effective feature selection and class imbalance handling has been shown to improve the performance of ML models in thyroid cancer detection [12]. However, a comprehensive evaluation of the combination of different feature selection techniques and class imbalance handling methods is still needed to identify the best approach in this context.

This study aims to analyze the performance of feature selection methods and class imbalance handling techniques in the context of thyroid cancer recurrence detection using ML such as Neural Network (NN), Tree, Support Vector Machine (SVM), Naive Bayes, AdaBoost, Neural Network (NN), and Logistic Regression (LR). By comparing the performance of various combinations of these methods, it is expected that the most optimal approach can be found to build an accurate, sensitive, and specific early detection model for Thyroid cancer, thereby contributing to increasing patient survival.

## II. METHOD

This section describes the procedures and stages of the research. The research stage begins with data collection. Before the data is input into the model, data preprocessing needs to be done, such as cleaning data that does not have complete information. At this stage, the data is cleaned and adjusted for processing to the next step using class balance using SMOTE and feature selection using Information Gain (IG), Gain Ratio (GR), Gini Decrease (GD), and Chi-Square (CS). Furthermore, the selected features are used for Thyroid cancer classification using ML algorithms such as Tree, SVM, Naive Bayes, AdaBoost, Neural Network (NN), and Logistic Regression (LR). The design of the proposed system is illustrated in Figure 1.

After the thyroid cancer classification is performed, a comparative analysis is conducted by calculating the accuracy, precision, recall and computing time values required by the ML algorithm for building the model.

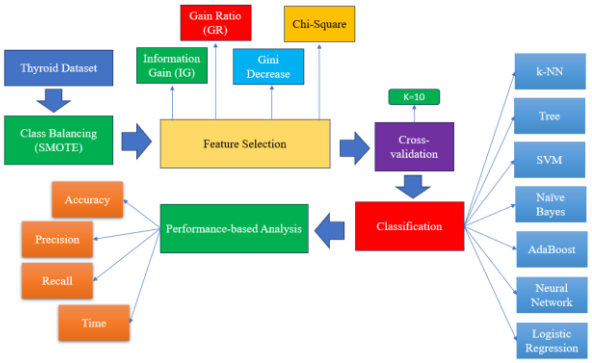


Figure 1. The design of the proposed method.

### A. Thyroid Cancer Recurrence Dataset

The dataset was collected in 2023. This dataset has two classes and thirteen clinical features to predict thyroid cancer recurrence. The dataset was collected over a period of 15 years, with each patient was followed for a minimum of 10 years. This dataset has 383 records, 16 features, and one class that imbalanced classes. The number of majority classes is 175 or 61.8%, and the minority is 108 or 38.1%. Dataset link: https://archive.ics.uci.edu/dataset/915/differentiated+thyroid+cancer+recurrence.

### B. Synthetic Minority Over-sampling Technique (SMOTE)

SMOTE is a method to address the problem of class imbalance [13]. SMOTE is a development of the oversampling method, where the way this method works is by generating new samples from the minority class to make the class proportions more balanced by resampling the minority class samples [14]. The integration between effective feature selection and class imbalance handling has been shown to improve the performance of ML models [15]. In this section, we briefly describe the SMOTE over-sampling algorithm developed by Chawla st al. The SMOTE over-sampling algorithm is illustrated in Figure 2.

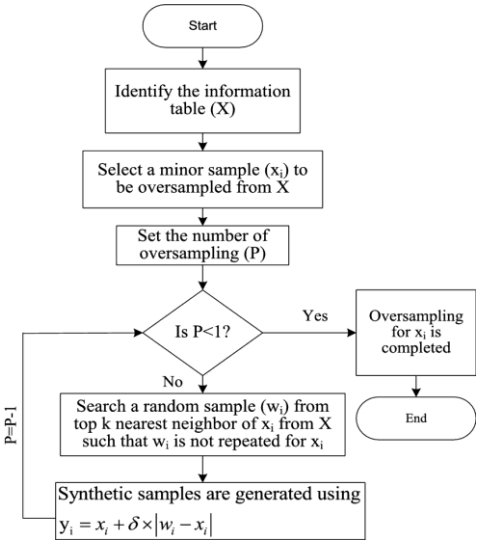


Figure 2. Flowchart of SMOTE algorithm.

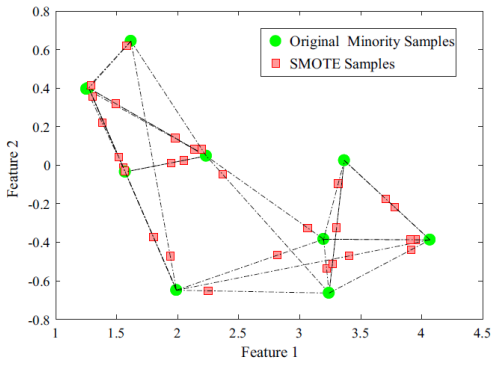


Figure 3. The SMOTE interpolation mechanism displaying the original minority samples and the SMOTE generated patterns

The SMOTE pattern lies on the connection line between the minority class samples and the K-nearest variable values. The position of SMOTE directs inwards, making it more contracted compared to the original distribution, shown in Figure 3.

## C. Features Selection

One of the most important aspects in classification is determining the features to get the best accuracy results [15]. Datasets used in the ML process usually contain redundant and irrelevant features, which do not improve accuracy [16], have no positive effect on the learning model, and may even degrade the performance of the learning model [17]. Therefore, performing relevant feature analysis is essential.

### *1) Information Gain (IG)*

Information gain is the change in class entropy from a previous state to a state when an attribute value occurs. It is applied here to demonstrate how features are pertinent [18]. Decision tree induction is the foundation of this method. Information gain is used as a criterion for choosing attributes. The information gain method has a faster time in the feature selection process than other methods. The features with the most information will be ranked highly in this method; otherwise, and low [19]. IG measures how much the entropy (uncertainty) of a target variable (V17) decreases when a feature is used as a data splitter. The higher the IG value, the more relevant the feature is for determining the class. The determination of IG is based on (1).

$$IG = (Y, X) = H(Y) - H(Y|X) \quad (1)$$

The variable H(Y) is the entropy of the target, and H(Y|X) is the conditional entropy after dividing the data based on feature X.

### *2) Gain-Ratio (GR)*

After dividing the data, the entropy value of the probability distribution subset is calculated using the GR, which normalizes the information gain acquired [15]. When selecting features, the GR considers the dataset's number and size [20]. The GR modifies the information gain, which lessens its bias. The GR chooses an attribute based on the number and size of branches. By accounting for the inherent information of a split, it corrects information gain. GR is an extension of IG. The IG technique tends to be biased towards features with a large number of unique categories. GR improves this by normalizing the IG value against Split Information, so that features with many categories are not automatically considered the most informative. GR calculation uses (2).

$$GR(Y, X) = \frac{IG(Y, X)}{SI(X)} \quad (2)$$

The variable SI(X) is the split information, which measures variation with respect to the division of data from feature X.

### *3) Gini Decrease (GD)*

The Gini Decrease or Mean Decrease Gini (MDG) technique is a measure of how much each feature contributes to the homogeneity of nodes in a decision tree [21]. It measures how much a feature reduces homogeneity when used to split the data in a forest tree. Features with higher Gini Decrease values are considered more important to the predictive power of the model [22].

### *4) Chi-Square ($X^2$) (CS)*

The Chi-square test is a statistical method used for feature selection in machine learning, especially when dealing with categorical data. This test helps determine statistically significant relationships between features and the target variable, enabling identification and selection of the most relevant features for the ML model [23]. Chi-Square evaluates whether there is a significant relationship between features (independent variables, V1-V16) and the class (dependent variable, V17). The larger the $\chi^2$ value, the stronger the association between the features and the target class (V17). The chi-square $\chi^2$ is calculated using (3).

$$x^2 = \sum \frac{(Oij - Eij)^2}{E_{ij}} \quad (3)$$

where $O_{ij}$ represents the observed value and $E_{ij}$ represents the expected value.

## D. Cross-validation (k-folds)

Classification is a popular data mining technique that functions similarly to other methods, such as decision trees and neural networks. These strategies use various methods to assess the available data to produce their prediction [15]. After the feature selection stage and the features that affect the class label based on ranking are obtained, the next stage is classification. Furthermore, the accuracy, precision, recall, error rate, and time required in the classification process are compared using the method of Naive Bayes. J48, AdaBoost, Random Tree, Random Forest, and Support Vector Machine (SVM). The classification model will be validated using k-fold cross-validation. The cross-validation method is commonly used for training sets [24]. Figure 4 shows the k-fold cross-validation.

## E. Performance Analysis

This study examines how the confusion matrix might be used to gauge accuracy and mistake rate. A confusion matrix of size $n \times n$ coupled to a classifier, where $n$ is the total number of classes, displays the anticipated and

actual categorization [25]. The confusion matrix for $n$=2 is shown in Table 1.

Calculating prediction accuracy, precision, and recall is another method for evaluating and comparing classifiers. Both values can be obtained from the confusion matrix Table 1 and calculated using (4), (5), and (6).

$$Accuracy = \frac{TP + TN}{TP + FP + FN + TN} \quad (4)$$

$$Precission = \frac{TP}{TP + FP} \quad (5)$$

$$Recall = \frac{TP}{TP + FN} \quad (6)$$

## III. Result and Discussion

### A. SMOTE Evaluation

We used Orange software (version 3.3.8). This platform simplifies the construction of various data analysis techniques. Orange provides capabilities for categorization, regression, feature selection, association rule mining, and data classification [26].

We performed class balancing using the SMOTE method oversampling approach, using k-fold=10 because we found this option to be the best for ML algorithm classification. The results of the difference in the number of classes after class balancing are shown in Table 2.

Table 2 shows that the application of SMOTE increases the number of classes and records. The number of "Yes" classes increases by 50% and the "No" class remains the same. The total number of records increases by 28.20%. Next, we tested the accuracy, precision, and recall of each ML algorithm using the non-SMOTE and SMOTE datasets shown in Table 3.

Table 1
Confusion Matrix

| Class | Predicted positives | Predicted negatives |
|---|---|---|
| Actual positives instances | Number of true positives instances (TP) | Number of false negatives instances (FN) |
| Actual negatives instances | Number of false positives (FP) | Number of true negatives instances (TN) |

Table 2
Number of Class (Recurred Feature)

| Class (Recurred) | Non-SMOTE | SMOTE |
|---|---|---|
| Yes | 108 | 216 |
| No | 275 | 275 |
| **Total** | **383** | **491** |

Based on Table 3, the application of SMOTE is proven to improve the performance of the ML algorithms k-NN, Tree, SVM, Naive Bayes, AdaBoost, Neural Network (NN), and Logistic Regression (LR). In addition, the ML algorithm experienced a decrease in performance.

However, in general, the application of SMOTE increased performance by 0.61%. Accuracy increased by 0.61%, precision and recall by 0.60%. This finding is in line with other studies [13]. The SMOTE technique does not cause information loss, can avoid overfitting, builds larger decision regions, and improves the accuracy of minority class prediction. Figure 5 presents a comparative analysis of the ML algorithms.

Based on the accuracy, precision, and recall values, we conducted a comparative analysis of the performance of ML algorithms. Several algorithms experienced an increase in performance, such as k-NN increased by 4%, Tree by 0.50%, Naive Bayes by 0.17%, and AdaBoost by 1.50%. In addition, there was a decrease. In addition, we conducted an evaluation of the performance of the algorithm using feature selection such as IG, GR, GD, and CS.

### B. Feature Selection (FS)

FS is the process of selecting the most relevant subset of features from a set of features in a dataset. The main purpose of feature selection is to measure model performance, speed up the training process, and avoid overfitting [27]. By eliminating irrelevant or redundant features, the model becomes simpler, easier to understand, and more efficient. After the initial dataset was class-balanced using the SMOTE technique, we performed feature selection to determine features that greatly influenced early detection of Thyroid Cancer. We used the Information Gain (IG), Gain Ratio (GR), Gini Decrease (GD), and Chi-Square (CS) feature selection methods. These methods rank features that have the highest to the lowest weights, so that the feature weights are obtained in Table 4.

Table 4
Ranking Weighting (W) and Feature Selection (F)

| Information Gain (IG) | | Gain Ratio (GR) | | Gini Decrease (GD) | | Chi-Square (CS) | |
|---|---|---|---|---|---|---|---|
| F | W | F | W | F | W | F | W |
| f16 | 0.806 | f16 | 0.487 | f16 | 0.421 | f13 | 274.57 |
| f11 | 0.526 | f11 | 0.393 | f11 | 0.304 | f15 | 169.32 |
| f13 | 0.387 | f13 | 0.326 | f13 | 0.24 | f16 | 164.94 |
| f8 | 0.338 | f14 | 0.216 | f8 | 0.21 | f12 | 116.27 |
| f12 | 0.314 | f8 | 0.214 | f12 | 0.184 | f11 | 83.41 |
| f15 | 0.166 | f15 | 0.183 | f15 | 0.099 | f10 | 38.48 |
| f10 | 0.136 | f12 | 0.139 | f10 | 0.9 | f3 | 33.35 |

Table 3
Comparison performance algorithm ML with Non-SMOTE and SMOTE

| No | Algorithm | Non-SMOTE | | | SMOTE | | |
|---|---|---|---|---|---|---|---|
| | | Accuracy | Precision | Recall | Accuracy | Precision | Recall |
| 1 | k-NN | 87.4 | 87.6 | 87.4 | 91.4 | 91.7 | 91.4 |
| 2 | Tree | 96.0 | 96.0 | 96.0 | 96.5 | 96.5 | 96.5 |
| 3 | SVM | 96.6 | 96.6 | 96.6 | 96.4 | 96.4 | 96.4 |
| 4 | Naïve Bayes | 92.1 | 92.2 | 92.1 | 92.3 | 92.3 | 92.3 |
| 5 | AdaBoost | 93.8 | 93.8 | 93.8 | 95.3 | 95.3 | 95.3 |
| 6 | Neural Network | 96.4 | 96.4 | 96.4 | 95.6 | 95.6 | 95.6 |
| 7 | Logistic Regression | 97.1 | 97.2 | 97.2 | 96.2 | 96.2 | 96.2 |
| | Average | **94.20** | **94.26** | **94.21** | **94.81** | **94.86** | **94.81** |

TABLE 5
COMPARISON OF ML ALGORITHM ACCURACY USING FEATURE SELECTION

| No. | Algorithm | All Features (%) | Information Gain (IG) | Gain Ratio (GR) | Gini Decrease (GD) | Chi-Square (CS) |
|---|---|---|---|---|---|---|
| 1 | k-NN | **91.4** | 93.5 | 93.6 | 92.2 | 93.1 |
| 2 | Tree | 96.5 | 95.4 | 95.4 | 96.7 | 95.9 |
| 3 | SVM | 96.4 | 96 | 96.4 | 96.5 | 96.6 |
| 4 | Naïve Bayes | 92.3 | 91.6 | 91.8 | 92.5 | 93.1 |
| 5 | AdaBoost | 95.3 | 95.4 | 96.4 | 95.9 | 96.4 |
| 6 | **Neural Network** | 95.6 | 94.9 | 96.2 | 96.5 | **97** |
| 7 | Logistic Regression | 96.2 | 96.2 | 96.7 | 96.6 | 96.5 |
| | Average | **94.81** | **94.71** | **95.21** | **95.27** | **95.51** |

Feature selection using four methods produces different features. The IG, GR, and GD techniques produce the same highest weight feature, which is f16, but different from the selection results for the CS technique, which is f13. Next, we compare the performance of the ML algorithm using the selected features. The five highest-weight features sorted by DESC are used to measure the performance of the algorithm.

### C. A Comparison Analysis of Algorithm ML

In this test, we apply k-fold cross-validation = 10, training data (80%), and test data (20%) because our findings show that this value produces optimal accuracy for each algorithm test. Performance measure testing uses a Confusion Matrix to evaluate how well the ML algorithm performs. The results of the comparison of accuracy performance based on the selection of IG, GR, GD, and CS features can be seen in Table 5.

We found differences in the accuracy of each ML algorithm's performance. In this case, the algorithm performance is more optimal using the Chi-Square (CS) technique compared to IG, GR, and GD feature selection. This is different from other studies [15], which use the same approach.

Based on the average accuracy, there is an increase in accuracy after feature selection. The GR technique increases the accuracy by 0.4%, GD by 0.465, and CS by 0.7%, except for the IG technique -0.10%. In addition, we found that the k-NN algorithm with all features is the worst combination, but the Neural Network (NN) + Chi-Square (CS) algorithm is the best combination in this dataset. This is because the NN algorithm can learn and improve its performance over time with more data given. NN can be used in various applications, ranging from speech recognition, text, to images. Next, we conducted a comparative analysis of the performance of the four feature selection techniques shown in Figure 6.

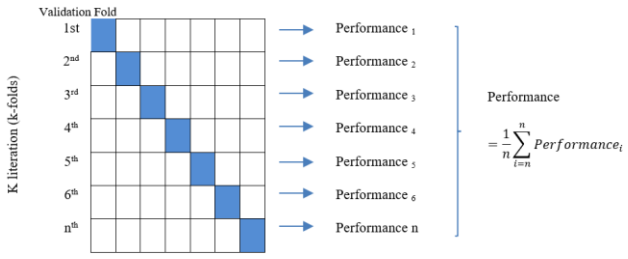


Figure 4. The procedure of k-fold validation.

Figure 7 shows that the feature selection technique using Chi-Square (CS) optimally increases the accuracy of the ML algorithms. Meanwhile, the IG technique exhibits the weakest performance. Figure 8 presents the precision evaluation of the ML algorithms.

Figure 8 shows the difference in precision using all features and feature selection. Precision analysis using feature selection techniques produces better performance except for IG feature selection. Consistently, the combination of NN+CS algorithms is still the best, and k-NN+all features is still the worst. Next, we evaluate the recall of the ML algorithm in Figure 9.

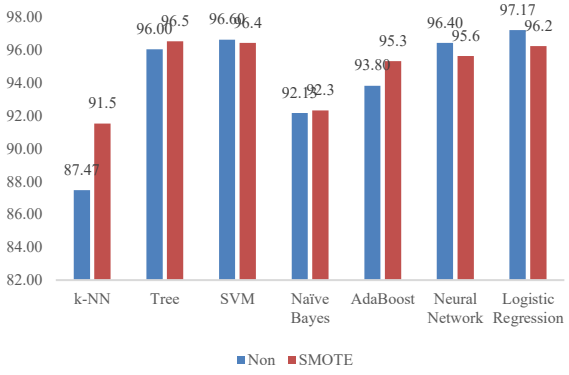


Figure 5. A Comparative Non-SMOTE, and SMOTE.

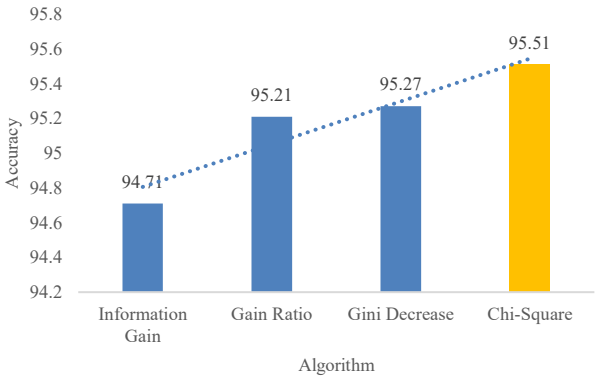


Figure 6. Performance comparison of feature selection techniques.

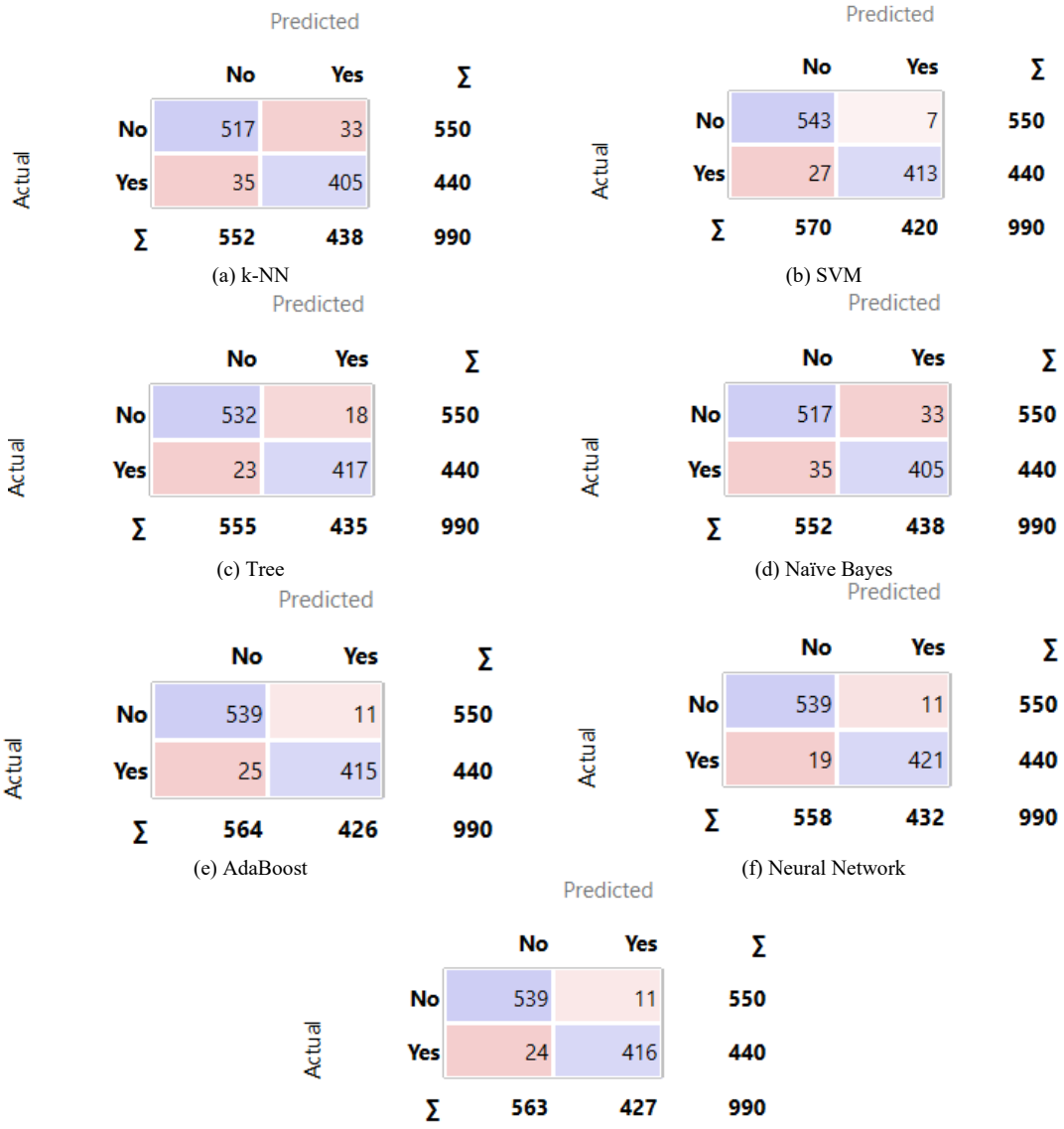


Figure 7. Confusion Matrix of algorithm (a) k-NN (b) SVM (c) Tree (d) Naïve Bayes (e) AdaBoost (f) Neural Network (g) Logistic Regression + Chi-Square.

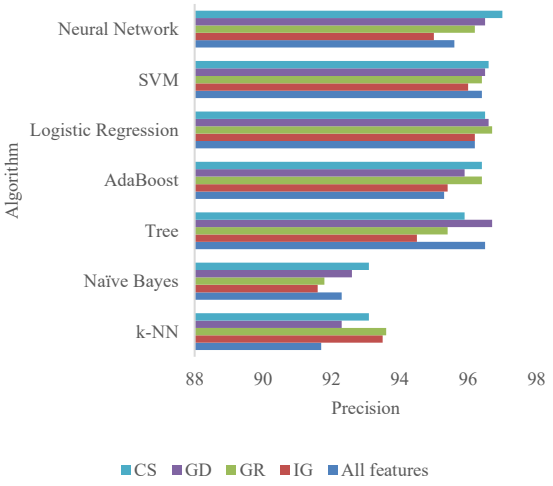


Figure 8. A Comparison precision of algorithm ML

Figure 9 shows the difference in recall using all features and feature selection. The recall value using feature selection has better performance except for IG feature selection. In this evaluation, the combination of NN+CS algorithms is still the best, and k-NN+all features are the worst. Next, we evaluate the time to build a model of the ML algorithm in Figure 8.

Figure 10 shows the time required to build each ML model. The evaluation results indicate that the Naive Bayes method required the least time. This finding aligns with the research in [15], as Naive Bayes can determine probabilities precisely by calculating the probability of one class for each attribute group. Meanwhile, the Random Forest algorithm takes the longest time compared to other algorithms. This is because the Random Forest (RF) method uses a kernel that searches for a hyperplane, so the processing time is longer. In addition, the test results show that feature selection affects the time required. The use of the IG technique

takes less time when compared to using all features and GR, GD, and CS techniques.

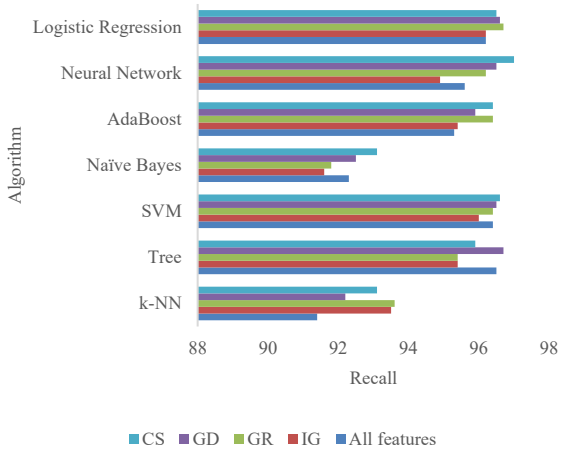


Figure 9. A Comparison recall of algorithm ML.

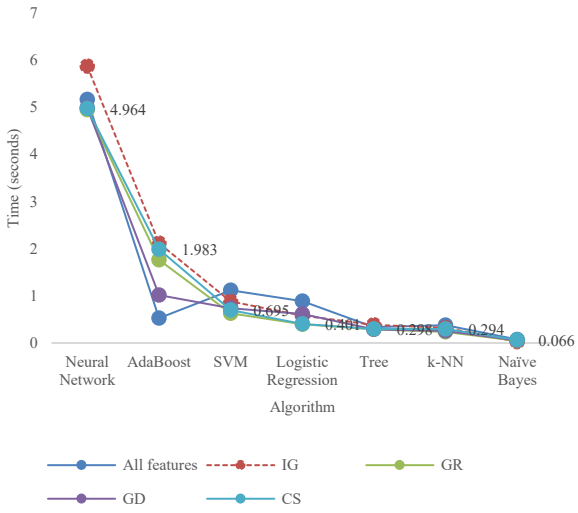


Figure 10. A comparison of consuming time (seconds) of build model the algorithm ML.

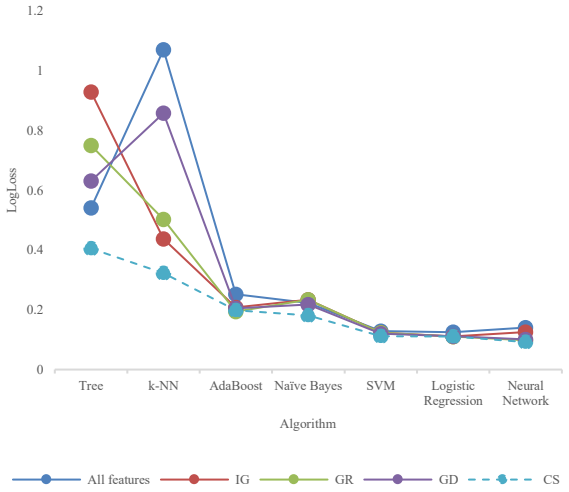


Figure 11. A comparison of Log Loss the algorithm ML.

TABLE 6
STATISTICAL TEST

| Information | | N | Percent |
|---|---|---|---|
| Sample | Training | 290 | 71,3% |
| | Holdout | 117 | 28,7% |
| Valid | | 492 | 100,0% |

Next, we perform a performance comparison against log loss. This technique is used to calculate how close our model's predicted probability is to the actual label. The further the predicted probability is from the actual value, the higher the log loss value. The log loss comparison of ML algorithms is shown in Figure 11.

Figure 11 presents the log loss in calculating the model prediction probability. The evaluation results show that the NN method has the best log loss. This is because the NN algorithm can determine probabilities more precisely by calculating the possibility of one class for each existing attribute group. Meanwhile, the k-NN algorithm has the worst log loss compared to other algorithms. This is because the k-NN algorithm requires high computation for large datasets, is sensitive to outliers, and has difficulty in handling high-dimensional data. In addition, the test results show that feature selection affects the log loss value. The use of the CS technique has a better log loss when compared to all other features and feature selection techniques. Next, we conducted statistical tests on the dataset. We conducted a statistical analysis on the dataset we used. The tool we used was SPSS. The results of our analysis are presented in Table 6.

Based on Table 6, the proportion of data split between training and holdout is considered ideal for predictive modeling with 290 observations (71.3%). This portion is used to train the model so it can learn patterns from the data. This proportion is close to the general standard (around 70-80%) in data splitting practices. The holdout (Testing or Validation) consists of 117 observations (28.7%). This portion is set aside and not used during the training phase, but is used to test or validate the model's performance. The 28.7% proportion is relatively large (usually 20-30%), making it quite representative for model evaluation. With 407 valid data points, the sample size is sufficient for analysis, although the results still need to be considered in light of the model's complexity. Figure 12 presents a visualization of the focal record points. Figure 12 shows a lower-dimensional projection of the predictor space, which contains a total of 16 predictors

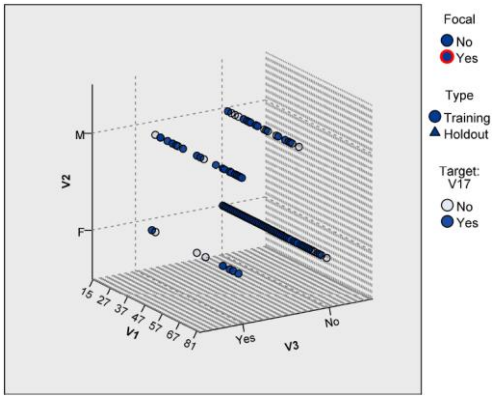


Figure 12. Select points to use as focal records.

## IV. CONCLUSION

Based on the comparative analysis conducted, it can be concluded that class balancing in the Thyroid Cancer Recurrence (TCR) Dataset significantly impacts classification model performances. The SMOTE balancing technique proved effective in improving algorithm performance. Although not significant, the accuracy variable increased by 0.61%, precision and recall by 0.60%. Logistic Regression (LR) gave the best results consistently with an accuracy of 97.2%.

In addition, we conducted a feature selection evaluation using four techniques, namely IG, GR, GD, and CS to select features that have the highest weight correlation to the class. The feature selection results varied across methods. The IG technique produced eight features that had the highest weights, namely f16, f11, f13, f8, f12, f15, f10, and f9. This is different from the selection using the GR technique which produced features namely f16, f11, f13, f14, f8, f15, f12, and f10. The GD technique produced features f16, f11, f13, f8, f12, f15, f10, and f1. The CS technique produced features f13, f15, f16, f12, f11, f10, f3, and f4. Based on these selected features, we conducted an evaluation of the ML algorithm to obtain comprehensive information on the performance of the algorithm.

The evaluation results of the TCR dataset that has been selected for features and class balancing, we found that the combination of the NN algorithm with CS feature selection is the best combination for the best accuracy. In addition, we evaluated precision and recall. Based on the results of the performance comparison, we found that the combination of the NN + CS algorithm was consistently the best. In addition, the evaluation of the model building time showed that the Naive Bayes algorithm had the best time, and the Neural Network took longer. The number of features affects the time required. The use of the IG technique takes less time when compared to all features and GR, GD, and CS techniques. Evaluation of the log loss in calculating the probability of model prediction, the NN algorithm has the best value.

Class imbalance and feature selection handling strategies should be integral components of the medical classification system development pipeline, as they impact overall interpretability and detection accuracy. Further research is recommended to evaluate the effectiveness of other imbalance handling techniques, such as ensemble-based sampling or cost-sensitive learning, and to apply these approaches to larger and more complex clinical datasets.

## DECLARATIONS

### Conflict of Interest

No conflict of interest declared.

### CRediT Authorship Contribution

First Author is **Agus Wantoro** (Conceptualization, Methodology, Supervision). He conceived the study, formulated the research framework, and supervised all stages of the research and manuscript development. The second Author is **Wahyu Caesarendra** (Data Curation, Formal Analysis, Software). He was responsible for preparing and preprocessing the dataset, implementing feature selection algorithms, and conducting performance evaluations using machine learning models. The third author is **Admi Syarif** (Investigation, Validation, Writing, Original Draft). He carried out the experimental investigations, validated the classification results, and contributed to the initial writing of the manuscript. The fourth author is **Hari Soetanto** (Visualization, Writing – Review & Editing, Resources). He developed the graphical and tabular visualizations, reviewed and edited the manuscript critically for intellectual content, and provided clinical insights and domain-specific resources on thyroid cancer recurrence.

### Funding

This research is supported by Universitas Aisyah Pringsewu (UAP) through the 2025 Independent Research Scheme. Additional support is provided by the Faculty of Engineering and Informatics, through independent research funding for the development of medical AI systems.

### Acknowledgment

The authors express their deepest gratitude to the Faculty of Health, Universitas Aisyah Pringsewu (UAP), for providing access to the anonymous thyroid cancer patient data used in this study. We also thank the Machine Learning (ML) Research Group at UAP for their valuable input during the model development and evaluation phases. Special appreciation is extended to the physicians whose clinical insights contributed significantly to the interpretation of the thyroid cancer recurrence prediction results.

## REFERENCES

[1] A. Schindele *et al.*, "Interpretable machine learning for thyroid cancer recurrence prediction: Leveraging XGBoost and SHAP analysis," *Eur. J. Radiol.*, vol. 186, May 2025, doi: 10.1016/j.ejrad.2025.112049.

[2] A. H. Barfejani *et al.*, "Predicting overall survival in anaplastic thyroid cancer using machine learning approaches," *Eur. Arch. Oto-Rhino-Laryngology*, vol. 282, no. 3, pp. 1653–1657, 2025, doi: 10.1007/s00405-024-08986-2.

[3] D. W. Chen, B. H. H. Lang, D. S. A. McLeod, K. Newbold, and M. R. Haymart, "Thyroid cancer," *Lancet*, vol. 401, no. 10387, pp. 1531–1544, May 2023, doi: 10.1016/S0140-6736(23)00020-X.

[4] A. Kuang, V. L. Kouznetsova, S. Kesari, and I. F. Tsigelny, "Diagnostics of Thyroid Cancer Using Machine Learning and Metabolomics," *Metabolites*, vol. 14, no. 1, 2024, doi: 10.3390/metabo14010011.

[5] R. Iacob *et al.*, "Evaluating the Role of Breast Ultrasound in Early Detection of Breast Cancer in Low- and Middle-Income Countries: A Comprehensive Narrative Review," *Bioengineering*, vol. 11, no. 3, 2024, doi: 10.3390/bioengineering11030262.

[6] Y.-M. Huang *et al.*, "Correction: Huang et al. Systemic Anticoagulation and Inpatient Outcomes of Pancreatic Cancer: Real-World Evidence from U.S. Nationwide Inpatient Sample. Cancers 2023, 15, 1985," *Cancers*, vol. 16, no. 6, 2024, doi: 10.3390/cancers16061181.

[7] I. O. Lixandru-Petre *et al.*, "Machine Learning for Thyroid Cancer Detection, Presence of Metastasis, and Recurrence Predictions—A Scoping Review," *Cancers (Basel)*, vol. 17, no. 8, pp. 1–27, 2025, doi: 10.3390/cancers17081308.

[8] S. Li, Z. Tang, L. Yang, M. Li, and Z. Shang, "Application of deep reinforcement learning for spike sorting under multi-class imbalance," *Comput. Biol. Med.*, vol. 164, pp. 107253, 2023, doi: 10.1016/j.compbiomed.2023.107253.

[9] X. Song *et al.*, "Evolutionary computation for feature selection in classification: A comprehensive survey of solutions, applications and challenges," *Swarm Evol. Comput.*, vol. 90, pp. 101661, 2024, doi: 10.1016/j.swevo.2024.101661.

[10] W. Chen, K. Yang, Z. Yu, Y. Shi, and C. L. P. Chen, "A survey on imbalanced learning: Latest research, applications and future directions," *Artif. Intell. Rev.*, vol. 57, no. 6, pp. 137, 2024, doi: 10.1007/s10462-024-10759-6.

[11] L. C. M. Liaw, S. C. Tan, P. Y. Goh, and C. P. Lim, "A histogram SMOTE-based sampling algorithm with incremental learning for imbalanced data classification," *Inf. Sci. (NY).*, vol. 686, pp. 121193, 2025, doi: 10.1016/j.ins.2024.121193.

[12] K. E. Setiawan, "Predicting Recurrence in Differentiated Thyroid Cancer: a Comparative Analysis of Various Machine Learning Models Including Ensemble Methods With Chi-Squared Feature Selection," *Commun. Math. Biol. Neurosci.*, vol. 2024, no. Scenario 1, pp. 1–29, 2024, doi: 10.28919/cmbn/8506.

[13] G. Husain *et al.*, "SMOTE vs. SMOTEENN: A Study on the performance of resampling algorithms for addressing class imbalance in regression models," *Algorithms*, vol. 18, no. 1, pp. 1–16, 2025, doi: 10.3390/a18010037.

[14] M. F. Ijaz, G. Alfian, M. Syafrudin, and J. Rhee, "Hybrid prediction model for type 2 diabetes and hypertension using DBSCAN-based outlier detection, synthetic minority over sampling technique (SMOTE), and random forest," *Appl. Sci.*, vol. 8, no. 8, 2018, doi: 10.3390/app8081325.

[15] H. Sulistiani, A. Syarif, K. Muludi, and Warsito, "Performance evaluation of feature selections on some ML approaches for diagnosing the narcissistic personality disorder," *Bull. Electr. Eng. Informatics*, vol. 13, no. 2, pp. 1383–1391, 2024, doi: 10.11591/eei.v13i2.6717.

[16] J. Wang, S. Zhou, Y. Yi, and J. Kong, "An improved feature selection based on effective range for classification," *Sci. World J.*, vol. 2014, 2014, doi: 10.1155/2014/972125.

[17] S. Bashir, Z. S. Khan, F. H. Khan, A. Anjum, and K. Bashir, "Improving heart disease prediction using feature selection approaches," in *2019 16th International Bhurban Conference on Applied Sciences and Technology (IBCAST)*, 2019, pp. 619–623. doi: 10.1109/IBCAST.2019.8667106.

[18] J. Gao, Z. Wang, T. Jin, J. Cheng, Z. Lei, and S. Gao, "Information gain ratio-based subfeature grouping empowers particle swarm optimization for feature selection," *Knowledge-Based Syst.*, vol. 286, pp. 111380, 2024, doi: 10.1016/j.knosys.2024.111380.

[19] P. Bhat and K. Dutta, "A multi-tiered feature selection model for android malware detection based on feature discrimination and information gain," *J. King Saud Univ. - Comput. Inf. Sci.*, vol. 34, no. 10, Part B, pp. 9464–9477, 2022, doi: 10.1016/j.jksuci.2021.11.004.

[20] M. Trabelsi, N. Meddouri, and M. Maddouri, "A new feature selection method for nominal classifier based on formal concept analysis," *Procedia Comput. Sci.*, vol. 112, pp. 186–194, 2017, doi: 10.1016/j.procs.2017.08.227.

[21] Y. Sang and X. Dang, "Grouped feature screening for ultrahigh-dimensional classification via Gini distance correlation," *J. Multivar. Anal.*, vol. 204, pp. 1–25, 2024, doi: 10.1016/j.jmva.2024.105360.

[22] Y. Zhang *et al.*, "Feature selection based on neighborhood rough sets and Gini index," *PeerJ Comput. Sci.*, vol. 9, pp. e1711, 2023, doi: 10.7717/peerj-cs.1711.

[23] A. Abdo, R. Mostafa, and L. Abdel-Hamid, "An Optimized Hybrid Approach for Feature Selection Based on Chi-Square and Particle Swarm Optimization Algorithms," *Data*, vol. 9, no. 2, 2024, doi: 10.3390/data9020020.

[24] T. Yan, S.-L. Shen, A. Zhou, and X. Chen, "Prediction of geological characteristics from shield operational parameters by integrating grid search and k-fold cross validation into stacking classification algorithm," *J. Rock Mech. Geotech. Eng.*, vol. 14, no. 4, pp. 1292–1303, 2022, doi: 10.1016/j.jrmge.2022.03.002.

[25] M. Ohsaki, P. Wang, K. Matsuda, S. Katagiri, H. Watanabe, and A. Ralescu, "Confusion-matrix-based kernel logistic regression for imbalanced data classification," *IEEE Trans. Knowl. Data Eng.*, vol. 29, no. 9, pp. 1806–1819, 2017, doi: 10.1109/TKDE.2017.2682249.

[26] I. Popchev and D. Orozova, "Algorithms for Machine Learning with Orange System," *Int. J. Online Biomed. Eng.*, vol. 19, no. 4, pp. 109–123, 2023, doi: 10.3991/ijoe.v19i04.36897.

[27] F. Miao, Y. Wu, G. Yan, and X. Si, "Dynamic multi-swarm whale optimization algorithm based on elite tuning for high-dimensional feature selection classification problems," *Appl. Soft Comput.*, vol. 169, pp. 112634, 2025, doi: 10.1016/j.asoc.2024.112634.