**Jurnal Teknik Informatika (JUTIF)**
P-ISSN: 2723-3863
E-ISSN: 2723-3871

Vol. 6, No. 4, Agustus 2025, Page. 2625-2637
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2025.6.4.5166

# Optimizing Type 2 Diabetes Classification with Feature Selection and Class Balancing in Machine Learning

**Agus Wantoro*[1], Aviv Fitria Yuliana[2], Dwi Yana Ayu Andini[3], Ikna Awaliyani[4], Wahyu Caesarendra[5]**

[1]Informatics, Universitas Aisyah Pringsewu, Indonesia
[2,3]Software Engineering, Universitas Aisyah Pringsewu, Indonesia
[4]Information Technology Education, Universitas Aisyah Pringsewu, Indonesia
[5]Departement of Mechanical and Mechatronics Curtin, University Malaysia

Email: [1]aguswantoro@aisyahuniversity.ac.id

Received: Jul 23, 2025; Revised: Aug 8, 2025; Accepted: Aug 16, 2025; Published: Aug 24, 2025

**Abstract**

Type 2 Diabetes (T2DM) is a crucial factor in patient survival and treatment effectiveness. Errors in diabetes detection lead to disease severity, high costs, prolonged healing time, and a decline in service quality. Additionally, a major challenge in developing Machine Learning (ML)-based detection decision support systems is the class imbalance in medical data as well as the high feature dimensionality that can affect the accuracy and efficiency of the model. This research proposes an approach based on feature selection (FS) and handling class imbalance to improve performance in type 2 diabetes. Several feature selection techniques such as Information Gain (IG), Gain Ratio (GR), Gini Decrease (GD), Chi-Square (CS), Relief-F, and FCBF can perform feature selection based on weighting ranking. Furthermore, to address the imbalanced class distribution, we utilize the Synthetic Minority Over-Sampling Technique (SMOTE). ML classification models such as Support Vector Machine (SVM), Gradient Boosting (GB), Tree, Neural Network (NN), Random Forest (RF), and AdaBoost were tested and evaluated based on the confusion matrix including accuracy, precision, recall, and time. The experimental results show that the combination of strategies for handling imbalanced classes significantly improves the predictive performance of ML algorithms. In addition, we found that the combination of feature selection techniques IG+AdaBoost consistently demonstrates optimal performance. This study emphasizes the importance of data preprocessing and the selection of the right algorithms in the development of machine learning-based T2DM detection systems. Accurate detection can reduce the severity of disease, lower treatment costs, speed up the healing process, and improve healthcare services.

***Keywords:*** *Diabetes, Feature selection, Imbalance class, Machine Learning.*



## 1. INTRODUCTION

Diabetes Mellitus (DM) is one of the chronic diseases whose prevalence continues to increase globally, including in Indonesia. DM is a chronic metabolic disorder characterized by increased blood sugar levels in the body. This is caused by a disruption in insulin secretion. In 2023, approximately 415 million people aged between 20 and 79 years are reported to suffer from DM [1]. Diabetes mellitus (DM) is generally categorized into three: Type 1, Type 2, and Gestational Diabetes (GDM). Type 1 diabetes (T1DM) affects 5% to 10%. This is characterized by autoimmune damage to the insulin-producing beta cells in the pancreas. [2]. Type 2 diabetes (T2DM) accounts for about 90% of all diabetes cases. In T2DM, the response to insulin is reduced. T2DM is mainly seen in people over the age of 45. It is increasingly observed in children, adolescents, and adults due to rising body weight, lack of physical activity, and a high-energy diet [3]

Early detection and accurate diagnosis of T2DM is very important to prevent serious complications that can arise from this disease [4]. In recent years, rapid advances in the field of Machine

Learning (ML) have paved the way for the development of TD2M detection systems. ML based approaches have been widely used to improve the accuracy of diabetes predictions [5]. ML algorithms have the ability to analyze medical data, including genetic data, medical images, and clinical histories, to identify complex patterns that may not be visible to the human eye. The potential of ML can enhance the accuracy of diagnoses and predict disease risks in various studies [6]

However, the application of ML in T2DM faces two main challenges: feature selection (FS) and class imbalance. T2DM datasets are often used as a basis for evaluating breast cancer diagnostic models [7]. This dataset has seventeen features, where all features need to be analyzed for their relevance to model accuracy. The presence of irrelevant features can lead to decreased model performance, increased computational complexity, and overfitting [8]. Therefore, effective feature selection (FS) techniques are crucial to identify the most informative subset of features that influence model performance and reduce noise [9]

In addition, the T2DM dataset has class imbalance, where the number of 'Yes' class is lower than that of '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 actually the primary target of early detection [7]. Various studies have highlighted the negative impact of class imbalance on ML performance and have proposed methods to address it, such as oversampling [8]

To address the issue of class imbalance, over-sampling techniques such as SMOTE (Synthetic Minority Over-sampling Technique) can be a choice. This technique can enhance the representation of the minority class in the dataset, allowing the model to learn better [10]. The integration of effective feature selection and handling class imbalance has proven to improve the performance of Machine Learning models in breast cancer diagnosis [11]. However, a comprehensive evaluation of the combination of various feature selection techniques and methods for dealing with class imbalance is still needed to identify the best approach in this context.

This research aims to conduct a performance-based analysis of various FS methods such as Information Gain (IG), Gain Ratio (GR), Gini Decrease (GD), Chi-Square (CS), Relief-F, and Fast Correlation Based Filter (FCBF) with the class imbalance handling technique SMOTE in the context of T2DM using ML algorithms such as Support Vector Machine (SVM), Gradient Boosting (GB), Decision Tree, Neural Network (NN), Random Forest (RF), and AdaBoost. A comparative analysis of the six feature selection techniques, and SMOTE has not been reported. Therefore, by comparing the performance of algorithms from various combinations of oversampling methods and feature selection, it is hoped to provide new findings in the form of the most optimal approach to build an accurate, sensitive, and specific T2DM detection model, thus contributing to science and improving the quality of healthcare services for T2DM patients.

## 2. METHOD

This section explains the research stages. The first stage is data collection. We used the T2DM dataset, which we then performed a class balancing process using the SMOTE technique. Next, we performed feature selection using IG, GR, GD, CS, Relief-F, and FCBF techniques. The selected features were used for diabetes classification using cross-validation. The proposed design is shown in Figure 1.

After classifying T2DM, a comparative analysis was then conducted by calculating the accuracy, precision, recall, and the computation time required by the ML algorithm in building the model.

### 2.1. Dataset

The Type 2 Diabetes Mellitus dataset is a collection of data that contains information related to type 2 diabetes. This dataset was collected by Neha Prerna Tigga and Dr. Shruti Garg of the Department

Vol. 6, No. 4, Agustus 2025, Page. 2625-2637
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2025.6.4.5166

of Computer Science and Engineering [7]. This dataset has 952 records, 17 varying features, and 2 classes. Table 1 shows the features in the T2DM dataset.

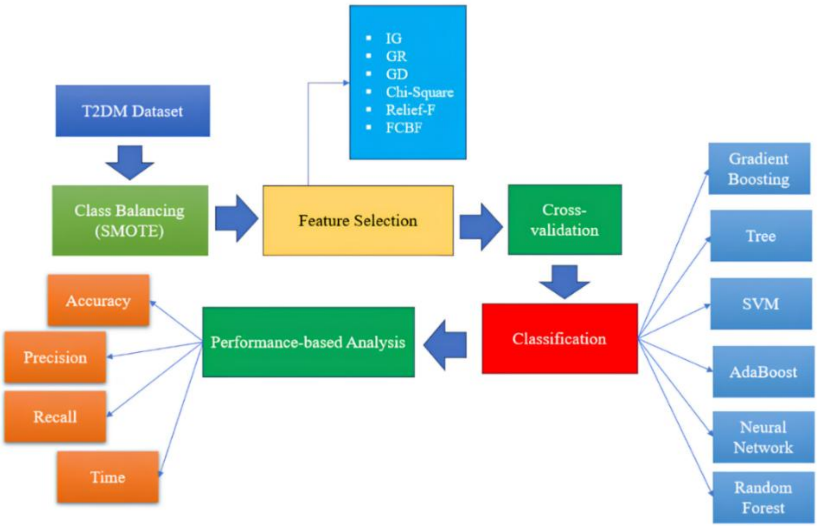


Figure 1. The design of the proposed method

Table 1. Feature T2DM dataset

| Code | Feature |
|---|---|
| f1 | Age (Ag) |
| f2 | Gender (Gen) |
| f3 | Family Diabetes (FD) |
| f4 | High Blood Pressure (Hbp) |
| f5 | Physically Active (PA) |
| f6 | BMI (BM) |
| f7 | Smoking (Smo) |
| f8 | Alcohol (Alc) |
| f9 | Sleep (Sle) |
| f10 | Sound Sleep (SS) |
| f11 | Regular Medicine (RM) |
| f12 | Junk Food (JF) |
| f13 | Stress (Str) |
| f14 | Blood Pressure Level (BpL) |
| f15 | Pregnancies (Pre) |
| f16 | Pre-diabetes (Pd) |
| f17 | Urination Freq (UF) |

### 2.2. SMOTE

Synthetic Minority Over-sampling Technique (SMOTE) is a method to address the issue of class imbalance. [11]. SMOTE is an extension of the oversampling method, where the way this method works is by generating new samples that originate from the minority class to create a more balanced class proportion by resampling the minority class samples [12]. The integration of effective feature selection and handling class imbalance has been proven to improve the performance of ML models [13]

P-ISSN: 2723-3863
https://jutif.if.unsoed.ac.id
E-ISSN: 2723-3871
DOI: https://doi.org/10.52436/1.jutif.2025.6.4.5166

### 2.3. Feature Selection (FS)

One of the most important things in classification is determining the features to achieve the best optimum results [13]. ML datasets often contain redundant features that are irrelevant to the class. This can degrade algorithm performance [14], and does not have an effect on the learning model [15], Therefore, it is important to conduct relevant feature analysis. We perform FS using six best techniques is Information Gain (IG), Gain Ratio (GR), Gini Decrease (GD), Chi-Square (CS), Relief-F, and Fast Correlation Based Filter (FCBF).

#### 2.2.1. Information Gain (IG)

IG is a technique in ML used to measure how well a feature separates data based on different classes or targets [16]. Information Gain measures the amount of uncertainty (entropy) reduction that occurs after knowing the value of a feature. The higher a feature's value, the better it helps the model to predict or classify the data [17].

#### 2.2.2. Gain-Ratio (GR)

GR is a technique in machine learning that is used to select the best features by partitioning the data [13]. GR is a modification of Information Gain (IG) aimed at addressing the weaknesses in handling attributes with many values (discrete). GR helps to reduce the bias that may occur in IG, especially in data that has many categories [18]. The GR modifies the IG technique, which lessens its overfitting. How GR works is by selecting attributes based on quantity and size by taking into account inherent information.

#### 2.2.3. Gini Decrease (GD)

Gini Decrease or Mean Decrease Gini (MDG) is a measure of how much each feature contributes to the homogeneity of nodes in a decision tree [19]. This measures how much a feature reduces homogeneity when used to split data in a forest tree. Features with higher Gini Decrease values are considered more important for the predictive power of the model [20].

#### 2.2.4. 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 the statistically significant relationship between features and the target variable, allowing for the identification and selection of the most relevant features for the ML model [21]

#### 2.2.5. Fast Correlation Based Filter (FCBF)

The FCBF technique is a feature selection algorithm based on the idea that a good feature is one that is relevant to the class but not redundant with other relevant features. Therefore, FCBF uses two approaches by measuring the correlation between two random variables, namely based on classical linear correlation and linear correlation coefficient based on information theory [22]

#### 2.2.6. Relief-F

Relief-F utilizes a weighting technique to measure significance. Features with a weight value above the threshold will be selected to represent an instance. The retrieval process uses the Mahala Nobis distance method, which measures distance by taking into account the correlation between attributes using the inverse of the variance-covariance matrix [23]

**Jurnal Teknik Informatika (JUTIF)**
P-ISSN: 2723-3863
E-ISSN: 2723-3871
Vol. 6, No. 4, Agustus 2025, Page. 2625-2637
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2025.6.4.5166

### 2.3. Cross-Validation (k-folds)

A technique in machine learning for assessing model performance by dividing the data into multiple parts (folds) and repeatedly training and testing the model on different combinations of folds. The goal is to obtain a more accurate estimate of how the model will perform on new, previously unseen data and to prevent overfitting [24]. The correct k-fold selection affects the performance of the ML algorithm. Figure 2 shows the cross-validation stages in ML

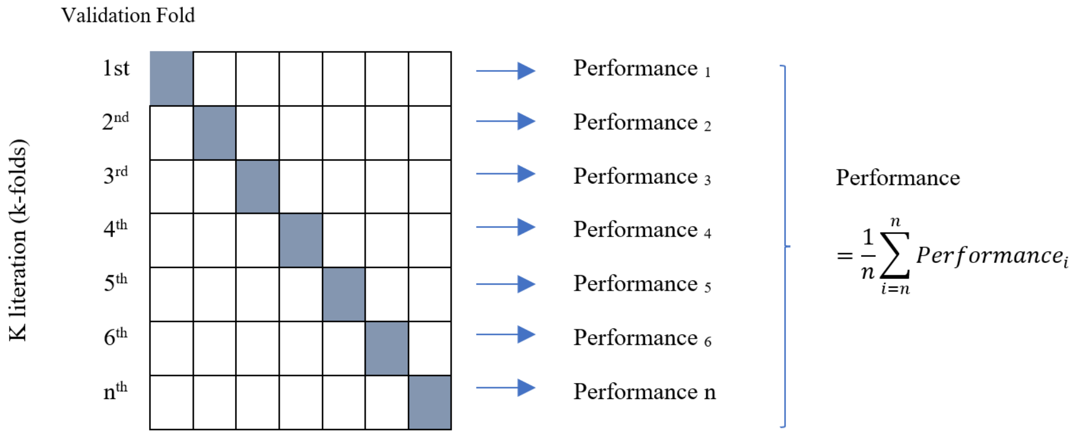


Figure 2. The stage cross-validation

### 2.4. Performance Analysis

This study examines how the Confusion Matrix (CM) is used to measure accuracy, precision, and recall. The matrix is $n \times n$ combined by class, where $n$ is the total number of classes [25]. The CM shown in Table 2.

Table 2. Confusion Matrix

| Class | Positives | Negatives |
|---|---|---|
| Positives | Number of True Positives (TP) | Number of False Positives (FP) |
| Negatives | Number of False Negatives (FN) | Number of True Negatives (TN) |

Accuracy, precision, and recall are other methods to evaluate and compare classifiers. These variables can be obtained by using equations (1-3)

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

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

$$Recall = \frac{TP}{TP+FN} \quad (3)$$

## 3. RESULT

### 3.1. SMOTE Evaluations

We use Orange software (version 3.3.8). This platform simplifies the construction of several analytical data techniques. Orange has the ability to categorize, perform regression, classification, feature elimination, create association rules, and class adjustment on datasets. [26]. We performed class balancing using the oversampling method SMOTE with k-fold = 5 because our findings indicate that

P-ISSN: 2723-3863
https://jutif.if.unsoed.ac.id
E-ISSN: 2723-3871
DOI: https://doi.org/10.52436/1.jutif.2025.6.4.5166

this choice is the best for ML algorithm classification. The results of the difference in the number of classes after class balancing are shown in Table 3.

Table 3. Imbalance class with SMOTE

| Class | Non-SMOTE | SMOTE |
|---|---|---|
| No | 686 | 686 |
| Yes | 266 | 532 |
| **Total** | **952** | **1218** |

Table 3 shows that the application of SMOTE increases the number of classes and records. The number of “Yes” class increases by 50% while the “No” class remains the same. The total number of records increases by 27.94%. We used this dataset and divided it into 2 parts. The first part for training data and the second part for testing data. The percentage of training data is 80% and testing data is 20%. Based on our testing, this value is the best for the performance of the ML algorithm. Furthermore, we conducted performance testing of the ML algorithm using both SMOTE and non-SMOTE datasets. The testing results include accuracy, precision, and recall, which are displayed in Table 4.

Table 4. Comparison performance algorithm ML with Non-SMOTE and SMOTE

| No | Algorithm | Non-SMOTE | | | SMOTE | | |
|---|---|---|---|---|---|---|---|
| | | Accuracy | Precision | Recall | Accuracy | Precision | Recall |
| 1 | Random Forest | 0.947 | 0.947 | 0.947 | 0.965 | 0.965 | 0.965 |
| 2 | Neural Network | 0.959 | 0.959 | 0.959 | 0.961 | 0.961 | 0.961 |
| 3 | Tree | 0.931 | 0.929 | 0.93 | 0.953 | 0.953 | 0.953 |
| 4 | Gradient Boosting | 0.935 | 0.934 | 0.934 | 0.936 | 0.936 | 0.936 |
| 5 | SVM | 0.933 | 0.932 | 0.932 | 0.912 | 0.912 | 0.916 |
| 6 | AdaBoost | **0.964** | 0.964 | 0.964 | **0.970** | 0.970 | 0.971 |

Based on the data in Table 3, the performance of the ML algorithms shows varying accuracies. The AdaBoost algorithm consistently stands out as the best. Furthermore, the implementation of SMOTE has proven effective in improving the performance of the ML algorithms. The increases were 1.8% for the Random Forest, 0.2% for Neural Network, 2.2% for Tree, 0.1% for Gradient Boosting, and 0.6% for AdaBoost, except for SVM which decreased by 2.1%. This is because the SVM algorithm has a high training time, making it unsuitable for fairly large datasets. In addition, the SVM classifier does not work well with overlapping classes. If SVM is combined with SMOTE techniques, then SMOTE cannot consider the majority class when generating synthetic samples. Furthermore, this method can create synthetic samples among noise samples. As a result, the augmented dataset will have more noise than the original data, thus reducing performance. Next, we conducted a comparative analysis of the average values of the ML algorithms displayed in Figure 3.

Based on the average values of accuracy, precision, and recall, the application of the SMOTE technique has been proven to improve accuracy by 0.6%, precision by 0.7%, and recall by 0.6%. This finding is in line with other research [11], This is because the SMOTE technique does not cause information loss. The SMOTE technique can prevent overfitting, build a larger decision area, and can increase the accuracy of minority class predictions. Next, we performed feature ranking using six selection techniques to obtain the best features in this case.

Vol. 6, No. 4, Agustus 2025, Page. 2625-2637
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2025.6.4.5166

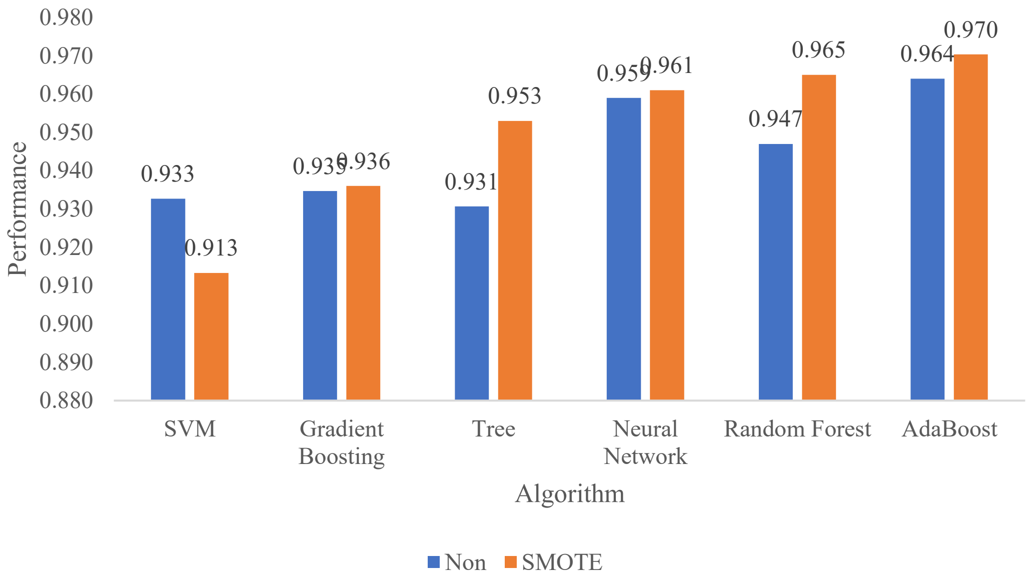


Figure 3. A Comparative Non-SMOTE, and SMOTE

### 3.2. Feature Selection Evaluation

Table 5. Best rank from feature selection techniques

| IG | | GR | | GD | | CS | | Relief-F | | FCBF | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| f | w | f | W | f | w | f | w | f | w | f | w |
| RM | 0.324 | RM | 0.323 | RM | 0.205 | RM | 546.962 | PA | 0.324 | RM | 0.483 |
| Ag | 0.274 | BpL | 0.162 | Ag | 0.169 | Ag | 431.832 | Ag | 0.260 | Ag | 0.229 |
| BpL | 0.174 | Ag | 0.140 | BpL | 0.111 | Hbp | 132.367 | Str | 0.205 | Bpl | 0.203 |
| Hbp | 0.114 | Hbp | 0.130 | Hbp | 0.076 | Pre | 130.565 | FD | 0.178 | FD | 0.069 |
| FD | 0.064 | Pd | 0.087 | FD | 0.043 | Bpl | 59.510 | JF | 0.132 | Pre | 0.043 |
| Str | 0.061 | FD | 0.065 | Str | 0.040 | FD | 50.562 | Bpl | 0.122 | Hbp | 0.000 |
| Pre | 0.042 | Pre | 0.041 | Pre | 0.028 | Pd | 18.701 | Pre | 0.116 | Str | 0.000 |
| PA | 0.014 | Str | 0.036 | PA | 0.009 | BM | 11.686 | BM | 0.107 | Pd | 0.000 |
| Pd | 0.012 | UF | 0.011 | Pd | 0.008 | UF | 11.046 | Gen | 0.104 | UF | 0.000 |
| UF | 0.010 | JF | 0.007 | UF | 0.007 | Str | 10.735 | RM | 0.077 | PA | 0.000 |
| BM | 0.010 | PA | 0.007 | BM | 0.006 | PA | 8.221 | SS | 0.072 | BM | 0.000 |
| JF | 0.006 | BM | 0.005 | JF | 0.004 | Alc | 3.058 | Sle | 0.064 | JF | 0.000 |
| Sle | 0.003 | Alc | 0.003 | Sle | 0.002 | Gen | 0.620 | Hbp | 0.058 | Alc | 0.000 |
| Alc | 0.002 | Sle | 0.002 | Alc | 0.002 | SS | 0.393 | Smo | 0.054 | Sle | 0.000 |
| Gen | 0.001 | Gen | 0.001 | Gen | 0.001 | JF | 0.239 | Alc | 0.044 | Gen | 0.000 |
| SS | 0.001 | SS | 0.000 | SS | 0.000 | Smo | 0.180 | Pd | 0.016 | SS | 0.000 |
| Smo | 0.000 | Smo | 0.000 | Smo | 0.000 | Sle | 0.009 | UF | 0.014 | Smo | 0.000 |

We conducted feature selection using several techniques, namely Information Gain (IG), Gain Ratio (GR), Gini Decrease (GD), Chi-Square (CS), Relief-F, and FCBF. The result of feature selection is the weight (w) of the best feature (f) related to the performance of the algorithm. The best features are displayed in Table 5.

Feature selection techniques produce different ranks. Feature selection using IG Technique results in 16 best features, and one feature "smo" has no impact on the performance of the ML algorithm. GR

**Jurnal Teknik Informatika (JUTIF)**
P-ISSN: 2723-3863
E-ISSN: 2723-3871
Vol. 6, No. 4, Agustus 2025, Page. 2625-2637
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2025.6.4.5166

and GD Techniques yield 15 features that have influence except for features “ss”, and “smo”. CS Technique results in 14 best features, and 3 features have no influence. Relief-F Technique produces 15 best features, and FCBF technique generates 5 best features and 12 features that do not influence the performance of the ML algorithm. Furthermore, we conduct an evaluation of the ML algorithm's performance to obtain a comprehensive picture after feature selection using several techniques. The performance of the ML algorithm using feature selection techniques is shown in Table 6.

Table 6. A Compare the accuracy of ML algorithms using feature selection

| Algorithm | All Features (%) | IG | GR | GD | CS | Relief-F | FCBF |
|---|---|---|---|---|---|---|---|
| Random Forest | 0.965 | 0.968 | 0.967 | 0.967 | 0.963 | 0.964 | **0.894** |
| Neural Network | 0.961 | 0.960 | 0.959 | 0.959 | 0.950 | 0.961 | 0.855 |
| Tree | 0.953 | 0.955 | 0.944 | 0.944 | 0.948 | 0.947 | 0.876 |
| Gradient Boosting | 0.936 | 0.934 | 0.933 | 0.934 | 0.946 | 0.940 | 0.866 |
| SVM | 0.912 | 0.918 | 0.919 | 0.919 | 0.911 | 0.925 | 0.773 |
| AdaBoost | **0.970** | **0.971** | **0.971** | **0.970** | **0.970** | **0.971** | 0.892 |
| Average | **0.949** | **0.951** | **0.949** | **0.949** | **0.948** | **0.951** | **0.859** |

The choice of feature selection techniques affects the performance of the algorithm. The performance of the AdaBoost algorithm consistently shows the best results, using all features and feature selection. The application of feature selection techniques IG, GR, and Relief-F has proven to provide an increase of 0.1 or 0.971, while the feature selection techniques GD, CS, and FCBF do not provide any contribution. Overall, the IG, and Relief-F feature selection techniques show the best results. Furthermore, we conducted a comparative analysis on the precision values displayed in Figure 4.

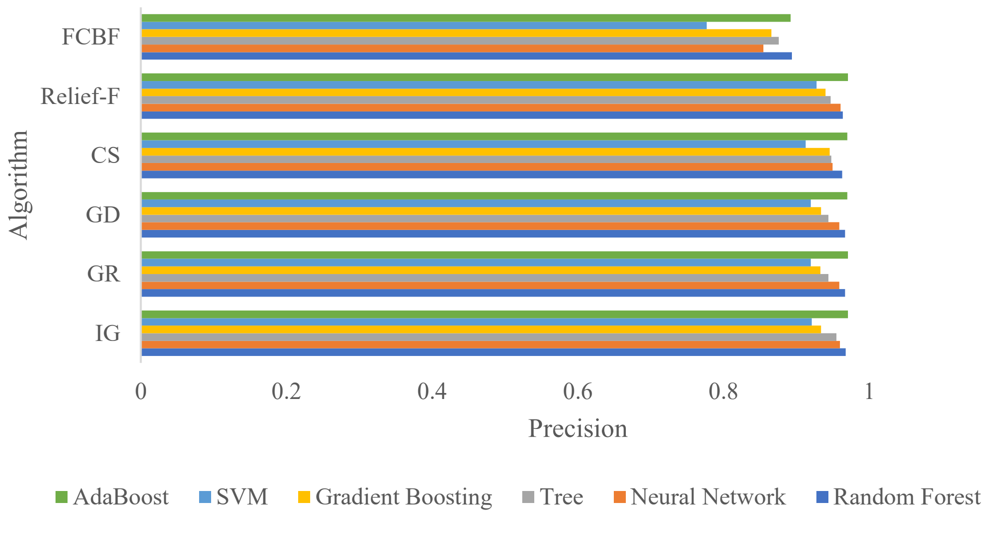


Figure 4. A Comparison of ML algorithm performance on precision values

The use of feature selection techniques in ML algorithms results in different precision. We found that the combination of AdaBoost algorithm + feature selection techniques (IG or Relief-F) is the best, while the combination of SVM + FCBF is the worst. In addition, we conducted a comparative analysis of recall displayed in Figure 5.

**Jurnal Teknik Informatika (JUTIF)**
P-ISSN: 2723-3863
E-ISSN: 2723-3871

Vol. 6, No. 4, Agustus 2025, Page. 2625-2637
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2025.6.4.5166

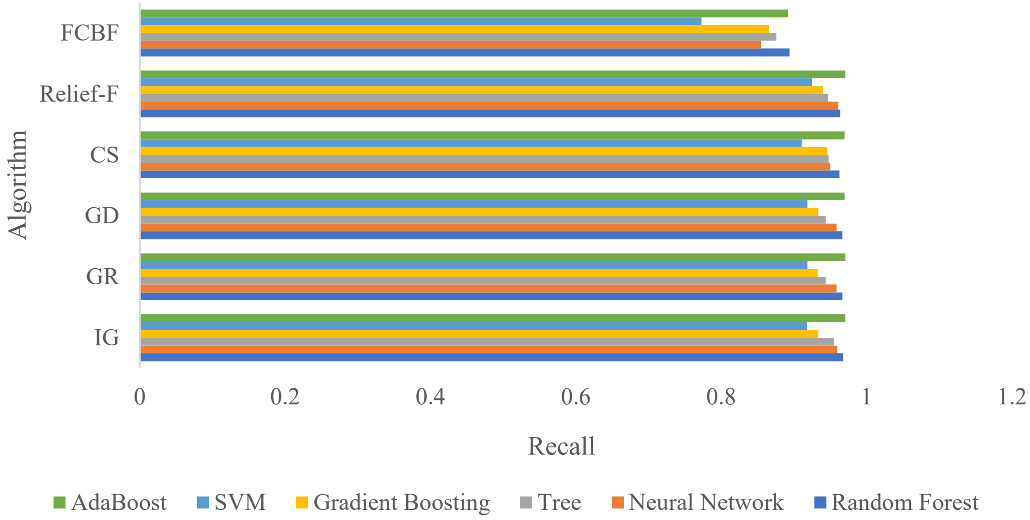


Figure 5. A Comparison of ML algorithm performance on recall values

The comparison results in the figure above are almost the same as the comparison results in the previous figure, where the combination of AdaBoost + (IG, GR, Relief-F) consistently performed the best and SVM+FCBC performed the worst. Next, we conducted an analysis of the time required by ML to build the model. The time taken by the algorithms is shown in Figure 6.

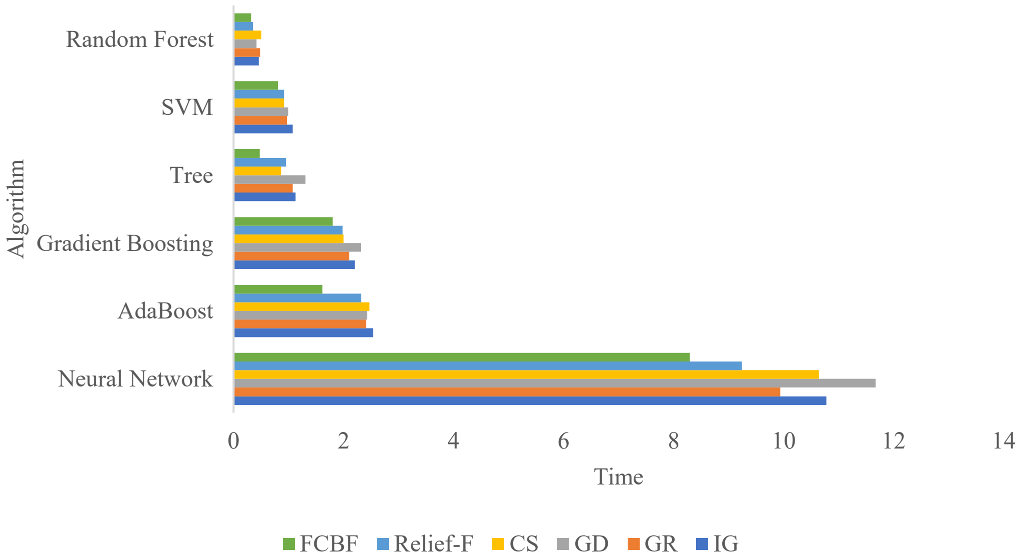


Figure 6. A Comparison of algorithm performance based on model building time

The time required for ML algorithms and feature selection techniques in building models varies. In this case, we found that the RF algorithm, and FCBF had the best time. Although in terms of accuracy, precision, and recall the FCBF technique had the worst performance, in terms of time, this technique was the fastest. Meanwhile, the NN+GD algorithm was the worst combination.

### 3.4. T-Test

We conducted a t-test using SPSS. A t-test is a statistical test used to compare the means of two sets of data and determine whether the observed differences are statistically significant or simply due to chance. Using the t-test, we found that the independent variables that had a strong influence on the

Vol. 6, No. 4, Agustus 2025, Page. 2625-2637
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2025.6.4.5166

dependent variable (diabetes) were age and BMI. The mean for age was 46.81 and BMI was 25.90, with a mean of 6.48 for all variables. The standard deviations for age were 10.2, BMI 5.33, and 1.68 for all variables, respectively. We also tested the standard error of the mean, which was 0.049.

### 3.5. Discussion

We proposed a combination of FS techniques, and imbalance class methods on the T2DM dataset. We found that the combination of AdaBoost+SMOTE+(IG or Relief-F) is the best formation with an accuracy level of 0.971 or 97.1%. We compared this success with the results of other studies, which we present in Table 7.

Table 7. Performance ML pada T2DM dataset

| No | Authors | Years | Algorithm | Accuracy (%) |
|---|---|---|---|---|
| 1 | Joshi et al. [27] | 2021 | Logistic Regression (LR) | 78.26 |
| 2 | Fazakiz et al.[28] | 2021 | Weighted Voting LRRFs+NB, DT, ANN | 88.4 |
| 3 | Lu et al.[28] | 2022 | Logistic Regression, Naïve Bayes, DTree, Random Forest, XGBoost and ANN, k-NN, Support Vector Machine; | 91 |
| 4 | Rakibul et al. [29] | 2023 | k-NN, Random Forest (RF), SVM, DTree, Naïve Bayes (NB), and Histogram-Based Gradient Boosting (HBGB) | 90 |
| 5 | Islam et al. [30] | 2023 | Naïve Bayes (NB), J48, Multilayer Perceptron (MP), and Random Forest (RF) | 95.7 |
| 6 | Villanueva et al. [31] | 2023 | k-NN, Bernoulli Naïve Bayes (BNB), DTree, Logistic Regression (LR), and SVM | 79.6 |
| 7 | Agliata et al. [32] | 2023 | Grid Search, ADAM optimizer, NN | 86 |
| 8 | Lugner et al. [33] | 2024 | XGboost | 90 |
| **9** | **This research** | **2025** | **SMOTE+(IG/Relief-F)+AdaBoost** | **97.1** |

Table 7 shows that the approach we used performs better. We employed different approaches in conducting a comparative analysis of ML algorithms using a combination of over-sampling techniques (SMOTE), feature selection techniques, and classification algorithms. We found that the application of SMOTE+AdaBoost+(IG or Relief-F) consistently demonstrated the best performance. This result is expected to contribute new findings to ML science in the context of T2DM dataset classification. Strategies for handling class imbalance and feature selection should be an integral part of the medical classification system development pipeline, as they can affect the interpretability and overall detection accuracy.

## 4. CONCLUSION

Based on the results of the comparative analysis that has been conducted, it can be concluded that class balancing in the T2DM Dataset impacts the performance of the ML classification model. The SMOTE balancing technique has been proven effective in improving algorithm performance. Although not significant, accuracy increased by 0.6%, precision by 0.7%, and recall by 0.6%. The AdaBoost algorithm consistently provides the best results with an accuracy of 97%.

In addition, we performed feature selection using six techniques namely IG, GR, GD, CS, Relief-F, and FCBF to select features that have the highest weight correlation with the class. From the feature selection, we found differences. The feature selection using the IG technique resulted in 16 best features

Vol. 6, No. 4, Agustus 2025, Page. 2625-2637
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2025.6.4.5166

and one feature “smo” had no impact on the performance of the ML algorithm. The GR and GD techniques produced 15 features that had an impact except for the features “ss” and “smo”. The CS technique yielded 14 best features and 3 features had no impact. The Relief-F technique resulted in 15 best features, and the FCBF technique generated 5 best features and 12 features that had no impact on the performance of the ML algorithm. Based on the selected features, we conducted an evaluation of the ML algorithm to obtain comprehensive information on the performance of the algorithm.

The results of the evaluation of the T2DM dataset, which has undergone feature selection and class balancing, show that the combination of the AdaBoost algorithm + (IG or Relief-F) consistently performs the best. Additionally, we evaluated the precision values. We found that the combination of the AdaBoost algorithm + feature selection techniques (IG or Relief-F) remain the best, while the combination of the SVM algorithm + FCBF is the worst. Furthermore, we conducted a comparative analysis of the recall values. The evaluation results indicate that the combination of the AdaBoost algorithm + (IG, GR, Relief-F) consistently ranks the best, while SVM + FCBC ranks the worst. In addition, we conducted a comparison of the time taken by the ML algorithms in building the models. Based on the comparison results, we found that the Random Forest and FCBF algorithms have the best time. Although in terms of accuracy, precision, and recall the FCBF technique performed poorly, in terms of time, this technique was the fastest. Meanwhile, the NN+GD algorithm was the worst combination.

Strategies for handling class imbalance and feature selection must be an integral part of the pipeline for developing medical classification systems, as they can affect the interpretability and overall accuracy of detection. Although this research successfully demonstrates an improvement in the performance of type 2 diabetes classification models through the application of feature selection techniques and class balancing, there are several limitations that need to be noted. First, the availability and diversity of data pose a major challenge. The dataset used is sourced from one or several public repositories with a limited number of samples, which may not be representative of the global population. This can affect the generalization of the model when applied to data from clinically different demographic or geographic environments. Further research is recommended to evaluate the effectiveness of other imbalance handling techniques such as ensemble-based sampling or cost-sensitive learning, as well as to apply these approaches to larger and more complex clinical datasets.

## ACKNOWLEDGEMENT

The author wishes to express their deepest gratitude to the Faculty of Health at Universitas Aisyah Pringsewu (UAP) for providing access to anonymous Type 2 Diabetes patient data. We also thank the Machine Learning (ML) Research Group at UAP for their valuable input during the model development and evaluation phase. Special appreciation is given to the doctors whose clinical insights significantly contributed to the interpretation of diabetes prediction results.

## REFERENCES

[1] International Diabetes Federation, “International Diabetes Federation - Complications,” *Idf.Org*. p. 1, 2019. Accessed: Apr. 18, 2020. [Online]. Available: https://www.idf.org/aboutdiabetes/complications.html
[2] American Diabetes Association, “Pharmacologic Approaches to Glycemic Treatment: Standards of Medical Care in Diabetes-2020,” *Diabetes care*, vol. 43, no. January. pp. S98–S110, 2020. doi: 10.2337/dc20-S009.
[3] A. A. Yameny, “Diabetes Mellitus Overview 2024,” *J. Biosci. Appl. Res.*, vol. 10, no. 3, pp. 641–645, 2024, doi: 10.21608/jbaar.2024.382794.
[4] F. A. Ibrahim and O. A. Shiba, “Data Mining : WEKA Software ( an Overview ),” *J. Pure Appl. Sci.*, vol. 18, no. 3, pp. 54–58, 2019, [Online]. Available: www.Suj.sebhau.edu.ly

P-ISSN: 2723-3863 https://jutif.if.unsoed.ac.id
E-ISSN: 2723-3871 DOI: https://doi.org/10.52436/1.jutif.2025.6.4.5166

[5] H. Guan *et al.*, "The role of machine learning in advancing diabetic foot: a review," *Front. Endocrinol. (Lausanne).*, vol. 15, no. April, pp. 1–15, 2024, doi: 10.3389/fendo.2024.1325434.
[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] N. P. Tigga and S. Garg, "Prediction of Type 2 Diabetes using Machine Learning Classification Methods," *Procedia Comput. Sci.*, vol. 167, pp. 706–716, 2020, doi: https://doi.org/10.1016/j.procs.2020.03.336.
[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, p. 107253, 2023, doi: https://doi.org/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, p. 101661, 2024, doi: https://doi.org/10.1016/j.swevo.2024.101661.
[10] 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, p. 121193, 2025, doi: https://doi.org/10.1016/j.ins.2024.121193.
[11] 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.
[12] 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.
[13] 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.
[14] 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.
[15] 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.
[16] 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, p. 111380, 2024, doi: https://doi.org/10.1016/j.knosys.2024.111380.
[17] 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: https://doi.org/10.1016/j.jksuci.2021.11.004.
[18] 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.
[19] 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.
[20] Y. Zhang *et al.*, "Feature selection based on neighborhood rough sets and Gini index," *PeerJ Comput. Sci.*, vol. 9, p. e1711, 2023, doi: 10.7717/peerj-cs.1711.
[21] 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.
[22] Y. Liu, J. Zhang, and L. Ma, "A fault diagnosis approach for diesel engines based on self-adaptive WVD, improved FCBF and PECOC-RVM," *Neurocomputing*, vol. 177, pp. 600–611, 2016, doi: https://doi.org/10.1016/j.neucom.2015.11.074.
[23] N. Aggarwal *et al.*, "Mean based relief: An improved feature selection method based on ReliefF,"

Vol. 6, No. 4, Agustus 2025, Page. 2625-2637
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2025.6.4.5166

*Appl. Intell.*, vol. 53, no. 19, pp. 23004–23028, 2023, doi: 10.1007/s10489-023-04662-w.

[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: https://doi.org/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] R. D. Joshi and C. K. Dhakal, "Predicting Type 2 Diabetes Using Logistic Regression and Machine Learning Approaches," *International Journal of Environmental Research and Public Health*, vol. 18, no. 14. 2021. doi: 10.3390/ijerph18147346.

[28] N. Fazakis, O. Kocsis, E. Dritsas, S. Alexiou, N. Fakotakis, and K. Moustakas, "Machine Learning Tools for Long-Term Type 2 Diabetes Risk Prediction," *IEEE Access*, vol. 9, pp. 103737–103757, 2021, doi: 10.1109/ACCESS.2021.3098691.

[29] R. Islam, A. Sultana, M. N. Tuhin, M. S. H. Saikat, and M. R. Islam, "Clinical Decision Support System for Diabetic Patients by Predicting Type 2 Diabetes Using Machine Learning Algorithms," *J. Healthc. Eng.*, vol. 2023, no. 1, p. 6992441, Jan. 2023, doi: https://doi.org/10.1155/2023/6992441.

[30] M. M. Islam *et al.*, "Identification of the risk factors of type 2 diabetes and its prediction using machine learning techniques," *Heal. Syst.*, vol. 12, no. 2, pp. 243–254, Apr. 2023, doi: 10.1080/20476965.2022.2141141.

[31] O. Iparraguirre-Villanueva, K. Espinola-Linares, R. O. Flores Castañeda, and M. Cabanillas-Carbonell, "Application of Machine Learning Models for Early Detection and Accurate Classification of Type 2 Diabetes," *Diagnostics*, vol. 13, no. 14. 2023. doi: 10.3390/diagnostics13142383.

[32] A. Agliata, D. Giordano, F. Bardozzo, S. Bottiglieri, A. Facchiano, and R. Tagliaferri, "Machine Learning as a Support for the Diagnosis of Type 2 Diabetes," *International Journal of Molecular Sciences*, vol. 24, no. 7. 2023. doi: 10.3390/ijms24076775.

[33] M. Lugner, A. Rawshani, E. Helleryd, and B. Eliasson, "Identifying top ten predictors of type 2 diabetes through machine learning analysis of UK Biobank data," *Sci. Rep.*, vol. 14, no. 1, p. 2102, 2024, doi: 10.1038/s41598-024-52023-5.