OPEN ACCESS 1

ISSN 2356-5462
https//journals.telkomuniversity.ac.id/ijoict

Intl. Journal on ICT
Vol. 12, No. 1, Jun 2026. pp. 1-13
doi.org/10.21108/ijoict.v12i1.10405

Improving the Accuracy of the C4.5 Algorithm in Diabetes Risk Prediction Using Bagging and Information Gain

Ernawati [1], Ade Candra [2*], Syahril Efendi [3]

[1,2,3] *Fakultas Ilmu Komputer dan Teknologi Informasi, Teknik Informatika, Universitas Sumatera Utara, Medan, Indonesia*.

[*]ade_candra@usu.ac.id

Abstract

Class imbalance is a common challenge in data classification, where the majority class significantly outnumbers the minority class, leading to a decrease in algorithm performance, particularly for the C4.5 algorithm. This study aims to address this problem by proposing a combination of Bootstrap Aggregation (Bagging) and Information Gain (IG). The IG method is employed for feature selection using a threshold of > 0.02 to select the most relevant attributes, while Bagging functions to enhance the stability and accuracy of the classification model. The experiment was conducted using a diabetes dataset from UCI with 10-fold cross-validation. The results showed that the C4.5+Bagging model achieved the highest accuracy at 95.96%, while the proposed C4.5+IG+Bagging combination reached an accuracy of 94.42%, a significant increase from the baseline C4.5 algorithm's accuracy of 89.04%. These findings demonstrate that the proposed method combination is effective in improving classification performance on imbalanced data.

Keywords: Algorithm C4.5, Class imbalance, Bagging, Information Gain, Accuracy.

I. INTRODUCTION

The digital era has produced an enormous volume of data that continues to grow exponentially. To extract valuable information from these large datasets, a process known as data mining is required. One of the most fundamental techniques in data mining is classification, which aims to build a model capable of predicting the class of an object whose label is unknown [1]. Classification models have been widely applied in various fields, including medicine, finance, and marketing, to support decision-making processes. However, the effectiveness of classification models is often hindered by several challenges, one of the most significant being the class imbalance problem. Class imbalance occurs when the distribution of data in one class (the majority class) is far greater than that in another class (the minority class) [2]. This condition causes many classification algorithms, including the C4.5 decision tree algorithm, to produce models that are biased toward the majority class. As a result, a model may exhibit high overall accuracy while failing to correctly identify instances of the minority class, which are often more critical, such as in disease detection or fraud identification [3]. In the medical context, for example, failure to identify patients suffering from a disease (the minority class) can have severe consequences, including delayed treatment and poorer prognosis, making this issue particularly important to address.

To overcome the class imbalance problem, two main approaches are commonly employed. The first is the data-level approach, which utilizes resampling techniques such as oversampling the minority class or *undersampling* the majority class. The second is the algorithm-level approach, which focuses on modifying the classification algorithm itself or combining multiple classifiers into an ensemble model to increase sensitivity

Received on 20 Dec 2025. Revised on 18 Jan 2026. Accepted and Published on 19 Mar 2026.

toward the minority class [4]. This study adopts the algorithm-level approach by proposing a combination of two methods, namely Bootstrap Aggregation (Bagging) and Information Gain (IG). The C4.5 algorithm is inherently vulnerable to imbalanced data because its splitting criterion, information gain, tends to select attributes that best separate the majority class, thereby neglecting patterns in the minority class. Bagging is an ensemble method that has proven effective in improving model stability and accuracy and has demonstrated good performance in handling noisy and imbalanced data [5]. Meanwhile, Information Gain is used as a feature selection method to reduce data dimensionality by selecting only the most relevant and influential attributes, thereby simplifying the model and reducing computational complexity[6]. The hypothesis of this study is that the combination of IG-based feature selection and the application of Bagging can significantly improve the performance of the C4.5 algorithm in classifying imbalanced data.

Several previous studies have addressed the class imbalance problem. Ruangthong and Jaiyen (2016) proposed a hybrid ensemble model combining decision trees and Bayesian Networks for bank marketing data. Their results showed that data balancing by enhancing the minority class was effective, and the ensemble model produced high and efficient classification results [7]. Another study by Bader-El-Den et al. (2019) employed the BRAF algorithm with Random Forest to address class imbalance across nine datasets from the UCI Repository and demonstrated the effectiveness of the proposed method. BRAF is an ensemble method that combines Random Forest with Bayesian techniques, which allows for more accurate probabilities in predictions by addressing class imbalance, making it relevant for datasets that have an unbalanced class distribution, such as the one in this study[3]. Focusing on the Bagging method, Muslim et al. (2018) applied it in conjunction with the C4.5 algorithm for credit card risk prediction. Their findings showed that Bagging increased accuracy from 70.5% to 75.1%, indicating the strong potential of this method [1]. Arfiani and Rustam (2019) also found that Bagging achieved very high accuracy, reaching 100%, in ovarian cancer classification, reinforcing its effectiveness in the medical domain [2]. Similarly, Fakhruzi (2018) demonstrated that the performance of Neural Networks improved when combined with Bagging to handle class imbalance across three different datasets, even when implemented using MapReduce for scalability [4].

Although previous studies have demonstrated the effectiveness of Bagging and feature selection methods independently, a research gap remains in exploring the specific combination of Information Gain and Bagging to enhance the performance of the C4.5 algorithm on imbalanced medical data. In this study, a diabetes dataset from the UCI Machine Learning Repository is used, consisting of 520 instances and 17 clinical attributes. This dataset exhibits an imbalanced class distribution, with 320 positive instances (diabetes) and 200 negative instances (non-diabetes), making it a relevant and challenging case study. This study proposes an approach that combines Information Gain and Bagging within the C4.5 algorithm to address class imbalance in medical data. While several studies have used feature selection and ensemble techniques in similar contexts, this study distinguishes itself by directly integrating both Bagging and Information Gain techniques within C4.5. This approach focuses not solely on feature selection or ensemble use alone, but rather combines both within a single framework, allowing the model to reduce variance through Bagging while filtering out irrelevant features with Information Gain before classification. This contrasts with other filtering approaches such as Relief or LASSO, which are more complex and often require additional parameter selection or wrapper methods like Recursive Feature Elimination (RFE), which require iterative model evaluation. In terms of ensembles, many studies have favored techniques like Random Forest or Gradient Boosting, which, while more complex, do not always offer the same transparency as C4.5 models, which are easier for medical practitioners to understand and interpret. The dataset used in this study has 320 positive (diabetic) instances and 200 negative (non-diabetic) instances, with a class imbalance ratio of approximately 1.6:1 (61.5% vs. 38.5%).

While this distribution is not extreme compared to many other medical datasets, it remains a challenge in the context of the C4.5 algorithm, which tends to favor attributes that occur more frequently in the majority (non-diabetic) class, thereby reducing the accuracy of identifying the minority (diabetic) class. This class imbalance, while moderate, can still impact model performance, especially in medical applications where detection of rarer but crucial diseases, such as diabetes, is crucial. Thus, the novelty of the proposed approach lies in combining both Information Gain and Bagging techniques in a single, more integrated and transparent scheme, which provides the benefit of simpler and more interpretable feature selection compared to more complex ensemble methods.

This research seeks to address the gap by investigating how IG-based feature selection with a threshold greater than 0.02 can filter important attributes before the data is processed by the Bagging C4.5 model, which is expected to produce a more accurate and efficient classifier.

Based on the background, problem formulation, and literature review presented, the objective of this study is to improve the accuracy of the C4.5 algorithm by applying a combination of Bootstrap Aggregation (Bagging) and Information Gain to address class imbalance in a diabetes dataset. In addition, this study aims to compare the performance of the proposed model (C4.5 + IG + Bagging) with the baseline C4.5 model, C4.5 with IG, and C4.5 with Bagging separately. This research is expected to contribute to an effective alternative method for minimizing classification errors in imbalanced data and to serve as a reference for other researchers facing similar challenges, particularly in the healthcare domain.

II. LITERATURE REVIEW

Decision tree algorithms, particularly C4.5, are widely used for heart disease classification due to their high interpretability and ease of implementation. A recent study focusing on the use of C4.5 for heart disease classification demonstrated an accuracy of approximately 73%, underscoring the importance of calculating entropy and information gain in decision tree formation, as well as applying pruning to reduce overfitting. Other studies have also shown that decision trees remain competitive, especially when combined with appropriate feature selection, which further improves accuracy. These findings provide a strong basis for suggesting that C4.5 is a reasonable baseline for heart disease prediction, but there is still significant room for accuracy improvement, which aligns with the goal of this study to improve the performance of the C4.5 algorithm [8].

The existing literature strongly supports the effectiveness of using bagging in improving the performance of decision tree models, including C4.5, as well as other weak classifiers in heart disease prediction. Decision tree-based ensemble models, such as bagging and Random Forest, often achieve very high accuracy up to around 99% on combined datasets when combined with appropriate feature selection [9]. A specific study focusing on the use of ensembles on heart disease datasets showed that bagging and boosting can improve the accuracy of a weak classifier by up to around 7%, and provide an average improvement of around 2% compared to a baseline model without ensembles [10], [11]. In several studies, bagging with decision trees outperformed other algorithms such as KNN, SVM, Naive Bayes, and other ensemble algorithms. Research in the medical field, including on the prediction of stroke, diabetes, breast cancer, and dengue fever, also shows that the combination of C4.5 and bagging consistently improves accuracy by around 2-5% compared to using C4.5 alone. These findings provide strong theoretical justification for this study design, demonstrating that applying bagging over C4.5 almost always reduces variance and improves prediction stability.

From this literature review, several key points can be extracted that are relevant to the objectives of this article. First, pure C4.5 has moderate accuracy (around 70-80%) for heart disease prediction, indicating that while C4.5 can be a good baseline, there is still significant room for improvement. This accuracy improvement can be achieved by combining C4.5 with other techniques, such as bagging. Many studies have shown that using bagging on top of decision trees or C4.5 can improve accuracy, reduce variance, and address data imbalance, which aligns with the hypothesis that C4.5 combined with bagging will yield better results than C4.5 alone. Furthermore, information gain serves as a core mechanism for attribute selection in C4.5, and its use in global feature selection, such as in the preprocessing stage (selecting the best feature subset before bagging), can simplify the model and reduce noise. However, mixed results in some studies applying information gain suggest that its contribution to accuracy improvement is not always significant, opening the opportunity for this article to explore in more detail information gain threshold selection schemes and when feature selection can help or hurt the model. This research can fill a gap in the literature, considering that many studies focus on Random Forest, Gradient Boosting, XGBoost, or generic bagging, while research explicitly using C4.5 as a base learner in a bagging + information gain-based feature selection scheme in heart disease, especially in the Indonesian context, is relatively rare. Furthermore, this research can offer a simpler and more explainable alternative to clinical practitioners, while still being effective. A systematic analysis of the extent of the additional accuracy contribution derived from bagging compared to feature selection in C4.5 in the heart disease domain, could provide significant added value to the analysis of this research.

This literature review demonstrates several general frameworks that can be emphasized in this study, ranging from the importance of early heart disease prediction and the role of machine learning, particularly C4.5, in classification, to the weaknesses of standard C4.5, such as the risk of overfitting and sensitivity to noise. The use of bagging as a solution to reduce variance and improve accuracy in decision trees or weak learners is supported by extensive empirical evidence. Therefore, the hypothesis that C4.5 plus bagging will outperform C4.5 alone is supported by theory and previous experimental results. Furthermore, information gain, as a solution for selecting relevant attributes in C4.5, can improve classification models, but its effect is limited in some cases, necessitating further study of its contribution to feature selection schemes. This study aims to address this gap by combining C4.5, bagging, and information gain-based feature selection in heart disease prediction and evaluating the accuracy improvement compared to the baseline C4.5 model and other models that use feature selection techniques separately.

III. RESEARCH METHOD

This study was systematically designed to address predefined research questions. The methodology follows a structured workflow, starting from data collection through model evaluation, to ensure that the results obtained are valid and reliable. Overall, the research framework aims to measure how effective the combination of Bootstrap Aggregation (Bagging) and Information Gain (IG) is in improving the performance of the C4.5 algorithm on imbalanced datasets.

Research Stages

The research stages describe the sequence of steps undertaken to achieve the research objectives. This workflow begins with data preparation and ends with the analysis of final results. Figure 1 visually illustrates the overall process carried out in this study.

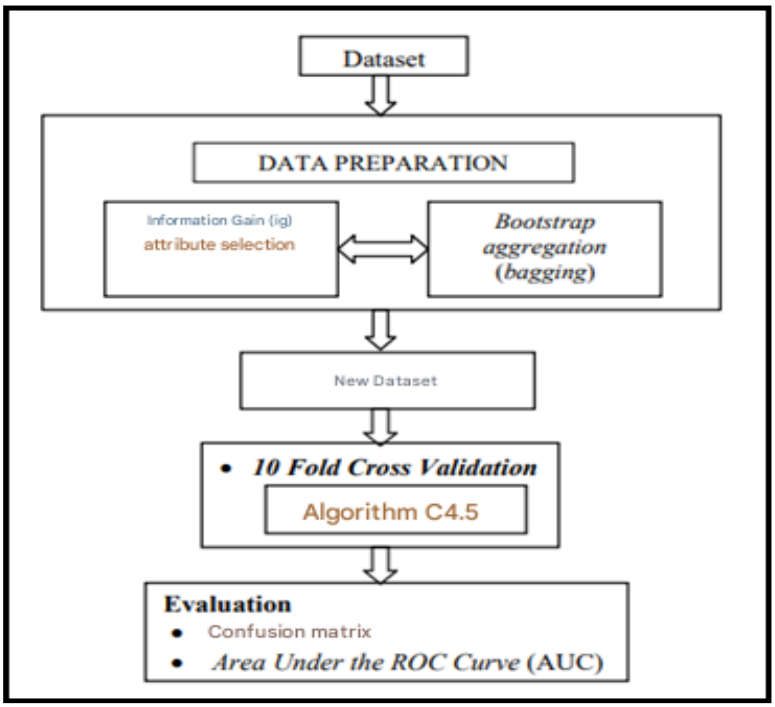

Figure 1. Research Stages

Based on Figure 1, the first stage is Data Preparation. In this stage, the dataset used in the study namely, the diabetes dataset from the UCI Machine Learning Repository—is collected. The characteristics of the dataset, such as the number of attributes, data types, and class distribution, are then examined to understand the level of class imbalance [12].

The second stage is Attribute Selection using Information Gain (IG). The objective of this stage is to reduce data dimensionality and select attributes that have the most significant influence on the target class. In this study, Information Gain (IG)-based feature selection uses a fixed threshold of >0.02 to select the most relevant attributes and remove attributes with lower IG scores. This threshold was chosen based on previous experiments showing that attributes with IG scores lower than 0.02 do not contribute significantly to predictions and tend to increase model complexity without improving accuracy. This threshold value is not the result of trial and error but rather based on domain knowledge and experimental results on similar datasets, which show that it is effective in identifying relevant features. Thus, the selection of this IG threshold is considered to improve model quality without adding irrelevant attributes that can cause overfitting or worsen model performance on imbalanced data.

The third stage, Modeling and Classification, is the section that illustrates the use of the C4.5 algorithm with the addition of Bagging and Information Gain to improve prediction accuracy. In the diagram, this can be represented by boxes indicating the steps in which data from which relevant features have been selected is processed using a classification model. The evaluated models include: (1) the C4.5 algorithm as the baseline model, (2) C4.5 with IG-based feature selection (C4.5 + IG), (3) C4.5 with the Bagging method (C4.5 + Bagging), and (4) the proposed model combining IG and Bagging (C4.5 + IG + Bagging). Each model is built using training data generated through validation techniques.

The final stage is Evaluation and Validation. The constructed models are tested using testing data. Evaluation is conducted using 10-fold cross-validation to ensure robust results and to avoid overfitting. The evaluation metrics include the Confusion Matrix to calculate accuracy, precision, and recall, as well as the Area Under the Curve (AUC) to measure overall classification performance.

Dataset and Algorithms

The research stages are implemented through a series of specific methods consisting of dataset description, classification algorithms, and evaluation techniques.

1. Dataset
   This study uses a public dataset available in the UCI Machine Learning Repository entitled the Early Stage Diabetes Risk Prediction Dataset. This dataset was selected due to its relevance to the class imbalance problem and its frequent use in medical classification research. The dataset consists of 520 instances (patient medical records) with 17 predictor attributes and 1 class attribute. The class attribute has two labels, namely Positive (diabetes) and Negative (non-diabetes), with an imbalanced distribution of 320 positive instances and 200 negative instances. All predictor attributes are categorical in nature.
2. C4.5 Algorithm
   The C4.5 algorithm is an extension of the ID3 (Iterative Dichotomiser 3) algorithm used to construct decision trees. In this study, the C4.5 algorithm was implemented to build decision trees used in data classification. C4.5 selects the attribute with the highest Information Gain to split the dataset, constructing a decision tree by selecting the branch that optimally separates the data. This decision tree is built recursively until it reaches a stopping condition, such as the maximum depth of the tree or the minimum number of samples per leaf. To improve model performance, the Bagging technique was applied by building an ensemble of 100 decision trees. Each tree was trained on a random subset of data using bootstrap sampling, where the training data for each iteration was different, and the final prediction was calculated by majority voting of the prediction results of all trees in the ensemble. Implementation details of the C4.5 decision tree in this study include setting the maximum tree depth to 10 to avoid overfitting and keep tree complexity under control. In addition, the minimum sample per leaf was set at 5 samples to prevent the formation of too few tree branches and keep the model more general. The attributes used for data splitting at each node were selected based on Information Gain, by selecting the attribute that provided the best separation. By providing detailed information regarding this implementation, it is hoped that it will improve the reproducibility of the experiment, as well as help the understanding and application of this method to similar problems in other studies [13].
3. Information Gain (IG) for Feature Selection
   Information Gain (IG) is a feature selection method that measures how much information an attribute provides about a class. IG calculates the reduction in entropy after a dataset is partitioned based on a specific attribute. In this study, IG was used to select the most relevant features. Attributes with IG

values below a threshold of 0.02 were considered insignificant and were removed from the dataset to simplify the model and reduce noise [7].

4. Bootstrap Aggregating (Bagging)
   Bagging is an ensemble learning method designed to improve the stability and accuracy of classification algorithms. The principle of Bagging is to generate multiple classification models from different subsets of the training data created through bootstrap sampling (random sampling with replacement). Each model is trained independently, and the final prediction is determined through a voting process across all models. This method effectively reduces variance and prevents overfitting, while also increasing the likelihood that the minority class becomes dominant in some data subsets [14].
5. Model Evaluation
   To measure the performance of the developed models, several standard evaluation metrics are employed. Model testing is conducted using 10-fold cross-validation, in which the dataset is divided into 10 parts; 9 parts are used for training and 1 part for testing, and this process is repeated 10 times [15]. The evaluation results are analyzed using a Confusion Matrix, which presents four values: True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN). From these values, accuracy, precision, and recall are calculated.[16] In addition, the Area Under the Curve (AUC) of the ROC curve is used to provide a comprehensive assessment of model performance, where an AUC value close to 1.0 indicates excellent classification performance [17].

IV. RESULTS AND DISCUSSION

The results of a series of experiments conducted to address the research objectives are presented in this section. The results are reported systematically, beginning with the baseline C4.5 model, followed by models with the gradual addition of methods (Information Gain and Bagging), and finally the proposed combined model. Each result is analyzed and discussed to understand the impact of each applied method in addressing the class imbalance problem in the diabetes dataset. Model performance is evaluated using the Confusion Matrix, Accuracy, Precision, Recall, and Area Under the Curve (AUC) metrics, employing a 10-fold cross-validation technique.

Analysis of Classification Model Performance

The performance analysis is carried out by comparing four main models: C4.5 (as the baseline), C4.5 + Information Gain (IG), C4.5 + Bagging, and C4.5 + IG + Bagging (the proposed model). The purpose of this analysis is to measure the extent to which each method contributes to improvements in accuracy and other evaluation metrics.

Performance of the Baseline C4.5 Model

Before applying any additional methods, the first step is to build a classification model using the standard C4.5 algorithm. This model serves as the baseline or reference point for comparing the performance of the modified models. The baseline C4.5 model is trained using all available features in the dataset, without feature selection or ensemble techniques. The performance of the baseline C4.5 model is evaluated using a Confusion Matrix, as presented in Table 1.

TABLE I
CONFUSION MATRIX OF THE BASELINE C4.5 MODEL

	Actual Positive	**Actual Negative**
Predicted Positive	292	29
Predicted Negative	28	171

Based on Table 1, the performance metrics of the model can be calculated. The model correctly classified 171 positive instances (True Positives) but failed to identify 29 other positive instances (False Negatives). For the negative class, the model correctly classified 292 instances (True Negatives) and incorrectly classified 28 instances as positive (False Positives). From these results, the model achieved an accuracy of 89.04%, a precision of 85.93%, and a recall of 85.50%. Although the overall accuracy appears relatively high, the recall

value of 85.50% indicates that the model still has limitations in identifying all instances of the minority (positive) class, which is a critical issue in class imbalance scenarios. To further understand how the model makes decisions, the decision tree generated by the C4.5 algorithm can be analyzed. Figure 2 illustrates the structure of the decision tree formed from the diabetes dataset.

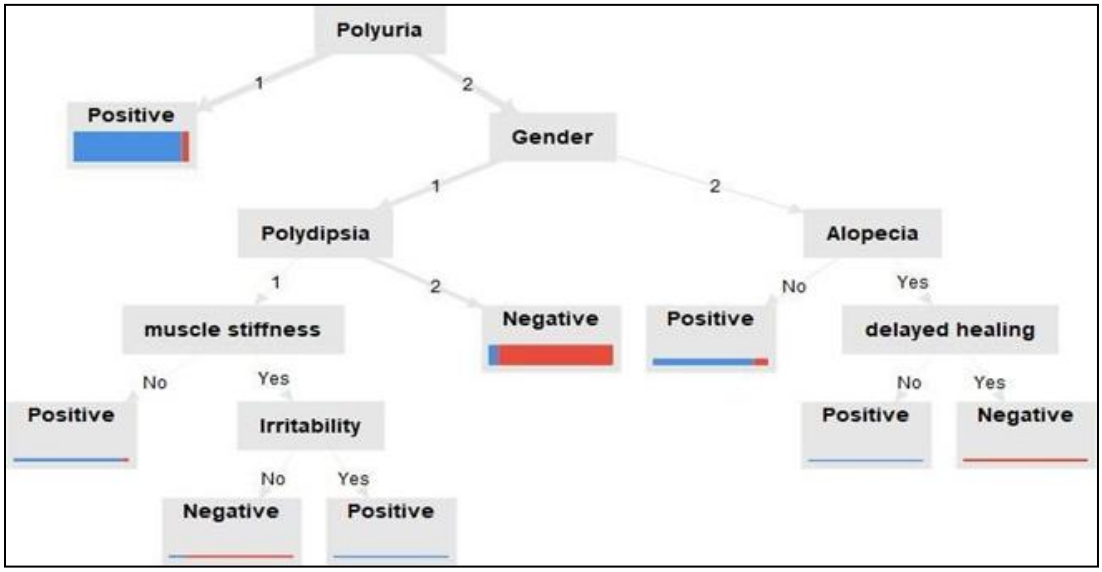

Figure 2. The Decision Tree

In Figure 2 showing the Decision Tree, it can be seen that only 7 features are actually used in the learned tree, even though earlier in the text it was mentioned that the base C4.5 model was trained using all available features in the dataset, without any feature selection or ensemble techniques. Whereas, in fact, although C4.5 uses all available features in the dataset to build the decision tree, the decision tree formation process naturally selects the most relevant attributes at each splitting step based on Information Gain. Therefore, even though all features are considered in the data split, the resulting decision tree may only use a subset of the available features, depending on how those attributes affect class separation. In this case, although the C4.5 model initially used all attributes, only 7 features were effectively used in the resulting decision tree, indicating that less informative attributes were not selected for further splits in the tree.

TABLE II
RANKING OF INFORMATION GAIN SCORES FOR EACH ATTRIBUTE

Rank	Attribute	IG Score
1	Polyuria	0.362
2	Polydipsia	0.359
3	Gender	0.163
4	Sudden weight loss	0.149
5	Partial paresis	0.145
…	…	…
16	Itching	0.0001

The decision tree generated produces eight main rules. One of the most prominent is the first rule: If Polyuria = 1, then the result is Positive. This rule covers 243 positive instances and only 15 negative instances, indicating that the symptom polyuria (excessive urination) is the strongest indicator for diabetes prediction in this dataset.

Other rules involve combinations of symptoms such as Gender, Polydipsia, Alopecia, and others to refine the classification process. Although this decision tree is easily interpretable, its performance can still be improved, particularly in reducing misclassification errors in the minority class.

Impact of Information Gain Feature Selection

The next stage involves applying feature selection using the Information Gain (IG) method to reduce data dimensionality and focus on the most relevant attributes. The objective is to simplify the model and eliminate noise that may arise from insignificant attributes. This process calculates the IG score for each attribute and selects attributes with scores above the threshold of 0.02. The ranking of attributes based on their IG scores is presented in Table II.

TABLE III
CONFUSION MATRIX OF THE C4.5+IG MODEL

	Actual Positive	**Actual Negative**
Predicted Positive	171	29
Predicted Negative	27	293

Table III, only a marginal improvement in performance is observed. Accuracy increased slightly from 89.04% to 89.23%, precision rose to 86.36%, while recall remained unchanged at 85.50%. This insignificant improvement indicates that although some attributes were removed, the remaining features were already sufficiently representative for building the model. However, feature selection alone was not sufficient to substantially address the existing class imbalance problem.

Application of the Bagging Ensemble Method

Next, the Bootstrap Aggregating (Bagging) ensemble method was applied without prior feature selection. The purpose of this step was to evaluate the independent effect of Bagging in improving the stability and accuracy of the C4.5 model, particularly on imbalanced data. Bagging works by constructing multiple decision trees from different subsets of the training data and combining their predictions through a voting mechanism. The results of the C4.5 + Bagging model are shown in Table IV.

TABLE IV
CONFUSION MATRIX OF THE BASELINE C4.5 +BAGGING MODEL

	Actual Positive	**Actual Negative**
Predicted Positive	191	9
Predicted Negative	12	308

The results in Table IV demonstrate a very significant performance improvement. Accuracy increased to 95.96%, precision reached 94.09%, and most importantly, recall rose sharply to 95.50%. The number of False Negatives decreased from 29 to only 9, and False Positives were also reduced. This improvement aligns with the theory that Bagging is effective in handling class imbalance, as each bootstrap sample may slightly alter the class distribution, allowing the minority class to be better learned by some models. These results confirm that Bagging is a highly powerful method for this classification task.

Analysis of the Combined Model (C4.5 + IG + Bagging)

In the final stage, the proposed model combining all three methods—Information Gain feature selection followed by the Bagging ensemble applied to the C4.5 algorithm—was evaluated. The expectation was that feature simplification through IG would allow the Bagging models to be more focused and efficient, resulting in the best overall performance. The results of the C4.5 + IG + Bagging model are presented in Table V.

TABLE V
CONFUSION MATRIX OF THE BASELINE C4.5 +IG+BAGGING MODEL

	Actual Positive	Actual Negative
Predicted Positive	188	12
Predicted Negative	17	303

This combined model achieved an accuracy of 94.42%, with a precision of 91.71% and a recall of 94.00%. Interestingly, its accuracy is slightly lower than that of the C4.5 + Bagging model without feature selection (95.96%). One possible explanation is that the IG-based feature selection process may have removed certain attributes that individually had low IG scores but were collectively useful for creating diversity among the ensemble models. Such diversity is a key factor in the success of ensemble methods and reducing it may slightly degrade overall performance. Nevertheless, the combined model still significantly outperforms the baseline model and the model using IG alone.

Comprehensive Comparison and ROC/AUC Analysis

To provide a comprehensive overview, all models are compared side by side in this subsection using various evaluation metrics.

TABLE VI
PERFORMANCE COMPARISON OF ALL MODELS

Model	Accuracy	Recall	Precision
C4.5	89.04%	85.50%	85.93%
C4.5 + IG	89.23%	85.50%	86.36%
C4.5 + Bagging	95.96%	95.50%	94.09%
C4.5 + IG + Bagging	94.42%	94.00%	91.71%

Table VI clearly shows that the C4.5 + Bagging model delivers the best performance across almost all metrics, followed by the proposed C4.5 + IG + Bagging model. The application of Bagging provides the most substantial performance improvement. This comparison is also visualized using bar charts in Figure 3 to facilitate interpretation.

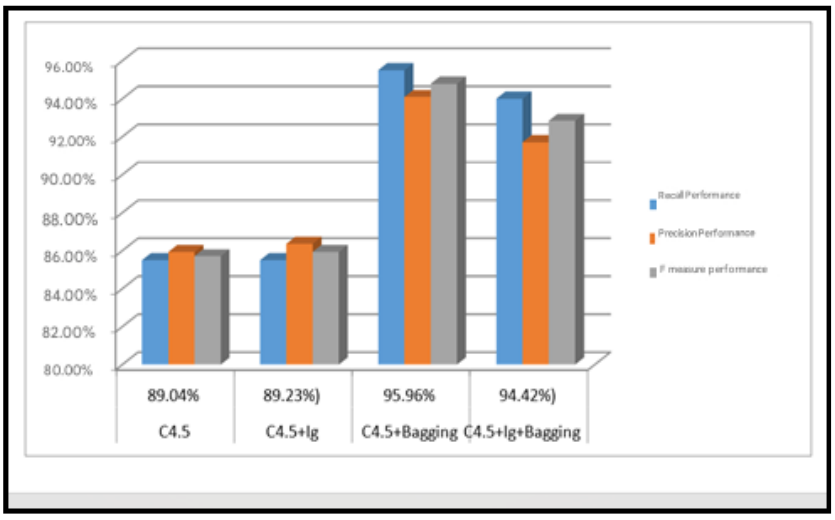

Figure 3. Comparison Chart of Model Accuracy

In addition to the metrics derived from the Confusion Matrix, analysis using the ROC (Receiver Operating Characteristic) curve and the AUC (Area Under the Curve) provides a more robust perspective on model performance, particularly for imbalanced datasets. The ROC curve illustrates the trade-off between the True Positive Rate (Recall) and the False Positive Rate. Figures 4, 5, and 6 present the ROC curves for each model.

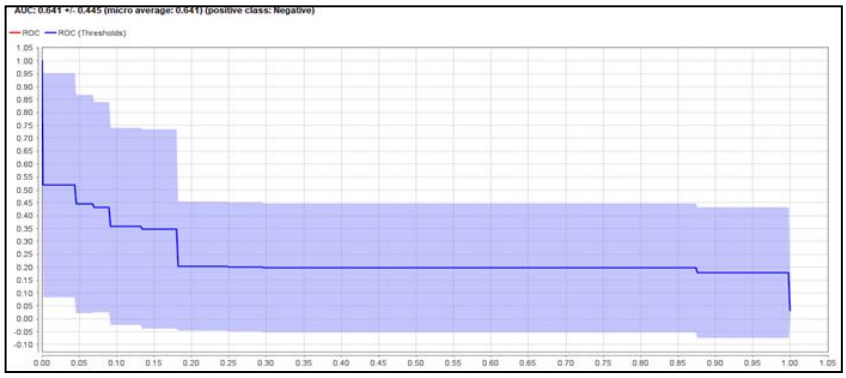

Figure 4. ROC Curve of the Baseline C4.5 Model

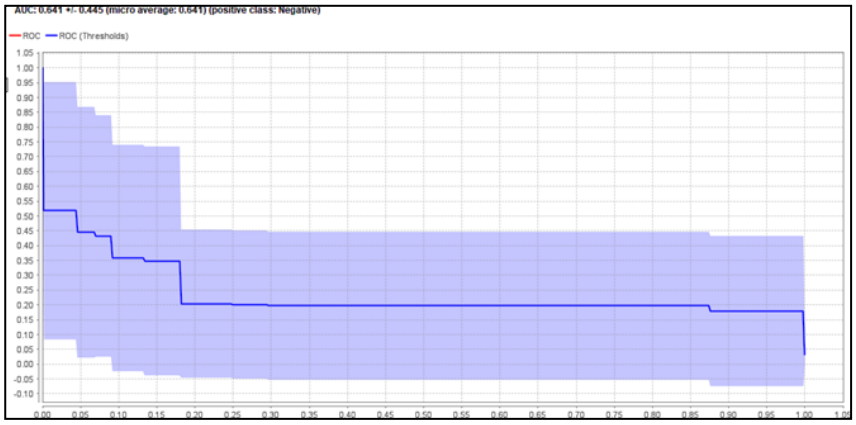

Figure 5. ROC Curve of the C4.5 + IG Model

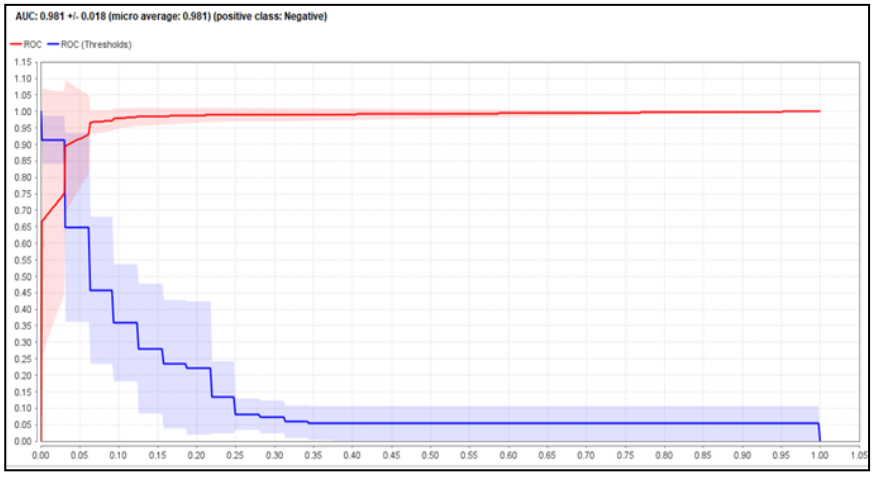

Figure 6. ROC Curve of the C4.5 + Bagging Model

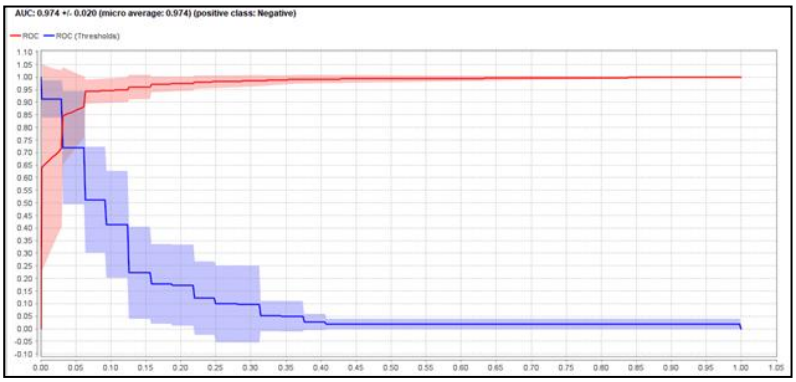

Figure 7. ROC Curve of the C4.5 + IG + Bagging Model

The ROC curves for the baseline C4.5 and C4.5 + IG models indicate relatively poor performance, with AUC values of 0.641, which are classified as poor classification. In contrast, the ROC curve for the C4.5 + Bagging model approaches the upper-left corner, achieving the highest AUC value of 0.981, which indicates excellent classification performance. A summary of the AUC values for all models is presented in Table VII.

TABLE VII
SUMMARY OF AUC VALUES FOR EACH MODEL

Model	AUC Value	Classification
C4.5	0.641	Poor
C4.5 + IG	0.641	Poor
C4.5 + Bagging	0.981	Excellent
C4.5 + IG + Bagging	0.972	Excellent

The AUC analysis reinforces the conclusion that the application of the Bagging method is the key factor in significantly improving classification performance on the imbalanced diabetes dataset. The C4.5 + Bagging model consistently demonstrates the best performance across all evaluation metrics, making it the most recommended model for similar cases. The proposed model (C4.5 + IG + Bagging) also exhibits very strong performance, although slightly inferior to the Bagging-only model. These findings provide valuable insight that, in certain cases, pre-ensemble feature selection does not necessarily enhance performance and may even slightly reduce the diversity required for ensemble learning.

V. CONCLUSION

This study has successfully addressed the main problem concerning the degradation of the C4.5 algorithm's performance due to class imbalance in a diabetes dataset. To overcome this challenge, a combined method integrating Information Gain (IG) feature selection and the Bootstrap Aggregation (Bagging) ensemble technique was proposed, with the aim of simplifying the model while improving its accuracy and robustness when dealing with imbalanced data. The experimental results show that the baseline C4.5 model achieved an accuracy of only 89.04% with an AUC value of 0.641, indicating poor classification performance, particularly in identifying the minority class. The application of Information Gain feature selection alone resulted in only a marginal improvement, suggesting that feature reduction by itself is insufficient to address the core class imbalance problem.

In contrast, the standalone application of the Bagging method produced a remarkable improvement, increasing accuracy to 95.96% and the AUC value to 0.981, which falls into the category of excellent classification performance. The proposed combined model, C4.5 + IG + Bagging, also demonstrated a

significant performance improvement, achieving an accuracy of 94.42% and an AUC of 0.972, although its performance was slightly lower than that of the C4.5 + Bagging model without feature selection. Based on these findings, it can be concluded that the combination of Bagging and Information Gain is effective in addressing the class imbalance problem in the C4.5 algorithm. However, this study also reveals that the primary contribution to the substantial performance improvement originates from the Bagging method itself, which plays a crucial role in enhancing recall and reducing misclassification errors in the minority class.

This research provides a practical contribution in the form of a reliable alternative solution for classifying imbalanced medical data, which can support medical practitioners in decision-making processes. Nevertheless, the finding that IG-based feature selection slightly reduced the performance of Bagging opens opportunities for future research, such as exploring alternative feature selection methods or different ensemble techniques to achieve more optimal results.

ACKNOWLEDGMENT

The authors would like to express their sincere gratitude to Dr. Eng. Ade Candra, S.T., M.Kom., as the primary supervisor, and Prof. Dr. Syahril Efendi, S.Si., M.IT., as the second supervisor, for their guidance, valuable suggestions, and continuous motivation throughout the completion of this research. The authors also thank the examiners for their constructive feedback. This research was made possible through the support of facilities provided by the Master's Program in Informatics Engineering, Faculty of Computer Science and Information Technology, Universitas Sumatera Utara.

REFERENCES

[1] H. B. Muslim, A. P. Wibawa, and A. Wibowo, "Improving accuracy of C4.5 algorithm using split feature reduction model and bagging ensemble for credit card risk prediction," in *Ii 2018 International Conference on Information and Communications Technology (ICOIACT*, 2018, pp. 141–145.

[2] A. Arfiani and Z. Rustam, "Ovarian cancer data classification using bagging and random forest," *AIP Conf. Proc.*, vol. 2168, no. 1, p. 20006, 2019.

[3] M. Bader-El-Den, E. Teitei, and T. Perry, "Biased Random Forest for Dealing with the Class Imbalance Problem," *IEEE Trans. Neural Networks Learn. Syst.*, vol. 30, no. 7, pp. 2163–2172, 2019.

[4] I. Fakhruzi, "An artificial neural network with bagging to address imbalance datasets on clinical prediction," in *2018 International Conference on Information and Communications Technology (ICOIACT*, 2018, pp. 895–898.

[5] M. Galar, A. Fernandez, E. Barrenechea, H. Bustince, and F. Herrera, "A review on ensembles for the class imbalance problem: Bagging-, boosting-, and hybrid-based approaches," *IEEE Trans. Syst. Man, Cybern. Part C (Applications Rev.*, vol. 42, no. 4, pp. 463–484, 2012.

[6] F. He, H. Yang, Y. Miao, and R. Louis, "A Hybrid Feature Selection Method Based on Genetic Algorithm and Information Gain," in *2016 International Conference on Computer Science and Network Technology (ICCSNT*, 2016, pp. 320–323.

[7] P. Ruangthong and S. Jaiyen, "Hybrid ensembles of decision trees and Bayesian network for class imbalance problem," in *2016 8th International Conference on Knowledge and Smart Technology (KST)*, 2016, pp. 39–42.

[8] E. Indra, K. Ho, R. Hakim, and D. Sitanggang, "Application of C4.5 Algorithm for Cattle Disease Classification," *J. Phys. Conf. Ser.*, vol. 1230, no. 1, p. 12070, 2019.

[9] T. M. Khoshgoftaar, J. Hulse, and A. Napolitano, "Comparing boosting and bagging techniques with noisy and imbalanced data," *IEEE Trans. Syst. Man, Cybern. - Part A Syst. Humans*, vol. 41, no. 3, pp. 552–568, 2011.

[10] J. Han, J. Pei, and M. Kamber, *Data Mining: Concepts and Techniques*, 3rd ed. Burlington, MA, USA: Morgan Kaufmann, 2011.

[11] L. Breiman, "Random Forests," *Mach. Learn.*, vol. 45, no. 1, pp. 5–32, 2001.
[12] N. V Chawla, K. W. Bowyer, L. O. Hall, and W. P. Keglmeyer, "SMOTE: Synthetic Minority Over-sampling Technique," *J. Artif. Intell. Res.*, vol. 16, pp. 321–357, 2002.
[13] S. García, J. Derrac, I. Triguero, and F. Herrera, "A survey of preprocessing techniques for imbalanced datasets," in *Data Preprocessing in Data Mining*, Springer, 2015, pp. 195–241.
[14] A. Fernández, S. García, F. Herrera, and N. V Chawla, "SMOTE for learning from imbalanced data: Progress and challenges, marking the 15-year anniversary," *J. Artif. Intell. Res.*, vol. 61, pp. 863–905, 2018.
[15] Z. Wang and Q. Li, "The Research of Imbalanced Data Set Sampling Based on K-Means and SMOTE," in *2020 IEEE International Conference on Artificial Intelligence and Computer Applications (ICAICA*, 2020, pp. 430–434.
[16] M. K. Duwairi and R. M. Abu-Errub, "Improved SMOTE for Mining Imbalanced Datasets," in *2021 International Conference on Information Technology (ICIT*, 2021, pp. 140–145.
[17] J. R. Quinlan, *C4.5: Programs for Machine Learning*. San Mateo, CA, USA: Morgan Kaufmann, 1993.