MEDINFTech JOURNAL MEDICAL INFORMATICS TECHNOLOGY
medinftech.org ISSN 2988-7003 (online)

Comparison of Classification Results of SVM, KNN, Decision Tree, and Ensemble Methods in Diabetes Diagnosis

Muhammad Iqbal Arsyad. H[1], *Ali Amran*[2*], *Anita Desiani*[3], *and Michael Jackson Napitu*[4]

[1,2,3,4] *Universitas Sriwijaya, Indonesia*

ARTICLE HISTORY

Received: 20 August 24
Final Revision: 07 September 24
Accepted: 09 September 24
Online Publication: 30 September 2024

KEYWORDS

Decision Tree, Ensemble, Diabetes, Support Vector Machine (SVM), K-Nearest Neighbor (KNN)

CORRESPONDING AUTHOR

ali_amran@mipa.unsri.ac.id

DOI

10.37034/medinftech.v2i3.62

ABSTRACT

This study aims to determine which algorithms and test techniques are the most optimal in detecting diabetes mellitus and obtaining the best results based on the value of accuracy, precision, and recall. In this study, approaches were used in early diagnosis of diabetes using KNN, SVM, Decision Tree, and Ensemble Majority Voting methods in Percentage Split and K-Fold Cross Validation methods. Diabetes is a disease characterized by high blood sugar (glucose) levels and can cause a variety of disease complications and damage to the body's organs if not treated immediately. Early diagnosis of diabetes is becoming crucial so that people can take immediate action to the hospital for immediate treatment. The data used is Healthcare-Diabetes from Kaggle. The results of this study have found that the K-Fold Cross Validation method is better because it can provide an average improvement in Ensemble accuracy of 13.42% compared to the Percentage Split method which only gives an average increase in Ensamble accuracy of 9.15%. The best algorithm for classifying diabetes disease is the Ensemble Majority Voting algorithm using the K-Fold Cross Validation method with a 98.81% accuracy rate. These excellent research results may contribute to detecting early symptoms of diabetes before it become too severe.

1. Introduction

Diabetes is one of the chronic diseases due to endogenous insulin secretion disorders and is characterized by higher blood sugar levels than normal blood glucose levels should be. If not treated immediately, diabetes can cause a variety of complications of diseases in organs such as eyes, blood vessels, and nerves [1]. The amount of movement, diet, and age also increases the risk of diabetes. The International Diabetes Federation (IDF) estimates that there are at least 463 million people aged 20-79 in the world suffering from diabetes in 2019 or equivalent to a prevalence rate of 9.3% of the total population at the same age. The prevalence of diabetes is estimated to increase as the age of the population rises to 19.9% or 111.2 million persons aged 66-79 years. The figure is predicted to rise to 578 million in 2030 and 700 million in 2045 [2].

Early-stage diagnosis is a prerequisite for diabetes to be treated properly immediately. Diagnosis of diabetes can be done by finding the relationship between patterns and tendencies through examination of a data set. Early detection of diabetes can use mathematical approaches to data mining to analyze data. Data mining has a variety of methods, one of which is classification that can be applied to early detection or prediction of diabetes. Classification is the process of finding a set of patterns or functions that describe and separate one data class from another and is used to predict data that does not yet have a particular data class [3]. Some of the algorithms that can be used in the classification of diabetes are the KNN, SVM, Decision Tree, and Ensemble Majority Voting.

KNN (K-Nearest Neighbor) is one of the most common and easy-to-use classification techniques. However, there is also a weakness of this method, which is the existence of k-bias values. Several previous studies have used KNN algorithms, such as Goddess Cahyanti, Alifah Rahmayani, and Syafira Ainy Husniar in the classification of breast cancer diseases that yield accuracy, precision and recall with performance values of around 93% [4]. However, the study only processed a small amount of data, 569 data. Muhammad Yunus and Ni Kadek Ari Pratiwi also applied KNN in the nutritional status forecast. However, the results of the study only processed a very small amount of data, 134 data, and yielded an accuracy value of 88,06% [5]. Other studies that used KNN for the classification of alcohol scent produced a precision of 96.4% for K=4 but were not done against data that amounted to thousands [6]. One of the shortcomings of the KNN method is the presence of a k-bias value that affects the value of the prediction result, so it is important to determine the value k that produces the best accuracy value [7]. Another method that can be used to classify diabetes is the Support Vector Machine (SVM) algorithm.

Unlike the KNN algorithm, one of the advantages of SVM is that it does not have a k-bias value. SVM is machine learning that has a function to separate data sets into different classes. A Support Vector Machine algorithm will group data that has the same characteristics into one class. SVM has high generalization capabilities without additional knowledge requirements, even with high dimensions of input space. SVM is a very useful technique for data classification and regression problems created [8]. There has been a previous study by Nofie Prasetiyo, Kiki Baihaqi, Santi Lestari, and Yana Cahyana using the SVM algorithm to classify plants affected by rat pests with 25% accuracy [9]. The disadvantage of the research is that in addition to the very low accuracy value obtained, the research also processes data in the form of images in a small amount. In addition, Dwi Sri Rahayu, Nursafika, Jihan Afifah, and Sri Intan also conducted research using the SVM algorithm in classifying diabetes mellitus with an accuracy of 82,01% [10]. However, the study only processed a small amount of data, 768 data. Other studies using SVM in the classification of repulsions produced accurations of 81.91%, but the datasets used are still few, that is, 470 data alone [11]. The weakness of SVM is that it can produce low accuracy if used on a dataset that has many overlapping target classes [12]. Another method that can be used to classify diabetes is using the Decision Tree algorithm. (DT).

One of the major advantages of the Decision Tree algorithm is that it can produce a Decision Tree that is easy to interpret, has an acceptable level of accuracy, and is efficient in dealing with both discrete and numerical attributes [13]. A previous study that used the Decision Tree algorithm to classify child nutrition, obtained an accuracy of 81.25% [14]. However, the study only performed processing on a small amount of data, namely 195 data. Other studies used the Decision Tree algorithm in classifying Alzheimer's disease with a precision of about 89%. However, this study only processed 373 data [15]. Another study that used the Decision Tree produced accurations of 98.47% but only used 5 indicator attributes in its research [16]. The weakness of the Decision Tree method is to produce a low accuracy of classification if there is a high degree of class imbalance [17]. The Decision Tree algorithm also has other weaknesses, namely many linear attributes with a lot of memory needed and in some cases can occur overfitting [15]. The KNN, SVM, and Decision Tree algorithms have their respective advantages and disadvantages, so the performance results of the third classification of the algorithm will be voted on using the Ensemble Majority Voting algorithm to get the best classification performance.

Ensemble Majority Voting is part of an ensemble learning that is commonly used as a comparison. Ensemble learning is a new field in machine learning and deep learning which is a combination of several different algorithms for which it is used to train data sets and choose final predictions based on the largest number of votes [18]. Voting is a concept in which a decision is taken by looking at the most number of values that appear. Majority voting means making decisions by seeing the value or prediction that appears on each method and choosing the prediction that appears the most [19].

Some studies use the majority vote algorithm, namely the study to predict heart disease with an accuracy of 85.71% [20]. However, the study compared only random forest, SVM, KNN, LSTM, and GRU methods with little data, 303 data only. Another study, which also used the majority voting algorithm, produced an accuracy of 98.78% but did not compare the results of the Decision Tree in its research [21]. Another study used the algorithm to predict breast cancer with an accuracy of 98.1% but did not compare the results of the Decision Tree's research [22].

In this study, testing of the four algorithms is used using the testing technique of Percentage Split and K-Fold Cross Validation. In the Percent Split, a split size of 80% is chosen for the training data and 20% for the test data. In a test technique of K-Fold Cross Validation, a k value of 4 is selected where the data will be divided into four groups and can be alternated as training data as well as test data four times. The research is aimed at determining which algorithms and test techniques are the most optimal in detecting diabetes mellitus and obtaining the best results based on criteria of accuracy, precision, and recall.

2. Research Method

The research methods carried out included data collecting, pre-processing data, data splitting, training and testing data, evaluating models, comparing models, and making decisions.

2.1. Collecting Data

The dataset used in this study is a dataset taken from Kaggle's website in CSV format (https://www.kaggle.com/datasets/nanditapore/healthcare-diabetes?).The target attributes in this dataset are the outcome attributes classified in numbers 0 and 1. Thus, there are 1816 classified data in outcome number 0 and 952 data in outcome number 1. A detailed description of the attribute used in the diabetes dataset is presented in Table 1.

Table 1. Attribute and Description

Attribute	Description
Id	Subject id
Pregnancies	Subjects's number of times pregnant
Glucose	Subjects's glucose concentration over 2 hours in an oral glucose tolerance test
Blood Pressure	Subject's blood pressure (mm Hg)
Skin Thickness	Subject's triceps skinfold thickness (mm)
Insulin	Subject's 2-Hour serum insulin (mu U/ml)
BMI	Body Mass Index (kg/m^2)
Diabetes Pedigree Function	The subject's genetic score for diabetes
Age	Subject's age in years
Outcome	0 = false, 1 = true

2.2. Pre-processing Data

The data preprocessing phase is very important in the data mining process. Data preprocessing means the process prepares raw data so there is noise on the data so that the results obtained in data mining can be better [23]. In this research, attributes that do not influence data mining results will be removed, namely the ID attribute.

2.3. Percentage Split

Percentage Split is a method of dividing or separating data into training and testing data in a certain percentage [24]. Percentage Split is usually used to evaluate an algorithm capable of predicting the percentage of data. The percentage of data used in this study is 20:80, which means that 20% of the data is used for testing data and 80% for random data training. Therefore, this method can be suitable to help optimize the accuracy of the predictive model.

2.4. K-Fold Cross Validation

K-Fold Cross Validation is an evaluation method that is performed after performing the Percentage Split method [25]. K-Fold Cross Validation means that the training data will be divided as k, in this study the value of k entered is 4. The training data is divided into 4 parts which means there is 1 part used as an experiment and 3 parts as training data.

2.5. Support Vector Machine (SVM)

Support Vector Machine (SVM) is a data classification technique with a training process (supervised learning). Support Vector Machine (SVM) is a method that compares a selection of standard parameters of discrete values called candidate sets. The SVM method works by defining the boundary between two classes with the maximum distance from the nearest data. The best hyperplane is one of the characteristics of the SVM classification method to obtain the maximum margin size. The core of the training process on the SVM is an attempt to locate the hyperplane. The use of SVM is limited to minor problems because SVM training algorithms tend to be slow and difficult to implement [26]. The Support Vector Machine formula can be seen in Equation (1).

$$f(x) = sign(wx + b) \tag{1}$$

Where w is weight, x is input, and b is the bias value.

2.6. K-Nearest Neighbor (KNN)

The way the K-Nearest Neighbor (KNN) works is by finding the closest distance between the data to be evaluated and the K-Nearest Neighbor in the training data. In this study, the K value used is 4. Classification uses the function of distance from new data to training data. The space is divided into sections based on the classification of the training data. Classification using the most votes among the classifications of k objects. The KNN algorithm uses the classification of inconsistency as the predictive value of the new query instance. The method of KNN is very simple, working on the shortest distance from the query instance to the training sample to determine the KNN [26]. As for the KNN formula, it can be seen in Equation (2).

$$d = \sqrt{\sum_{i=1}^{p}(x_1 - x_2)^2} \tag{2}$$

Where x_1 is sample data, x_2 is test data, i is variable, d is distance, and p is data dimension.

2.7. Decision Tree

Using a set of decision rules and decision tree structures, huge datasets can be divided into smaller record sets [27]. Here are the stages in the Decision Tree algorithm:

a. Prepare training datasets.

b. Determine the roots of the Decision Tree.

c. Choose the characteristic that will act as the root of the decision tree by calculating the value of the gain. Gains are calculated based on the highest Gain value of available attributes. The following Equation (3) can be used to calculate the Gain value.

$$Gain(S) = Entropy(S) - \sum_{i=1}^{N}\frac{|si|}{|S|} \times Entropy(S_i) \tag{3}$$

d. Procedure for each branch formed, repeat step two. On the other hand, to calculate the entropy value, use the corresponding equation. The equation that can be used is Equation (4).

$$Entropy(S) = \sum_{i=1}^{N} -\pi \times \log(2\pi) \tag{4}$$

e. The Decision Tree formation process ends when all branches of the N node have the same class.

2.8. Ensemble Majority Voting

The Ensemble Majority Voting method became essential in decision-making to produce the best classification performance. The easiest and most efficient decision-making in data mining is based on the majority voting rule which sets samples based on most class assignments. The majority vote rule assigns samples to the class associated with the highest prediction frequency unless a specific limitation on the percentage of class agreement is not met [28].

2.9. Confusion Matrix

The number of correctly classified test data and the number of incorrectly classified test data are given in a table called the Confusion Matrix. The Confusion matrix is also used as one of the test methods to calculate the performance of classification by obtaining Accuracy, Precision, and Recall [29]. The Confusion matrix to be used in this study can be seen in Table 2.

Table 2. Confusion Matrix

		Prediction	
		Negative	Positive
Actual	Negative	True Negative	False Positive
	Positive	False Negative	True Positive

Based on Table 2, True Negative (TN) is negative data that is predicted correctly and False Negative (FN) is positive data that is predicted as negative data. Meanwhile True Positive (TP) is positive data that is predicted correctly and False Positive (FP) is negative data that is predicted as positive data.

a. Accuracy

The error rate, also known as accuracy, is a number that shows whether a model makes a correct or wrong prediction of a data set. Accuracy is usually calculated using independent tests, but not always used in the learning process. To calculate the accuracy value use the formula in Equation (5).

$$accuracy = \frac{TP+TN}{TP+FN+FP+TN}\ x\ 100\% \tag{5}$$

b. Precision

Precision is the matrix used to calculate the ability of the system to generate critical data. It is a True Positive prediction ratio combined with all predicted positive results. Precision in data mining is the amount of TP data divided by the number of data recognized as positive. To calculate precision use the following Equation (6).

$$precision = \frac{TP}{TP+FP}\ x\ 100\% \tag{6}$$

c. Recall

Sensitivity or Recall is a True Positive prediction ratio combined with the total positive data. Sensitivity refers to the ability of testing to identify a positive outcome of several data that is supposed to be positive. Calculating the sensitivity or recall can be done using the following Equation (7).

$$recall = \frac{TP}{TP+FN}\ x\ 100\% \tag{7}$$

3. Result and Discussion

To be able to figure out which algorithms and methods will deliver the best results in the classification of diabetes, we need to compare the confusion matrix, precision, recall, and accuracy of SVM, KNN, Decision Tree, and Ensemble algorithms using the Percentage Split and K-Fold Cross Validation methods.

3.1. SVM Algorithm

The results of the diagnosis of diabetes using SVM algorithms are visible using the confusion matrix that will used to calculate the values of Precision, Recall, and Accuracy to measure the successful rate. The results of the classification using the SVM algorithm using the Percentage Split and K-Fold Cross Validation methods are visible in Table 3.

Table 3. Confusion Matrix, Precision, Recall, and Accuracy of SVM Algorithms with Percentage Split and K-Fold Cross Validation Methods

	Class		Prediction	
			0	1
Percentage Split	Actual	0	297	62
		1	87	272
	Label	Precision	Recall	Accuracy
	0	0.77	0.83	0.7925
	1	0.81	0.76	
	Class		Prediction	
			0	1
K-Fold Cross Validation	Actual	0	1619	197
		1	422	530
	Label	Precision	Recall	Accuracy
	0	0.79	0.89	0.7764
	1	0.73	0.56	

In Table 3, we can see a comparison of the results of SVM algorithm classification using both methods. The Percentage Split method successfully predicted correctly for 297 data labeled 0 and 272 data labeled 1. There were also predictive wrongly where 62 data labeled 0 predicted as 1, and 87 data labeled 1 predicted as 0. Using the K-Fold Cross Validation method, it was successfully predicted correctly 1619 data labeled 0 and 530 data labeled 1, and the predictable error of 197 data labeled 0 predicted as 1, and 422 data labeled 1 predicted as 0. From the confusion matrix, we can get that the Percentage Split method has an accuracy rate of 79.25%, whereas the K-Fold Cross Validation method has an accuracy rate of 77.64%.

3.2. KNN Algorithm

The results of the diagnosis of diabetes using KNN algorithms are visible using the confusion matrix that will used to calculate the values of Precision, Recall, and Accuracy to measure the successful rate. The results of the classification using the KNN algorithm using the Percentage Split and K-Fold Cross Validation methods are visible in Table 4.

Table 4. Confusion Matrix, Precision, Recall, and Accuracy of KNN Algorithms with Percentage Split and K-Fold Cross Validation

	Class		Prediction	
			0	1
Percentage Split	Actual	0	344	15
		1	14	345
	Label	Precision	Recall	Accuracy
	0	0.96	0.96	0.9596
	1	0.96	0.96	
	Class		Prediction	
			0	1
K-Fold Cross Validation	Actual	0	1740	76
		1	267	685
	Label	Precision	Recall	Accuracy
	0	0.87	0.96	0.8761
	1	0.90	0.72	

In Table 4, we can see a comparison of the results of KNN algorithm classification using both methods. The Percentage Split method successfully predicted correctly for 344 data labeled 0 and 345 data labeled 1. There were also predictive wrongly where 15 data labeled 0 predicted as 1, and 14 data labeled 1 predicted as 0. Using the K-Fold Cross Validation method, it was successfully predicted correctly 1740 data labeled 0 and 685 data labeled 1, and the predictable error of 76 data labeled 0 was predicted as 1, and 267 data labeled 1 predicted as 0. From the confusion matrix, we can get that the Percentage Split method has an accuracy rate of 95.96%, whereas the K-Fold Cross Validation method has an accuracy rate of 87.61%.

3.3. Decision Tree Algorithm

The results of the diagnosis of diabetes using Decision Tree algorithms are visible using the confusion matrix that will used to calculate the values of Precision, Recall, and Accuracy to measure the successful rate. The results of the classification using the Decision Tree algorithm using the Percentage Split and K-Fold Cross Validation methods are visible in Table 5.

Table 5. Confusion Matrix, Precision, Recall, and Accuracy of Decision Tree Algorithms with Percentage Split and K-Fold Cross Validation Methods

	Class		Prediction	
			0	1
Percentage Split	Actual	0	356	3
		1	28	331
	Label	Precision	Recall	Accuracy
	0	0.93	0.99	0.9568
	1	0.99	0.92	
	Class		Prediction	
			0	1
K-Fold Cross Validation	Actual	0	1798	18
		1	21	931
	Label	Precision	Recall	Accuracy
	0	0.99	0.99	0.9859
	1	0.98	0.98	

In Table 5, we can see a comparison of the results of Decision Tree algorithm classification using both methods. The Percentage Split method successfully predicted correctly for 356 data labeled 0 and 331 data labeled 1. There were also predictive wrongly where 3 data labeled 0 predicted as 1, and 28 data labeled 1 predicted as 0. Using the K-Fold Cross Validation method, it was successfully predicted correctly 1798 data labeled 0 and 931 data labeled 1, and the predictable error of 18 data labeled 0 predicted as 1, and 21 data labeled 1 predicted as 0. From the confusion matrix, we can get that the Percentage Split method has an accuracy rate of 95.68%, whereas the K-Fold Cross Validation method has an accuracy rate of 98.59%.

3.4 Ensemble Algorithm

The Ensemble algorithm used here is Majority Voting which means making decisions based on the most class assignments of the three previously used. The results of the diagnosis of diabetes using Ensemble algorithms are visible using the confusion matrix that will used to calculate the values of Precision, Recall, and Accuracy to measure the successful rate. The results of the classification using the Ensemble algorithm using the Percentage Split and K-Fold Cross Validation methods are visible in Table 6.

Table 6. Confusion Matrix, Precision, Recall, and Accuracy of Decision Tree Algorithms with Percentage Split and K-Fold Cross Validation Methods

	Class		Prediction	
			0	1
Percentage Split	Actual	0	355	4
		1	12	347
	Label	Precision	Recall	Accuracy
	0	0.97	0.99	0.9777
	1	0.99	0.97	
	Class		Prediction	
			0	1
K-Fold Cross Validation	Actual	0	1804	12
		1	21	931
	Label	Precision	Recall	Accuracy
	0	0.99	0.99	0.9881
	1	0.99	0.98	

In Table 6, we can see a comparison of the results of Ensemnle algorithm classification using both methods. The Percentage Split method successfully predicted correctly for 355 data labeled 0 and 347 data labeled 1. There were also predictive wrongly where 4 data labeled 0 predicted as 1, and 12 data labeled 1 predicted as 0. Using the K-Fold Cross Validation method, it was successfully predicted correctly 1804 data labeled 0 and 931 data labeled 1, and the predictable error of 12 data labeled 0 predicted as 1, and 21 data labeled 1 predicted as 0. From the confusion matrix, we can get that the Percentage Split method has an accuracy rate of 97.77%, whereas the K-Fold Cross Validation method has an accuracy rate of 98.81%.

3.5 Comparison of Four Algorithms

The Comparison of the results of four algorithms and two methods can be seen from the accuracy values of each SVM, KNN, Decision Tree, and Ensemble, as seen in Table 7.

Table 7. Comparison of Precision, Recall, and Accuracy Values of Each Algorithm with the Percentage Split and K-Fold Cross Validation Methods in Percent (%)

Method	Algorithm	Label	Precision	Recall	Accuracy
Percentage Split	SVM	0	77	83	79.25
		1	81	76	
	KNN	0	96	96	95.96
		1	96	96	
	Decision Tree	0	93	99	95.68
		1	99	92	
	Ensemble	0	97	99	97.77
		1	99	97	
K-Fold Cross Validation	SVM	0	79	89	77.64
		1	73	56	
	KNN	0	87	96	87.61
		1	90	72	
	Decision Tree	0	99	99	98.59
		1	98	98	
	Ensemble	0	99	99	98.81
		1	99	98	

Table 7 shows that the KNN and Decision Tree algorithms can give quite good results in classifying diabetes, whereas the SVM algorithms give poor results. It also appears that the Ensemble algorithm can improve the accuracy of the classification of diabetes. Graphic improvement of accuracy values of each algorithm with the Percentage Split and K-Fold Cross Validation methods visible in Figure 1.

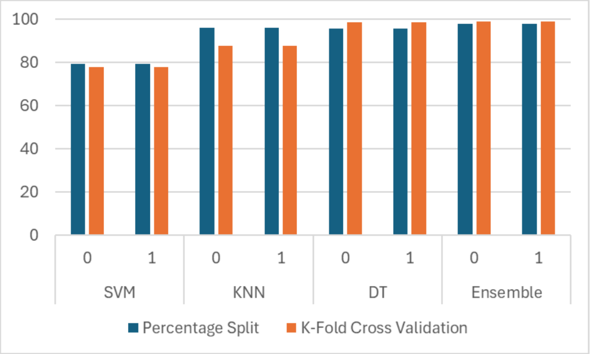

Figure 1. Graphic Comparison Value of Accuracy of each Algorithm with Percentage Split and K-Fold Cross Validation Methods.

Figure 1 shows that the Percentage Split method can provide better results for SVM and KNN algorithms. The respective accuracy values of the SVM algorithms and the KNN with the Split percentage method are 79.25% and 95.96%, while the K-Fold Cross Validation method only achieves accurations of 77.64% and 87.61%. On the other hand, K-Fold Cross Validation can provide a better result for the Decision Tree and Ensemble algorithms. The respective accuracy values of each of the Decision Tree and Ensemble algorithms with the K-Fold Cross Validation method are 98.59% and 98.81%, while in the Percentage Split method, the accurate values will be 95.68% and 97.77%. The Ensemble algorithm can improve the accuracy of the other three methods with an average increase of 9.15% with the Percentage Split method and 13.42% with the K-Fold Cross Validation method.

The results of this study are much better compared to previous studies that only used the C4.5 and SVM algorithms. The study also only divided the data with the Holdout method so that it only got an accuracy rate of 75.32% with the C4.5 algorithm, and 82.01% with the SVM algorithm. This study uses the SVM, KNN, and Decision Tree algorithms. This study also utilized the Ensemble Majority Voting process to improve the accuracy of results. We divided the data using the Percentage Split and K-Fold Cross Validation methods to get an accuracy rate of 98.81%.

4. Conclusion

The study compared classifications from the KNN, SVM, Decision Tree, and Ensemble Majority Voting algorithms with the Percentage Split and K-Fold Cross Validation methods. Based on the results of this study, it was concluded that the Ensemble Majority Voting algorithm with the K-Fold Cross Validation method could provide a better classification of diabetes.

The accuracy result is divided into two sections, namely the Percentage Split and K-Fold Cross Validation methods. The Ensemble Majority Voting algorithm successfully gives the highest accuracy value in both Percentage Split and K-Fold Cross Validation methods. K-Fold Cross Validation can be better because it delivers an average improvement of 13.42% in ensemble accuracy compared to the Percent Split method, which only provides an average increase of 9.15%. Besides that, the accuracy of the results has reached 98.81% using the Ensemble Majority Voting algorithm in the K-Fold Cross Validation method compared to the Percentage Split which only gives 97.77%, which means that the Ensemble Majority Voting algorithm and K-Fold Cross Validation method works best in classifying diabetes.

Further research can compare the results of other types of algorithms using the ensembles of other kinds in the Percentage Split and K-Fold Cross Validation methods to determine with certainty which method performs best consistently in the classification of diabetes.

References

[1] A. R. P. Abimanyu, A. D. Rahma, D. R. Putri, R. N. Ilham, W. A. Audia, and M. Arfania, “Pengaruh Terapi Pada Penderita Diabetes Mellitus Sebagai Penurunan Kadar Gula Darah: Review Artikel”, *Innovative*, vol. 3, no. 2, pp. 8931–8949, Jun. 2023..

[2] S. Rammang, Nurhikmah, and N. Reza, “Pengendalian Diabetes Melitus Melalui Edukasi dan Pemeriksaan Kadar Gula Darah Sewaktu,” *Jurnal Pendidikan Tambusai*, vol. 7, no. 1, pp. 133–137, 2023.

[3] P. B. N. Setio, D. R. S. Saputro, and Bowo Winarno, “Klasifikasi

Dengan Pohon Keputusan Berbasis Algoritme C4.5," *Prisma, Prosiding Seminar Nasional Matematika*, vol. 3, pp. 64–71, 2020.

[4] D. Cahyanti, A. Rahmayani, and S. Ainy Husniar, "Analisis Performa Metode Knn pada Dataset Pasien Pengidap Kanker Payudara," *Indonesian Journal of Data Science*, vol. 1, no. 2, pp. 39–43, July 2020.

[5] M. Yunus and N. K. A. Pratiwi, "Prediksi Status Gizi Balita Dengan Algoritma K-Nearest Neighbor (KNN) di Puskemas Cakranegara," *JTIM: Jurnal Teknologi Informasi dan Multimedia*, vol. 4, no. 4, pp. 221–231, 2023.

[6] F. T. Admojo and Ahsanawati, "Klasifikasi Aroma Alkohol Menggunakan Metode KNN," *Indonesian Journal of Data Science*, vol. 1, no. 2, pp. 34–38, July 2020.

[7] M. Saputra, J. P. Sidabuke, R. P. Sinulingga, and R. B. Tamba, "Analisis Metode Algoritma K-Nearest Neighbor (KNN) dan Naive Bayes Untuk Klasifikasi Diabetes Mellitus," *Jurnal TEKINKOM*, vol. 6, no. 2, pp. 723–729, 2023.

[8] S. Talib, S. Sudin, and M. D. Suratin, "Penerapan Metode Support Vector Machine (SVM) pada Klasifikasi Jenis Cengkeh Berdasarkan Fitur Tekstur Daun," *Jurnal RESTIA*, vol. 2, no. 1, pp. 17-27, February 2024.

[9] N. Prasetiyo, K. Ahmad Baihaqi, S. Arum, P. Lestari, and Y. Cahyana, "Classification of Rice Plants Affected by Rats Using the Support Vector Machine (SVM) Algorithm," *Jurnal Teknik Informatika (JUTIF)*, vol. 5, no. 2, pp. 637–643, April 2024.

[10] D. S. Rahayu, Nursafika, J. Afifah, and S. Intan, "Classification of Diabetes Mellitus Using C4.5 Algorithm, Support Vector Machine (SVM) and Linear Regression," *SENTIMAS: Seminar Nasional Penelitian dan Pengabdian Masyarakat*, vol. 2, pp. 56–63, August 2023.

[11] H. N. Irmanda and Ria Astriratma, "Klasifikasi Jenis Pantun dengan Metode Support Vector Machines (SVM)," *Jurnal RESTI (Rekayasa Sistem dan Teknologi Informasi)*, vol. 4, no. 5, pp. 915–922, 2021.

[12] N. Arifin, U. Enri, and N. Sulistiyowati, "Penerapan Algoritma Support Vector Machine (SVM) dengan TF-IDF N-Gram untuk Text Classification," *STRING (Satuan Tulisan Riset dan Inovasi Teknologi*, vol. 6, no. 2, pp. 129-136, December 2021.

[13] H. Rifa'i, Ryan Hamonangan, Dian Ade Kurnia, Kaslani, and Mulyawan, "Implementasi Algoritma Decision Tree Dalam Klasifikasi Kompetensi Siswa," *KOPERTIP: Jurnal Ilmiah Manajemen Informatika dan Komputer*, vol. 6, no. 1, pp. 15–20, February 2022.

[14] M. Ula, A. F. Ulva, M. Mauliza, M. A. Ali, and Y. R. Said, "Application of Machine Learning in Determining the Classification of Children'S Nutrition With Decision Tree," *Jurnal Teknik Informatika (JUTIF)*, vol. 3, no. 5, pp. 1457–1465, October 2022.

[15] A. A. Mortara, M. Permatasari, A. Desiani, Y. Andriani, and M. Arhami, "Comparison of C4.5 and Adaptive Boosting Algorithms in Alzheimer's Disease Classification," *Jurnal Teknologi dan Informasi*, vol. 13, no. 2, pp. 196–207, September 2023.

[16] D. Fatmawati, W. Trisnawati, Y. Jumaryadi, and G. Triyono, "Klasifikasi Tingkat Kepuasan Penggunaan Layanan Teknologi Informasi Menggunakan Decision Tree," *KLIK: Kajian Ilmiah Informatika dan Komputer*, vol. 3, no. 6, pp. 1056–1062, June 2023.

[17] Y. Crismayella, N. Satyahadewi, and H. Perdana, "Algoritma Adaboost pada Metode Decision Tree untuk Klasifikasi Kelulusan Mahasiswa," *Jambura Journal of Mathematics*, vol. 5, no. 2, pp. 278–288, August 2023.

[18] Kade Bramasta Vikana Putra, I Putu Agung Bayupati, and Dewa Made Sri Arsa, "Klasifikasi Citra Daging Menggunakan Deep Learning dengan Optimisasi Hard Voting," *Jurnal RESTI (Rekayasa Sistem dan Teknologi Informasi)*, vol. 5, no. 4, pp. 656–662, 2021.

[19] T. I. Rais, "Analisis Sentimen Terhadap Komentar Video Youtube Raiden Shogun-Judgment of Euthymia Menggunakan Metode Majority Voting," Bachelor thesis, Fakultas Sains dan Teknologi, Universitas Islam Negeri Syarif Hidayatullah, Banten, 2022.

[20] I. Javid, A. K. Z. Alsaedi, and R. Ghazali, "Enhanced Accuracy of Heart Disease Prediction using Machine Learning and Recurrent Neural Networks Ensemble Majority Voting Method," *(IJACSA) International Journal of Advanced Computer Science and Applications*, vol. 11, no. 3, pp. 540–551, 2020.

[21] A. M. Bamhdi, I. Abrar, and F. Masoodi, "An Ensemble Based Approach for Effective Intrusion Detection using Majority Vvoting," *TELKOMNIKA (Telecommunication, Computing, Electronics and Control)*, vol. 19, no. 2, pp. 664–671, April 2021.

[22] M. A. Naji, S. El Filali, M. Bouhlal, E. H. Benlahmar, R. A. Abdelouhahid, and O. Debauche, "Breast Cancer Prediction and Diagnosis through a New Approach based on Majority Voting Ensemble Classifier," *Procedia Computer Science*, vol. 191, pp. 481–486, 2021.

[23] N. A. R. Putri and Ardiansyah, "Analisis Sentimen Terhadap Kemajuan Kecerdasan Buatan di Indonesia Menggunakan BERT dan RoBERTa," *Jurnal Sains dan Informatika*, vol. 9, no. 2, pp. 136–145, November 2023.

[24] A. Septiarini, R. Saputra, A. Tejawati, and M. Wati, "Deteksi Sarung Samarinda Menggunakan Metode Naive Bayes Berbasis Pengolahan Citra," *Jurnal RESTI (Rekayasa Sistem dan Teknologi Informasi)*, vol. 5, no. 5, pp. 927–935, 2021.

[25] N. Yudistira and A. F. Putra, "Algoritma Decision Tree Dan Smote Untuk Klasifikasi Serangan Jantung Miokarditis Yang Imbalance," *Jurnal Litbang Edusaintech*, vol. 2, no. 2, pp. 112–122, October 2021.

[26] R. Umar, I. Riadi, and D. A. Faroek, "Komparasi Image Matching Menggunakan Metode K-Nearest Neighbor (KNN) dan Metode Support Vector Machine (SVM)," *Journal of Applied Informatics an Computing*, vol. 4, no. 2, pp. 124–131, December 2020.

[27] A. I. Putri, Y. Syarif, P. Jayadi, F. Arrazak, and F. N. Salisah, "Implementation of Decision Tree and Support Vector Machine (SVM) Algorithm for Stunting Risk Prediction," *MALCOM: Indonesian Journal of Machine Learning and Computer Science*, vol. 3, no. 2, pp. 349–357, October 2024.

[28] A. Wibowo, M. Makruf, I. Virdyna, and F. C. Venna, "Penentuan Klaster Koridor TransJakarta dengan Metode Majority Voting pada Algoritma Data Mining," *Jurnal RESTI (Rekayasa Sistem dan Teknologi Informasi)*, vol. 5, no. 3, pp. 565–575, 2021.

[29] A. Wibowo, S. Wardani, R. W. Dewantoro, W. Wesly, and Leonardo, "Komparasi Tingkat Akurasi Random Forest dan Decision Tree C4.5 Pada Klasifikasi Data Penyakit Infertilitas," *KLIK: Kajian Ilmiah Informatika dan Komputer*, vol. 4, no. 1, pp. 218–224, August 2023.