

# Improvement Performance of the Random Forest Method on Unbalanced Diabetes Data Classification Using Smote-Tomek Link

Hairani Hairani<sup>a\*</sup>, Anthony Anggrawan<sup>b</sup>, Dadang Priyanto<sup>c</sup>

<sup>abc</sup> Department of Computer Science, Universitas Bumigora, Mataram, 83127, Indonesia

Corresponding author: Hairani@universitasbumigora.ac.id

**Abstract**— Most of the health data contained unbalanced data that affected the performance of the classification method. Unbalanced data causes the classification method to more easily classify the majority data and ignore the minority class. One of the health data that has unbalanced data is Pima Indian Diabetes. Diabetes is a deadly disease caused by the body's inability to produce enough insulin. Complications of diabetes can cause heart attacks and strokes. Early diagnosis of diabetes is needed to minimize the occurrence of more severe complications. In the diabetes dataset used, there is an imbalance of data between positive and negative diabetes classes. Diabetes negative class data (500 data) is more than diabetes positive class (268) so that it can affect the performance of the classification method. Therefore, this study aims to apply the Smote-TomekLink and Random Forest methods in the classification of diabetes. The research methodology used is the collection of diabetes data obtained from Kaggle as many as 768 data with 8 input attributes and 1 output attribute as a class, pre-processing data is used to balance the dataset with Smote-TomekLink, classification using the random forest method, and performance evaluation based on accuracy, sensitivity, precision, and F1-score. Based on the tests carried out by dividing data using 10-fold cross-validation, the Random forest algorithm with Smote-TomekLink gets the highest accuracy, sensitivity, precision, and F1-score compared to Random Forest with Smote. The Random Forest algorithm with Smote-TomekLink has 86.4% accuracy, 88.2% sensitivity, 82.3% precision, and 85.1% F1-score. Thus, using Smote-TomekLink can improve the performance of the random forest method based on accuracy, sensitivity, precision, and F1-score.

**Keywords**— Class Imbalance; Smote-TomekLink; Random Forest Method; Diabetest Disease.

*Manuscript received dd mm yyyy; revised dd mm yyyy; accepted dd mm yyyy. Date of publication dd mm yyyy.  
International Journal on Informatics Visualization is licensed under a Creative Commons Attribution-Share Alike 4.0 International License.*



## I. INTRODUCTION

Most of the Health data contained unbalanced data that affected the performance of the classification method. Unbalanced data causes the classification method to more easily classify the majority data and ignore the minority class. One of the health data that has unbalanced data is Pima Indian Diabetes. Diabetes is a deadly disease caused by increased blood sugar in the body. Diabetes is caused by the body's inability to produce enough insulin. Complications of diabetes can cause heart attacks and strokes. One way to improve the performance of the classification method is to handle balanced data by adding minority data so that the number is equal to the majority class. The diabetes dataset has 768 instances of data. However, the problem is that there is an imbalance of data in the dataset, namely the negative diabetes class with 500 data (majority class), while the positive diabetes class with 268 data (minority class). Data imbalance is the amount of data in one class more than in the other class. The problem of data imbalance causes the classification method to be more

dominant in classifying the majority class than the minority class, or in other words, the classification method ignores the minority class. The problem of unbalanced data can be handled with a data sampling approach.

Several data sampling methods that can be used to solve the problem of data imbalance are oversampling [1][2], [3][4], undersampling [5][6], and Hybrid Sampling[6],[7]. Oversampling works by adding the minority class, while Undersampling works by removing the majority class so as to produce balanced data. However, both methods have their respective weaknesses. The weakness of the oversampling method is that there are too many repetitions of samples that can cause overfitting of the classification method, while the weakness of undersampling is that it will lose information from most of the samples in the dataset and cannot take full advantage of the available information[9].

To avoid overfitting the oversampling method, the Smote method was developed to overcome these weaknesses. Smote is an oversampling method to generate new synthesis training data by linear interpolation on minority classes[10]. However,

the Smote method has a weakness, namely overgeneralization, and the addition of a minority class randomly can generate noise data, because it does not differentiate between classes[11]. Therefore, the undersampling method is used to improve the performance of the oversampling method by cleaning the noise data in the majority class. The noise data is the majority class instance which is closest to the minority class instance. Usually, noise data reduces the level of accuracy for classification methods[5]. One method to remove noise data in the majority class is Tomeklink[12]. Tomeklink is an undersampling method that cleans noise data from the majority class which has similar characteristics and overlapping. However, Tomeklink only deletes instances defined as “Tomek Links” so that the analyzed data cannot be balanced and in its implementation the method is combined with other methods. Combining Tomeklink and Smote oversampling can improve accuracy better than individual performance[25].

Several previous studies that have discussed the classification of diabetes, namely Research [13] predicts diabetes using the k-NN method with an accuracy of 83%. The weakness of the research is that it does not address the problem of data imbalance. Research [14] classifying diabetes using the C4.5 method with an accuracy of 75.65%. The weakness of the research is that the accuracy obtained is low so that it can be improved, and also does not address the problem of data imbalance. Research [15] Using XGBoost to predict diabetes with 74% accuracy. The weakness of the research is that the accuracy obtained is low so that it can be improved, and also does not address the problem of data imbalance.

Research [16] using the Correlated Naïve Bayes method with correlation-based feature selection to predict diabetes with an accuracy of 69.51%. The weakness of the research is that the accuracy obtained is low so that it can be improved, and also does not address the problem of data imbalance. Research [17] using the C4.5 method for diabetes detection with an accuracy of 68%.

Research [18] used logistic regression and smote methods to detect diabetes with 82% accuracy, 81% precision, 79% recall, and 80% F1-score. The weakness of the research is that the accuracy is good but can be improved using Tomeklink to clean noise data in the majority class. Research [19] using the C4.5 and Smote methods to predict diabetes with 82% accuracy, 80% precision, and 86% sensitivity. Research [20] used logistic and Smote-ENN methods to predict kidney disease with 75.2% accuracy, 70.6% recall, 4.9% precision, and 30% F1-score. The weakness of the research is the low accuracy so that it can be improved using Tomeklink to clean noise data in the majority class. Research [21] SME-XGBoost with Smote-ENN for heart disease prediction with 80% AUC.

Based on previous research, this study proposes the Smote-Tomeklink method to resolve the imbalance of diabetes data. Smote-Tomeklink is a good way to avoid the drawbacks of SMOTE and Tomeklink teknik techniques [9]. The classification method used in this research is Random Forest. The Random Forest method was chosen because it has several advantages, namely high accuracy [22], the ability to handle noise data, fast performance in training data, overfitting control, and easy to implement [23]. Therefore, the combination method of Smote and Tomeklink (Smote-Tomeklink) is applied to balance the data on diabetes data so as to improve the

performance of the Random forest classification method based on accuracy, sensitivity (recall), precision, and F1-score.

## II. MATERIALS AND METHOD

This research consists of several stages as shown in Figure 1.

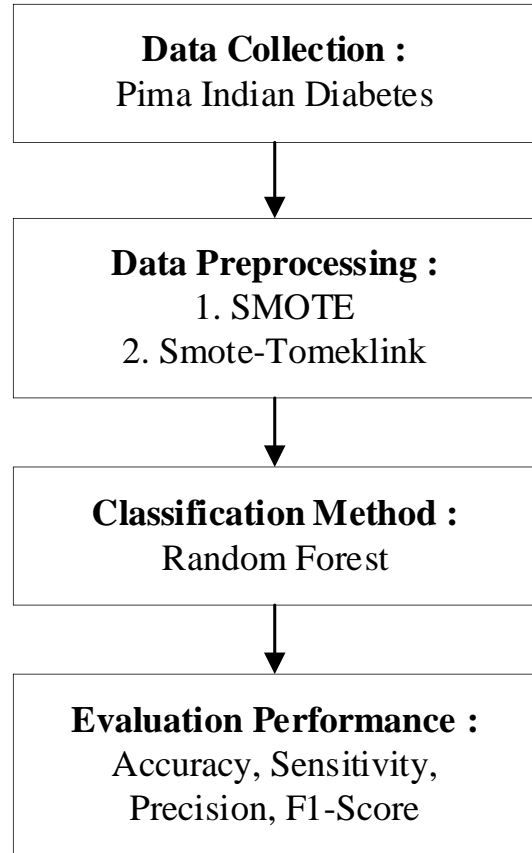


Fig 1. Research Stages

### A. Data Collection

The dataset used in this study is a diabetes dataset obtained from Kaggle, which consists of 768 instances and 9 attributes. The description of the attributes and the sample data used are shown respectively in Table I and Table II.

TABLE I  
DESCRIPTION ATRIBUT DATASET

No	Atribute	Description	Label
1	Pregnancies	Number of Pregnancy	X1
2	Glucose	Glucose level 2 hours after eating	X2
3	Blood Pressure	Blood Pressure	X3
4	Skin Thickness	Skin Thickness	X4
5	Insulin	Insulin	X5
6	BMI	Body Massa Index	X6
7	Diabetes Pedigree Function	Diabetes Pedigree Function	X7
8	Age	Age	X8
9	Outcome	Diabetes Status ( 1 = Positive Diabetes, 2 = Negative Diabetes	Y

TABLE III  
SAMPLE DATASET

No	X1	X2	X3	X4	X5	X6	X7	X8	Y
1	6	148	72	35	0	33.6	0.627	50	1
2	1	85	66	29	0	26.6	0.351	31	0
3	8	183	64	0	0	23.3	0.672	32	1
4	1	89	66	23	94	28.1	0.167	21	0
5	0	137	40	35	168	43.1	2.288	33	1
6	5	116	74	0	0	25.6	0.201	30	1
7	3	78	50	32	88	31	0.248	26	1
8	10	115	0	0	0	35.3	0.134	29	0
9	2	197	70	45	543	30.5	0.158	53	1
10	8	125	96	0	0	0	0.232	54	1
11	4	110	92	0	0	37.6	0.191	30	0
12	10	168	74	0	0	38	0.537	34	1
13	10	139	80	0	0	27.1	1.441	57	0
14	1	189	60	23	846	30.1	0.398	59	1
15	5	166	72	19	175	25.8	0.587	51	1
..	..	...	..	..	...	..	.....	..	..
754	0	181	88	44	510	43.3	0.222	26	1
755	8	154	78	32	0	32.4	0.443	45	1
756	1	128	88	39	110	36.5	1.057	37	1
757	7	137	90	41	0	32	0.391	39	0
758	0	123	72	0	0	36.3	0.258	52	1
759	1	106	76	0	0	37.5	0.197	26	0
760	6	190	92	0	0	35.5	0.278	66	1
761	2	88	58	26	16	28.4	0.766	22	0
762	9	170	74	31	0	44	0.403	43	1
763	9	89	62	0	0	22.5	0.142	33	0
764	10	101	76	48	180	32.9	0.171	63	0
765	2	122	70	27	0	36.8	0.34	27	0
766	5	121	72	23	112	26.2	0.245	30	0
767	1	126	60	0	0	30.1	0.349	47	1
768	1	93	70	31	0	30.4	0.315	23	0

### B. Data Pre-processing

Data Pre-processing is one of the important stages in data mining to improve the quality of datasets. This study focuses on dealing with unbalanced data contained in the diabetes dataset. The dataset used has 268 instances of negative diabetes and 500 instances of Positive Diabetes. The algorithms used to handle unbalanced data in the dataset are SMOTE (Synthetic Minority Oversampling Technique) and Smote-Tomeklink.

SMOTE is one of the most commonly used oversampling methods to solve the problem of data distribution imbalance in machine learning modeling. SMOTE aims to balance the distribution of classes by increasing the number of minority classes randomly by creating synthetic data for oversampling purposes [10]. Creating new data on the minority class using the equation (1).

$$Y' = Y^i + (Y^j - Y^i) * \gamma \quad (1)$$

$Y'$  representation of the addition of the minority class.  $Y^i$  representasi kelas minoritas,  $Y^j$  is a value chosen at random from the k-nearest neighbors of the minority class on  $Y^i$ , and  $\gamma$  is a value in a randomly selected vector with a range of 0 to 1 [2].

SMOTE generates new synthesis training data by linear interpolation for the minority class. Synthesis training data is generated by randomly selecting one or more of the k-nearest neighbors for each sample in the minority class as shown in Figure 2.

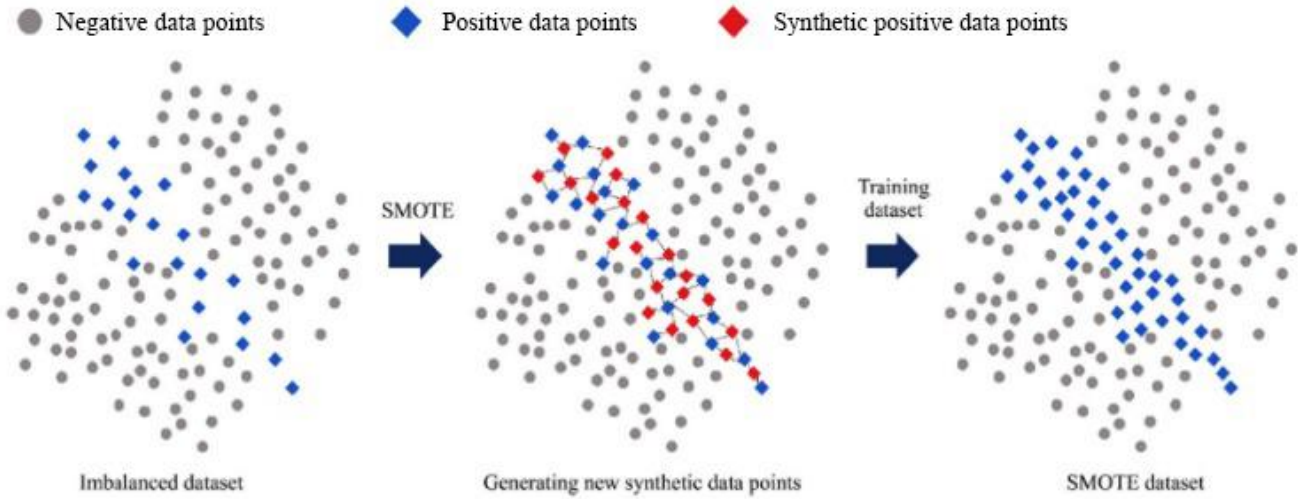


Fig 2. Process of Synthetic Minority Oversampling Technique (SMOTE) Algorithm [24]

Tomeklink is an undersampling method that cleans noise data from the majority class that has similar characteristics and overlapping[12]. Tomeklink works by eliminating the majority class instances that are closer to the minority class by applying the nearest neighbor rule to select instances. The combination of Tomeklink and Smote oversampling can improve accuracy better than individual performance [25].

### C. Random Forest Method

Random Forest is a decision tree-based ensemble learning method [26]. The Random Forest method has the advantages of high accuracy, the ability to handle noise data, fast performance in training data, overfitting control, and easy to implement [23]. The working process of the Random Forest method in classifying a data is shown in Figure 3.

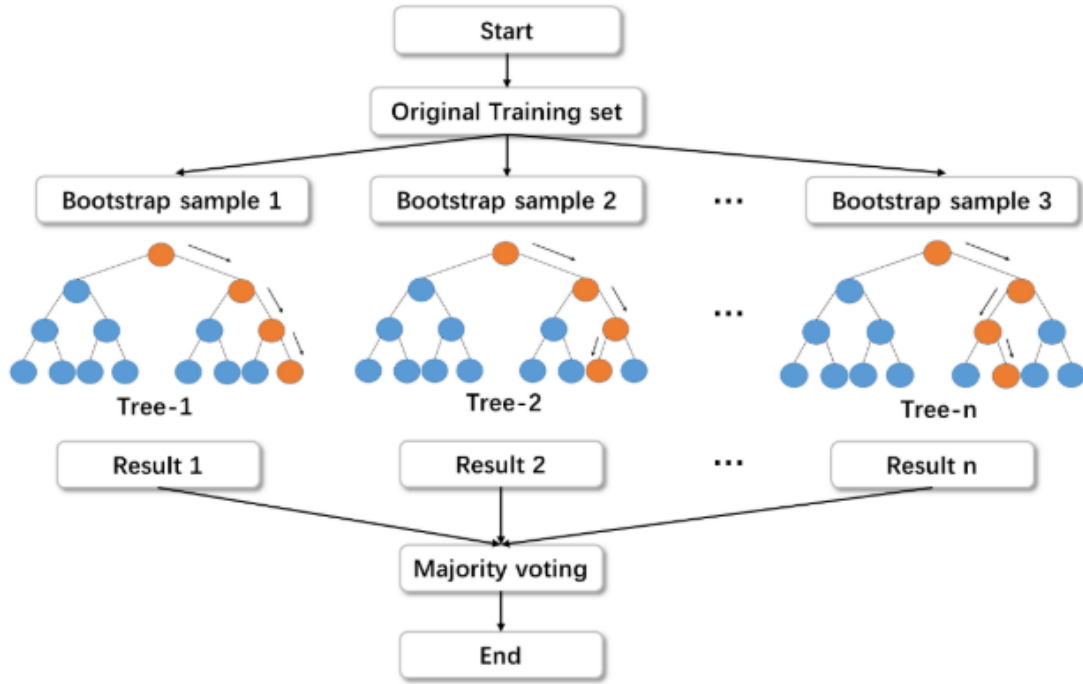


Fig 1. Working Process of Random Forest Method[23]

Figure 3 shows how the Random Forest algorithm works by creating a set of decision trees from a randomly selected subset, getting predictions from each decision tree, voting for each predicted outcome, and choosing the best prediction result based on the most votes assigned as final prediction

#### D. Evaluation Performance

Performance testing uses a confusion matrix table. The confusion matrix is a table that is used to describe the performance of the classification method on a dataset whose true value is known. The confusion matrix can visualize the amount of data that is classified as true and false as shown in the Table III[27].

TABLE III  
CONFUSION MATRIX

Actual	Predicted	
	Negative	Positive
Negative	TN	FP
Positive	FN	TP

Formula used to calculate *Accuracy* (6), *Sensitivity* (7), *Precision* (8) [28] [29][30], and *F1-score* (5)[31].

$$Accuracy = \frac{TP + TN}{TP + FN + TN + FP} \quad (6)$$

$$Sensitivity = \frac{TP}{TP + FN} \quad (7)$$

$$Precision = \frac{TP}{TP + FP} \quad (8)$$

$$F1-score = \frac{2 \times precision \times recall}{precision + recall} \quad (9)$$

True Positive (TP) is a class of positive diabetes that is predicted correctly. False Positive (FP) is a diabetes negative class but is predicted to be diabetes positive. True Negative (TN) is a diabetes negative class that is predicted correctly. False Negative (FN) is a positive diabetes class but is predicted to be diabetes negative.

### III. RESULT AND DISCUSSION

This research starts from the stages of data collection, data pre-processing, classification, and performance testing. The data used in this study is diabetes data obtained from Kaggle. The pre-processing of this study used the Smote and Smote-Tomeklink algorithms to deal with class imbalances in diabetes data. The classification method of this research is Random Forest. The performance test is based on accuracy, sensitivity, precision, and F1-score. The results of the comparison of the original data with the data from Smote and the results of Smote-Tomeklink are shown in Figure 4.

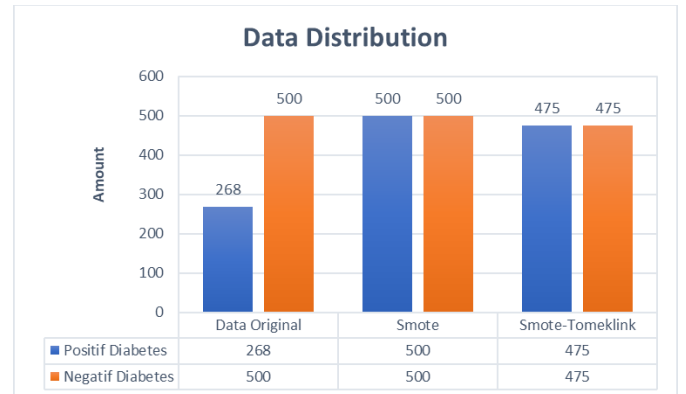


Fig 4. Data Distribution Result

The classification method of this research is Random Forest. Performance testing is based on accuracy, sensitivity, precision,

and F1-score using a confusion matrix table. Based on testing the Random Forest method using 10-fold cross-validation, the results obtained in the form of a confusion matrix table as shown in Table IV for the Random Forest method on the original data, Table V for the results of the Random Forest method with Smote, and Table VI for the results of the Random Forest method with Smote-Tomeklink. The results of the comparison of the performance of the Random Forest method as a whole are shown in Figure 5.

TABLE IV  
RESULT CONFUSION MATRIX OF RANDOM FOREST

Actual	Predicted	
	Negative	Positive
Negative	429	71
Positive	113	155

TABLE V  
RESULT CONFUSION MATRIX OF RANDOM FOREST AND SMOTE

Actual	Predicted	
	Negative	Positive
Negative	390	110
Positive	71	429

TABLE VI  
RESULT CONFUSION MATRIX OF RANDOM FOREST AND SMOTE-TOMEKLINK

Actual	Predicted	
	Negative	Positive
Negative	385	90
Positive	56	419

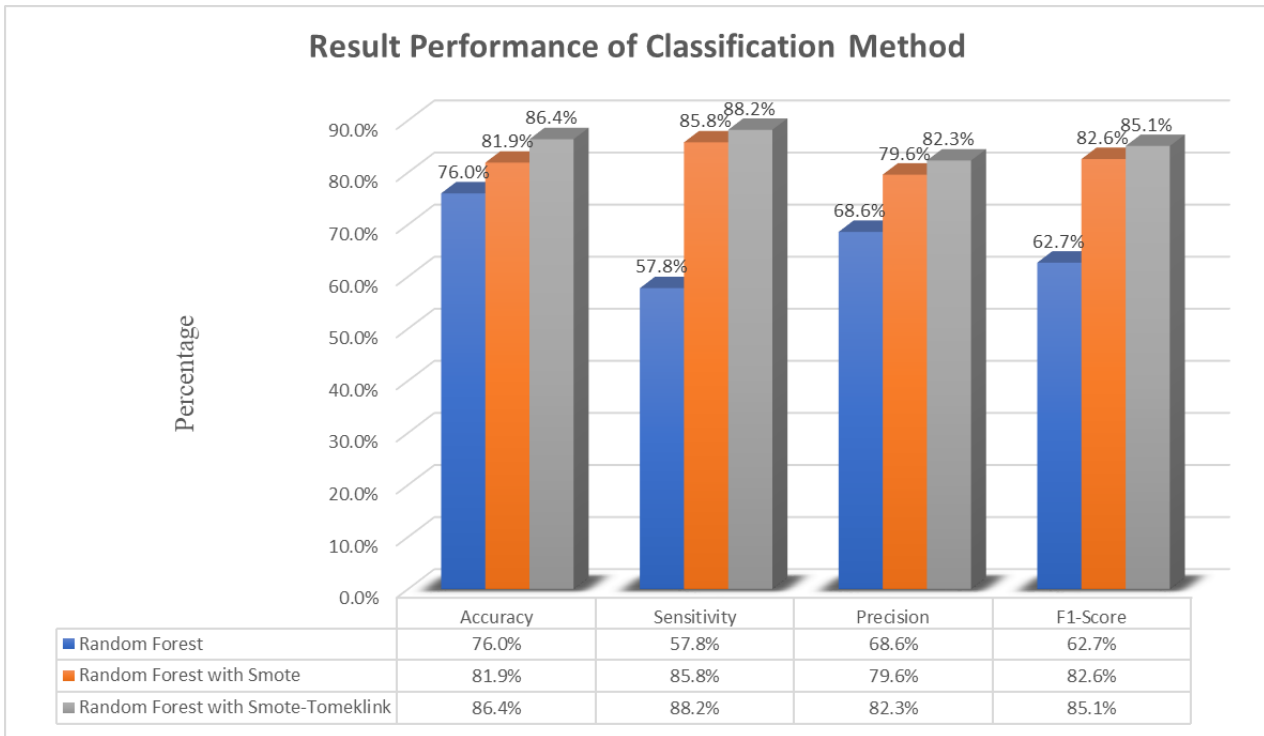


Fig 5. Result Performance of Classification Method

In Table IV, the Random Forest method succeeded in correctly classifying the negative class (TN) as many as 429 instances and the negative class classified incorrectly (FP) as many as 17 instances. While the correctly classified positive class (TP) is 155 instances and the incorrectly classified positive class is 113 instances.

In Table V, the Random Forest method with Smote succeeded in correctly classifying the negative class (TN) as many as 390 instances and the negative class classified incorrectly (FP) as many as 110 instances. While the positive class that is classified correctly (TP) is 429 instances and the positive class that is classified incorrectly is 71 instances.

In Table VI, the Random Forest method with Smote-Tomeklink succeeded in correctly classifying the negative class (TN) as many as 385 instances and the negative class classified incorrectly (FP) as 90 instances. While the positive class that is classified correctly (TP) is 419 instances and the positive class that is classified incorrectly is 56 instances.

Based on Figure 4, there was an increase in the performance of the Random Forest method with Smote-Tomeklink based on accuracy, sensitivity, precision, and F1-score. In the original dataset, the Random Forest method has 76% accuracy, 57.8% sensitivity, 68.6% precision, and 62.7% F1-score. The Random Forest method with Smote has an accuracy of 81.9%, sensitivity of 85.8%, precision of 79.6%, and F1-score of 82.6%. Meanwhile, the use of the Random Forest method with Smote-Tomeklink resulted in an accuracy of 86.4%, a sensitivity of 88.2%, a precision of 83.3%, and F1-score of 85.1%.

Sensitivity has a very important role to improve the accuracy and F1-score performance of the Random Forest method with Smote-Tomeklink. The Random Forest method with Smote-Tomeklink gives higher accuracy, sensitivity, precision, and F1-score results than smote and without sampling.

Random Forest method with Smote an increase in performance indicators accuracy, sensitivity, precision, and F1-score. The increase in accuracy scores is 5.9%, Sensitivity is

28%, precision is 11%, and F1-score is 19.9%. The Random Forest method with Smote-Tomeklink showed an increase in the indicators of accuracy by 10.4%, Sensitivity by 30.4%, precision by 13.7%, and F1-score by 22.4%. Therefore, the use of the Smote-tomeklink method can increase accuracy, sensitivity, precision, and F1-score in the Random Forest method [11][32][33].

#### IV. CONCLUSION

This study applies the Smote-Tomeklink algorithm to the Random Forest method for the classification of diabetes. The application of Smote-Tomeklink can improve the performance of accuracy, sensitivity, precision, and F1-score in the Random Forest method. The combination of Random Forest and Smote-Tomeklink got the best accuracy, sensitivity, and precision compared to Smote and without sampling for the classification of diabetes. Where, there was an increase in performance indicators of 10.4% accuracy, 30.4% sensitivity, 13.7% precision, and 22.4 F1-score. Further research can apply Smote-Tomeklink to deal with the problem of data imbalance in multiclass data.

#### REFERENCES

- [1] O. Heranova, "Synthetic Minority Oversampling Technique pada Averaged One Dependence Estimators untuk Klasifikasi Credit Scoring," *Jurnal RESTI (Rekayasa Sistem dan Teknologi Informatika)*, vol. 3, no. 3, pp. 443–450, 2019, doi: 10.29207/resti.v3i3.1275.
- [2] T. Zhu, Y. Lin, and Y. Liu, "Synthetic minority oversampling technique for multiclass imbalance problems," *Pattern Recognition*, vol. 72, pp. 327–340, Dec. 2017, doi: 10.1016/j.patcog.2017.07.024.
- [3] F. Last, G. Douzas, and F. Bacao, "Oversampling for Imbalanced Learning Based on K-Means and SMOTE," no. November, 2017, [Online]. Available: <http://arxiv.org/abs/1711.00837>.
- [4] G. A. Pradipta, R. Wardoyo, A. Musdholifah, and I. N. H. Sanjaya, "Radius-SMOTE: A New Oversampling Technique of Minority Samples Based on Radius Distance for Learning from Imbalanced Data," *IEEE Access*, vol. 9, pp. 74763–74777, 2021, doi: 10.1109/ACCESS.2021.3080316.
- [5] M. Kamaladevi, V. Venkataraman, and K. R. Sekar, "Tomek link Undersampling with Stacked Ensemble classifier for Imbalanced data classification," vol. 25, no. 4, pp. 2182–2190, 2021.
- [6] W. C. Lin, C. F. Tsai, Y. H. Hu, and J. S. Jhang, "Clustering-based undersampling in class-imbalanced data," *Information Sciences*, vol. 409–410, pp. 17–26, 2017, doi: 10.1016/j.ins.2017.05.008.
- [7] Z. Xu, D. Shen, T. Nie, and Y. Kou, "A hybrid sampling algorithm combining M-SMOTE and ENN based on Random Forest for medical imbalanced data," *Journal of Biomedical Informatics*, p. 103465, 2020, doi: 10.1016/j.jbi.2020.103465.
- [8] E. AT, A. M, A.-M. F, and S. M, "Classification of Imbalance Data using Tomek Link (T-Link) Combined with Random Under-sampling (RUS) as a Data Reduction Method," *Global Journal of Technology and Optimization*, vol. 01, no. S1, 2016, doi: 10.4172/2229-8711.s1111.
- [9] Z. Wang, C. Wu, K. Zheng, X. Niu, and X. Wang, "SMOTETomek-Based Resampling for Personality Recognition," *IEEE Access*, vol. 7, pp. 129678–129689, 2019, doi: 10.1109/ACCESS.2019.2940061.
- [10] N. V Chawla, K. W. Bowyer, and L. O. Hall, "SMOTE: Synthetic Minority Over-sampling TEchnique," *Journal of Artificial Intelligence Research*, vol. 16, pp. 341–378, 2002.
- [11] H. Hairani, K. E. Saputro, and S. Fadli, "K-means-SMOTE for handling class imbalance in the classification of diabetes with C4.5, SVM, and naive Bayes," *Jurnal Teknologi dan Sistem Komputer*, vol. 8, no. 2, pp. 89–93, 2020, doi: 10.14710/jtsiskom.8.2.2020.89-93.
- [12] I. Tomek, "Tomek Link: Two Modifications of CNN," *IEEE Trans. Systems, Man and Cybernetics*, pp. 769–772, 1976, [Online]. Available: <https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&number=4309452>.
- [13] R. Kaur, "Predicting diabetes by adopting classification approach in data mining," *International Journal on Informatics Visualization*, vol. 3, no. 2–2, pp. 218–221, 2019, doi: 10.30630/ijov.3.2-2.229.
- [14] A. Azrar, M. Awais, Y. Ali, and K. Zaheer, "Data mining models comparison for diabetes prediction," *International Journal of Advanced Computer Science and Applications*, vol. 9, no. 8, pp. 320–323, 2018, doi: 10.14569/ijcasc.2018.090841.
- [15] S. Barik, S. Mohanty, S. Mohanty, and D. Singh, "Analysis of prediction accuracy of diabetes using classifier and hybrid machine learning techniques," *Smart Innovation, Systems and Technologies*, vol. 153, no. January, pp. 399–409, 2021, doi: 10.1007/978-981-15-6202-0\_41.
- [16] H. Hairani, M. Innuddin, and M. Rahardi, "Accuracy Enhancement of Correlated Naive Bayes Method by Using Correlation Feature Selection (CFS) for Health Data Classification," in *2020 3rd International Conference on Information and Communications Technology (ICOIACT)*, 2020, pp. 51–55, doi: 10.1109/ICOIACT50329.2020.9332021.
- [17] C. Fiarni, E. M. Sipayung, and S. Maemunah, "Analysis and prediction of diabetes complication disease using data mining algorithm," *Procedia Computer Science*, vol. 161, pp. 449–457, 2019, doi: 10.1016/j.procs.2019.11.144.
- [18] Erlin, Y. N. Marlim, Junadhi, L. Suryati, and N. Agustina, "Early Detection of Diabetes Using Machine Learning with Logistic Regression Algorithm," *Jurnal Nasional Teknik Elektro dan Teknologi Informasi*, vol. 11, no. 2, pp. 88–96, 2022.
- [19] C. Azad, B. Bhushan, R. Sharma, A. Shankar, K. K. Singh, and A. Khamparia, "Prediction model using SMOTE, genetic algorithm and decision tree (PMSGD) for classification of diabetes mellitus," *Multimedia Systems*, vol. 28, no. 4, pp. 1289–1307, 2022, doi: 10.1007/s00530-021-00817-2.
- [20] X. Shi, T. Qu, G. Van Pottelbergh, M. van den Akker, and B. De Moor, "A Resampling Method to Improve the Prognostic Model of End-Stage Kidney Disease: A Better Strategy for Imbalanced Data," *Frontiers in Medicine*, vol. 9, no. March, pp. 1–9, 2022, doi: 10.3389/fmed.2022.730748.
- [21] K. Wang *et al.*, "Improving risk identification of adverse outcomes in chronic heart failure using smote +enn and machine learning," *Risk Management and Healthcare Policy*, vol. 14, no. May, pp. 2453–2463, 2021, doi: 10.2147/RMHP.S310295.
- [22] H. Kaur, H. S. Pannu, and A. K. Malhi, "A systematic review on imbalanced data challenges in machine learning: Applications and solutions," *ACM Computing Surveys*, vol. 52, no. 4. Association for Computing Machinery, pp. 1–34, Aug. 01, 2019, doi: 10.1145/3343440.
- [23] K. Guo, X. Wan, L. Liu, Z. Gao, and M. Yang, "Fault diagnosis of intelligent production line based on digital twin and improved random forest," *Applied Sciences (Switzerland)*, vol. 11, no. 16, pp. 1–18, 2021, doi: 10.3390/app1167733.
- [24] J. Chen, H. Huang, A. G. Cohn, D. Zhang, and M. Zhou, "Machine learning-based classification of rock discontinuity trace: SMOTE oversampling integrated with GBT ensemble learning," *International Journal of Mining Science and Technology*, vol. 32, no. 2, pp. 309–322, 2021, doi: 10.1016/j.ijmst.2021.08.004.
- [25] E. F. Swana, W. Doorsamy, and P. Bokoro, "Tomek Link and SMOTE Approaches for Machine Fault Classification with an Imbalanced Dataset," *Sensors*, vol. 22, no. 9, 2022, doi: 10.3390/s22093246.
- [26] Y. Sun, H. Zhang, T. Zhao, Z. Zou, B. Shen, and L. Yang, "A New Convolutional Neural Network with Random Forest Method for Hydrogen Sensor Fault Diagnosis," *IEEE Access*, vol. 8, pp. 85421–85430, 2020, doi: 10.1109/ACCESS.2020.2992231.
- [27] H. Hartono and E. Ongko, "Avoiding Overfitting dan Overlapping in Handling Class Imbalanced Using Hybrid Approach with Smoothed Bootstrap Resampling and Feature Selection," *International Journal on Informatics Visualization*, vol. 6, no. June, pp. 343–348, 2022.
- [28] H. Hairani, A. Anggrawan, A. I. Wathan, K. A. Latif, K. Marzuki, and M. Zulfikri, "The Abstract of Thesis Classifier by Using Naive Bayes Method," in *Proceedings - 2021 International Conference on Software Engineering and Computer Systems and 4th International Conference on Computational Science and Information Management, ICSECS-ICOCSIM 2021*, 2021, no. August, pp. 312–315, doi: 10.1109/ICSECS52883.2021.00063.
- [29] A. Luque, A. Carrasco, A. Martín, and A. de las Heras, "The impact of class imbalance in classification performance metrics based on the binary confusion matrix," *Pattern Recognition*, vol. 91, pp. 216–231, 2019, doi: 10.1016/j.patcog.2019.02.023.
- [30] H. Qteat and M. Awad, "Using Hybrid Model of Particle Swarm

- Optimization and Multi-Layer Perceptron Neural Networks for Classification of Diabetes,” *International Journal of Intelligent Engineering and Systems*, vol. 14, no. 3, pp. 11–22, 2021, doi: 10.22266/ijies2021.0630.02.
- [31] H. Hanafi, A. H. Muhammad, I. Verawati, and R. Hardi, “An Intrusion Detection System Using SDAE to Enhance Dimensional Reduction in Machine Learning,” *International Journal on Informatics Visualization*, vol. 6, no. June, pp. 306–316, 2022.
- [32] H. Hairani, A. S. Suweleh, and D. Susilowaty, “Penanganan Ketidak Seimbangan Kelas Menggunakan Pendekatan Level Data,” *MATRIK : Jurnal Manajemen, Teknik Informatika dan Rekayasa Komputer*, vol. 20, no. 1, pp. 109–116, 2020, doi: 10.30812/matrik.v20i1.846.
- [33] M. Y. Thanoun, M. T. Yaseen, and A. M. Aleesa, “Development of Intelligent Parkinson Disease Detection System Based on Machine Learning Techniques Using Speech Signal,” *International Journal on Advanced Science, Engineering and Information Technology*, vol. 11, no. 1, pp. 388–392, 2021.



Home > User > Author > Submissions > #1069 > Summary

## #1069 Summary

**SUMMARY** REVIEW EDITING

### Submission

Authors	Hairani Hairani, Anthony Anggrawan, Dadang Priyanto	
Title	Improvement Performance of the Random Forest Method on Unbalanced Diabetes Data Classification Using Smote-Tomek Link	
Original file	<a href="#">1069-2305-1-SM.DOCX</a> 2022-08-01	
Supp. files	None	<a href="#">ADD A SUPPLEMENTARY FILE</a>
Submitter	Hairani Hairani	
Date submitted	August 1, 2022 - 12:34 PM	
Section	Articles	
Editor	Alde Alanda	

### Status

Status	In Editing
Initiated	2022-12-13
Last modified	2022-12-20

### Submission Metadata

[EDIT METADATA](#)

#### Authors

Name	Hairani Hairani
ORCID iD	<a href="http://orcid.org/0000-0002-6756-5896">http://orcid.org/0000-0002-6756-5896</a>
Affiliation	Universitas Bumigora
Country	Indonesia
Bio Statement	—

Principal contact for editorial correspondence.

Name	Anthony Anggrawan
Affiliation	Universitas Bumigora
Country	Indonesia
Bio Statement	—

Name	Dadang Priyanto
Affiliation	Universitas Bumigora
Country	Indonesia
Bio Statement	—

#### Title and Abstract

Title Improvement Performance of the Random Forest Method on Unbalanced Diabetes Data Classification Using Smote-Tomek Link

Abstract  
Most of the health data contained unbalanced data that affected the performance of the classification method. Unbalanced data causes the classification method to more easily classify the majority data and ignore the minority class. One of the health data that has unbalanced data is Pima Indian Diabetes. Diabetes is a deadly disease caused by the body's inability to produce enough insulin. Complications of diabetes can cause heart attacks and strokes. Early diagnosis of diabetes is needed to minimize the occurrence of more severe complications. In the diabetes dataset used, there is an imbalance of data between positive and negative diabetes classes. Diabetes negative class data (500 data) is more than diabetes positive class (268) so that it can affect the performance of the classification method. Therefore, this study aims to apply the Smote-Tomeklink and Random Forest methods in the classification of diabetes. The research methodology used is the collection of diabetes data obtained from Kaggle as many as 768 data with 8 input attributes and 1 output attribute as a class, pre-processing data is used to balance the dataset with Smote-Tomeklink, classification using the random

#### QUICK MENU

- » [Editorial Team](#)
- » [Focus & Scope](#)
- » [Indexing](#)
- » [Author Guidelines](#)
- » [Peer Review Process](#)
- » [Author Fees](#)
- » [Publication Ethics](#)
- » [Online Submission](#)
- » [Open Access Statement](#)
- » [Plagiarism Policy](#)
- » [Special Issues](#)
- » [Licensing terms](#)
- » [Contact](#)



#### International Journal on Informatics Visualization



#### REQUEST INDEXING

- » **SCOPUS (ACCEPTED)**
  - » Submission Received: **March 3, 2020**
  - » Submission Accepted: **July 30, 2020**
  - » [SCOPUS CiteScore Tracker 2020](#)
- » **WoS / Web of Science**
  - » Latest submission: September 16, 2018
  - » [Web of Science](#) Citation Analysis
- » **IET INSPEC**



forest method, and performance evaluation based on accuracy, sensitivity, precision, and F1-score. Based on the tests carried out by dividing data using 10-fold cross-validation, the Random forest algorithm with Smote-TomekLink gets the highest accuracy, sensitivity, precision, and F1-score compared to Random Forest with Smote. The Random Forest algorithm with Smote-TomekLink has 86.4% accuracy, 88.2% sensitivity, 82.3% precision, and 85.1% F1-score. Thus, using Smote-TomekLink can improve the performance of the random forest method based on accuracy, sensitivity, precision, and F1-score.

## Indexing

Keywords Class Imbalance; Smote-TomekLink; Random Forest Method; Diabetes Disease  
Language en

## Supporting Agencies

Agencies —

## References

- References
- [1] O. Heranova, "Synthetic Minority Oversampling Technique pada Averaged One Dependence Estimators untuk Klasifikasi Credit Scoring," *Jurnal RESTI (Rekayasa Sistem dan Teknologi Informasi)*, vol. 3, no. 3, pp. 443–450, 2019, doi: 10.29207/resti.v3i3.1275.
  - [2] T. Zhu, Y. Lin, and Y. Liu, "Synthetic minority oversampling technique for multiclass imbalance problems," *Pattern Recognition*, vol. 72, pp. 327–340, Dec. 2017, doi: 10.1016/j.patcog.2017.07.024.
  - [3] F. Last, G. Douzas, and F. Bacao, "Oversampling for Imbalanced Learning Based on K-Means and SMOTE," no. November, 2017, [Online]. Available: <http://arxiv.org/abs/1711.00837>.
  - [4] G. A. Pradipta, R. Wardoyo, A. Musdholifah, and I. N. H. Sanjaya, "Radius-SMOTE: A New Oversampling Technique of Minority Samples Based on Radius Distance for Learning from Imbalanced Data," *IEEE Access*, vol. 9, pp. 74763–74777, 2021, doi: 10.1109/ACCESS.2021.3080316.
  - [5] M. Kamaladevi, V. Venkataraman, and K. R. Sekar, "Tomek link Undersampling with Stacked Ensemble classifier for Imbalanced data classification," vol. 25, no. 4, pp. 2182–2190, 2021.
  - [6] W. C. Lin, C. F. Tsai, Y. H. Hu, and J. S. Jhang, "Clustering-based undersampling in class-imbalanced data," *Information Sciences*, vol. 409–410, pp. 17–26, 2017, doi: 10.1016/j.ins.2017.05.008.
  - [7] Z. Xu, D. Shen, T. Nie, and Y. Kou, "A hybrid sampling algorithm combining M-SMOTE and ENN based on Random Forest for medical imbalanced data," *Journal of Biomedical Informatics*, p. 103465, 2020, doi: 10.1016/j.jbi.2020.103465.
  - [8] E. AT, A. M, A.-M. F, and S. M, "Classification of Imbalance Data using Tomek Link (T-Link) Combined with Random Under-sampling (RUS) as a Data Reduction Method," *Global Journal of Technology and Optimization*, vol. 01, no. S1, 2016, doi: 10.4172/2229-8711.s1111.
  - [9] Z. Wang, C. Wu, K. Zheng, X. Niu, and X. Wang, "SMOTETomek-Based Resampling for Personality Recognition," *IEEE Access*, vol. 7, pp. 129678–129689, 2019, doi: 10.1109/ACCESS.2019.2940061.
  - [10] N. V. Chawla, K. W. Bowyer, and L. O. Hall, "SMOTE: Synthetic Minority Over-sampling Technique," *Journal of Artificial Intelligence Research*, vol. 16, pp. 341–378, 2002.
  - [11] H. Hairani, K. E. Saputro, and S. Fadli, "K-means-SMOTE for handling class imbalance in the classification of diabetes with C4.5, SVM, and naive Bayes," *Jurnal Teknologi dan Sistem Komputer*, vol. 8, no. 2, pp. 89–93, 2020, doi: 10.14710/jtsiskom.8.2.2020.89-93.
  - [12] I. Tomek, "Tomek Link: Two Modifications of CNN," *IEEE Trans. Systems, Man and Cybernetics*, pp. 769–772, 1976, [Online]. Available: <https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&number=4309452>.
  - [13] R. Kaur, "Predicting diabetes by adopting classification approach in data mining," *International Journal on Informatics Visualization*, vol. 3, no. 2–2, pp. 218–221, 2019, doi: 10.30630/ijov.3.2-2.229.
  - [14] A. Azrar, M. Awais, Y. Ali, and K. Zaheer, "Data mining models comparison for diabetes prediction," *International Journal of Advanced Computer Science and Applications*, vol. 9, no. 8, pp. 320–323, 2018, doi: 10.14569/ijacsa.2018.090841.
  - [15] S. Barik, S. Mohanty, S. Mohanty, and D. Singh, "Analysis of prediction accuracy of diabetes using classifier and hybrid machine learning techniques," *Smart Innovation, Systems and Technologies*, vol. 153, no. January, pp. 399–409, 2021, doi: 10.1007/978-981-15-6202-0\_41.
  - [16] H. Hairani, M. Innuddin, and M. Rahardi, "Accuracy Enhancement of Correlated Naive Bayes Method by Using Correlation Feature Selection (CFS) for Health Data Classification," in *2020 3rd International Conference on Information and Communications Technology (ICOIACT)*, 2020, pp. 51–55, doi: 10.1109/ICOIACT50329.2020.9332021.
  - [17] C. Fiarni, E. M. Sipayung, and S. Maemunah, "Analysis and prediction of diabetes complication disease using data mining algorithm," *Procedia Computer Science*, vol. 161, pp. 449–457, 2019, doi: 10.1016/j.procs.2019.11.144.
  - [18] Erlin, Y. N. Marlim, Junadhi, L. Suryati, and N. Agustina, "Early Detection of Diabetes Using Machine Learning with Logistic Regression Algorithm," *Jurnal Nasional Teknik Elektro dan Teknologi Informasi*, vol. 11, no. 2, pp. 88–96, 2022.
  - [19] C. Azad, B. Bhushan, R. Sharma, A. Shankar, K. K. Singh, and A. Khamparia, "Prediction model using SMOTE, genetic algorithm and decision tree (PMSGD) for classification of diabetes mellitus," *Multimedia Systems*, vol. 28, no. 4, pp. 1289–1307, 2022, doi: 10.1007/s00530-021-00817-2.
  - [20] X. Shi, T. Qu, G. Van Pottelbergh, M. van den Akker, and B. De Moor, "A Resampling Method to Improve the Prognostic Model of End-Stage Kidney Disease: A Better Strategy for Imbalanced Data," *Frontiers in Medicine*, vol. 9, no. March, pp. 1–9, 2022, doi: 10.3389/fmed.2022.730748.
  - [21] K. Wang et al., "Improving risk identification of adverse outcomes in chronic heart failure using smote +enn and machine learning," *Risk Management and Healthcare Policy*, vol. 14, no. May, pp. 2453–2463, 2021, doi: 10.2147/RMHP.S310295.
  - [22] H. Kaur, H. S. Pannu, and A. K. Malhi, "A systematic review on imbalanced data challenges in machine learning: Applications and solutions," *ACM Computing Surveys*, vol. 52, no. 4. Association for Computing Machinery, pp. 1–34, Aug. 01, 2019, doi: 10.1145/3343440.
  - [23] K. Guo, X. Wan, L. Liu, Z. Gao, and M. Yang, "Fault diagnosis of intelligent production line based on digital twin and improved random forest," *Applied Sciences (Switzerland)*, vol. 11, no. 16, pp. 1–18, 2021, doi: 10.3390/app11167733.
  - [24] J. Chen, H. Huang, A. G. Cohn, D. Zhang, and M. Zhou, "Machine learning-based classification of rock discontinuity trace: SMOTE oversampling integrated with GBT ensemble learning," *International Journal of Mining Science and Technology*, vol. 32, no. 2, pp. 309–322, 2021, doi: 10.1016/j.ijmst.2021.08.004.
  - [25] E. F. Swana, W. Doorsamy, and P. Bokoro, "Tomek Link and SMOTE Approaches for Machine Fault Classification with an Imbalanced Dataset," *Sensors*, vol. 22, no. 9, 2022, doi: 10.3390/s22093246.
  - [26] Y. Sun, H. Zhang, T. Zhao, Z. Zou, B. Shen, and L. Yang, "A New Convolutional Neural Network with

» Added to review: **May 29, 2020**

» **Ei COMPENDEX**

» Submission: **February 10, 2021**

### PUBLICATION PARTNERS



### USER

You are logged in as...

**hairani10**

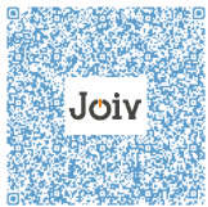
- » [My Profile](#)
- » [Log Out](#)

### AUTHOR

Submissions

- » [Active \(1\)](#)
- » [Archive \(0\)](#)
- » [New Submission](#)

- [ ] g, , , g.  
Random Forest Method for Hydrogen Sensor Fault Diagnosis," IEEE Access, vol. 8, pp. 85421–85430, 2020, doi: 10.1109/ACCESS.2020.2992231.
- [27] H. Hartono and E. Ongko, "Avoiding Overfitting dan Overlapping in Handling Class Imbalanced Using Hybrid Approach with Smoothed Bootstrap Resampling and Feature Selection," International Journal on Informatics Visualization, vol. 6, no. June, pp. 343–348, 2022.
- [28] H. Hairani, A. Anggrawan, A. I. Wathan, K. A. Latif, K. Marzuki, and M. Zulfikri, "The Abstract of Thesis Classifier by Using Naive Bayes Method," in Proceedings - 2021 International Conference on Software Engineering and Computer Systems and 4th International Conference on Computational Science and Information Management, ICSECS-ICOCSIM 2021, 2021, no. August, pp. 312–315, doi: 10.1109/ICSECS52883.2021.00063.
- [29] A. Luque, A. Carrasco, A. Martín, and A. de las Heras, "The impact of class imbalance in classification performance metrics based on the binary confusion matrix," Pattern Recognition, vol. 91, pp. 216–231, 2019, doi: 10.1016/j.patcog.2019.02.023.
- [30] H. Qteat and M. Awad, "Using Hybrid Model of Particle Swarm Optimization and Multi-Layer Perceptron Neural Networks for Classification of Diabetes," International Journal of Intelligent Engineering and Systems, vol. 14, no. 3, pp. 11–22, 2021, doi: 10.22266/ijies2021.0630.02.
- [31] H. Hanafi, A. H. Muhammad, I. Verawati, and R. Hardi, "An Intrusion Detection System Using SDAE to Enhance Dimensional Reduction in Machine Learning," International Journal on Informatics Visualization, vol. 6, no. June, pp. 306–316, 2022.
- [32] H. Hairani, A. S. Suweleh, and D. Susilowaty, "Penanganan Ketidak Seimbangan Kelas Menggunakan Pendekatan Level Data," MATRIK: Jurnal Manajemen, Teknik Informatika dan Rekayasa Komputer, vol. 20, no. 1, pp. 109–116, 2020, doi: 10.30812/matrik.v20i1.846.
- [33] M. Y. Thanoun, M. T. Yaseen, and A. M. Aleesa, "Development of Intelligent Parkinson Disease Detection System Based on Machine Learning Techniques Using Speech Signal," International Journal on Advanced Science, Engineering and Information Technology, vol. 11, no. 1, pp. 388–392, 2021.



**JOIV : International Journal on Informatics Visualization**

ISSN **2549-9610 (print) | 2549-9904 (online)**

Organized by [Department of Information Technology - Politeknik Negeri Padang](#), and [Institute of Visual Informatics - UKM](#) and [Soft Computing and Data Mining Centre - UTHM](#)

W : <http://joiv.org>

E : [joiv@pnp.ac.id](mailto:joiv@pnp.ac.id), [hidra@pnp.ac.id](mailto:hidra@pnp.ac.id), [rahmat@pnp.ac.id](mailto:rahmat@pnp.ac.id)

[View JOIV Stats](#)



is licensed under a [Creative Commons Attribution-ShareAlike 4.0 International License](#).

The screenshot shows a Gmail interface with a search bar at the top containing 'joiv'. The left sidebar includes navigation options like Mail (99+), Compose, Mail, Chat, Spaces, and Meet. The main content area displays an email with the following text:

Thank you for submitting the manuscript, "Improvement Performance of the Random Forest Method on Unbalanced Diabetes Data Classification Using Smote-Tomek Link" to **JOIV** : International Journal on Informatics Visualization. With the online journal management system that we are using, you will be able to track its progress through the editorial process by logging in to the journal web site:

Manuscript URL: <https://joiv.org/index.php/joiv/author/submission/1069>  
Username: hairani10

If you have any questions, please contact me. Thank you for considering this journal as a venue for your work.

In addition, due to the rising costs of academic publications, starting 2021, publication fees shall be implemented to all accepted papers. For more details, please email to [joiv \[at\] pnp.ac.id](mailto:joiv@pnp.ac.id). This journal charges the following author fees (Article Publication Fee):

- Indonesian authors: 3.500.000 IDR per article
- International authors: 280 USD per article

This fee includes:

- DOI registration for each paper
- Checking the article similarity by turnitin
- English proofreading

Editor  
2022-08-15 09:08 PM

Subject: [JOIV] Editor Decision

[DELETE](#)

Hairani Hairani:

We have reached a decision regarding your submission to JOIV : International Journal on Informatics Visualization, "Improvement Performance of the Random Forest Method on Unbalanced Diabetes Data Classification Using Smote-Tomek Link".

Our decision is: Revisions Required

initial review

JOIV requires authors to submit at least 6 pages article, excluding the references.

Alde Alanda  
(Scopus ID: 57203718850); Politeknik Negeri Padang, Sumatera Barat  
Phone 81267775707  
Fax 81267775707  
aldealanda@gmail.com

Alde Alanda

<http://joiv.org/index.php/joiv>

Author  
2022-08-17 06:04 AM

Subject: Improvement Performance of the Random Forest Method on Unbalanced Diabetes Data Classification Using Smote-Tomek Link

[DELETE](#)

Dear Editor,

We have fixed it according to the suggestion from the editor.

Regards

<http://joiv.org/index.php/joiv>

Editor  
2022-10-21 01:23 AM

Subject: [JOIV] Editor Decision

[DELETE](#)

Hairani Hairani:

We have reached a decision regarding your submission to JOIV : International Journal on Informatics Visualization, "Improvement Performance of the Random Forest Method on Unbalanced Diabetes Data Classification Using Smote-Tomek Link".

Our decision is: Revisions Required

Alde Alanda  
(Scopus ID: 57203718850); Politeknik Negeri Padang, Sumatera Barat  
Phone 81267775707  
Fax 81267775707  
aldealanda@gmail.com

Alde Alanda

Reviewer A:

From the manuscript it is not clear why this paper is written, what is the difference of the present submitted paper from the previous studies? Therefore, the aim of study is not clear and it should be extended and revised at end of introduction part.

Therefore, give details about the target of the study, clarify the needs of the study, and explain the difference of the submitted work from the previous studies. Explain all these questions at the end of the introduction part of the paper. Without this a reader do not know why do we need this and who needs this.

Literature review has to be improved by adding some recent literature. Put you research in the context of a bigger picture. I suggest the following references to improve your literature review: "The role of data mining techniques and tools in big data management in healthcare field", Sustainable Engineering and Innovation, vol. 4, no. 1, pp. 58-65, Feb. 2022.; "Analysis of student performances in online and face-to-face learning: A case study from a Bosnian public university", Heritage and Sustainable Development, vol. 4, no. 2, pp. 87-94, Jul. 2022.; "Bacterial endophytes of aloe vera and their potential applications", Heritage and Sustainable Development, vol. 4, no. 1, pp. 32-41, Jul. 2022.

Results and discussion part has to be improved in a way to compare your finding with previous studies. I suggest you to prepare a table in which you are going to provide comparative results and indicate coherent an incoherent points.

Some typos were noticed and I suggest a proofread.

I suggest publication of this paper only after the above required improvements are implemented.

<http://joiv.org/index.php/joiv>

Author  
2022-10-22 09:43 PM

Subject: Improvement Performance of the Random Forest Method on Unbalanced Diabetes Data Classification Using Smote Tomek Link

[DELETE](#)

2022 10 22 09:43 PM

Diabetes Data Classification Using Smote-Tomek Link

---

Dear Editors,

We have made improvements based on suggestions from reviewers. The improvements we've made, are highlighted in yellow.

Thank you.

---

<http://joiv.org/index.php/joiv>

Editor  
2022-11-01 12:58 AM

Subject: [JOIV] Editor Decision

[DELETE](#)

Hairani Hairani:

We have reached a decision regarding your submission to JOIV : International Journal on Informatics Visualization, "Improvement Performance of the Random Forest Method on Unbalanced Diabetes Data Classification Using Smote-Tomek Link".

Our decision is: Revisions Required

Alde Alanda  
(Scopus ID: 57203718850); Politeknik Negeri Padang, Sumatera Barat  
Phone 81267775707  
Fax 81267775707  
aldealanda@gmail.com

Alde Alanda

---

Reviewer A:

Provided missing details in this version of the manuscript make some improvement. An important part of the "big picture" is still missing, the work needs to be placed in a context of bigger picture by adding some adding some recent literature. Please update your Introduction part as suggested in previous comments. To improve this component I suggest the following references to improve your literature review: "The role of data mining techniques and tools in big data management in healthcare field ", Sustainable Engineering and Innovation, vol. 4, no. 1, pp. 58-65, Feb. 2022.; "Analysis of student performances in online and face-to-face learning: A case study from a Bosnian public university", Heritage and Sustainable Development, vol. 4, no. 2, pp. 87-94, Jul. 2022.; "Bacterial endophytes of aloe vera and their potential applications", Heritage and Sustainable Development, vol. 4, no. 1, pp. 32-41, Jul. 2022.

I suggest publication of this paper only after the above required improvements are implemented.

---

<http://joiv.org/index.php/joiv>

Author  
2022-11-03 03:05 PM

Subject: Improvement Performance of the Random Forest Method on Unbalanced Diabetes Data Classification Using Smote-Tomek Link

[DELETE](#)

Dear Editor,

Thank you reviewers for your suggestions to improve the quality of our manuscript. We have made improvements based on suggestions from reviewers by adding some suggested review literature. Revised manuscripts are highlighted in yellow.

Best Regard

---

<http://joiv.org/index.php/joiv>

Editor  
2022-12-13 01:06 AM

Subject: [JOIV] Editor Decision

[DELETE](#)

Hairani Hairani:

We have reached a decision regarding your submission to JOIV : International Journal on Informatics Visualization, "Improvement Performance of the Random Forest Method on Unbalanced Diabetes Data Classification Using Smote-Tomek Link".

Our decision is to: Accept Submission

Publication fees shall be implemented to all accepted papers. For more details, please email to joiv [at] pnp.ac.id. This journal charges the following author fees (Article Publication Fee):

- Indonesian authors: 3.500.000 IDR per article
- International authors: 280 USD per article

This fee includes:

- DOI registration for each paper
- Checking the article similarity by turnitin
- English proofreading

Alde Alanda  
(Scopus ID: 57203718850); Politeknik Negeri Padang, Sumatera Barat  
Phone 81267775707  
Fax 81267775707  
aldealanda@gmail.com

Alde Alanda

---

<http://joiv.org/index.php/joiv>

Close

# Improvement Performance of the Random Forest Method on Unbalanced Diabetes Data Classification Using Smote-Tomek Link

Hairani Hairani<sup>a\*</sup>, Anthony Anggrawan<sup>b</sup>, Dadang Priyanto<sup>c</sup>

<sup>abc</sup> Department of Computer Science, Universitas Bumigora, Mataram, 83127, Indonesia

Corresponding author: Hairani@universitasbumigora.ac.id

**Abstract**— Most of the health data contained unbalanced data that affected the performance of the classification method. Unbalanced data causes the classification method to more easily classify the majority data and ignore the minority class. One of the health data that has unbalanced data is Pima Indian Diabetes. Diabetes is a deadly disease caused by the body's inability to produce enough insulin. Complications of diabetes can cause heart attacks and strokes. Early diagnosis of diabetes is needed to minimize the occurrence of more severe complications. In the diabetes dataset used, there is an imbalance of data between positive and negative diabetes classes. Diabetes negative class data (500 data) is more than diabetes positive class (268) so that it can affect the performance of the classification method. Therefore, this study aims to apply the Smote-TomekLink and Random Forest methods in the classification of diabetes. The research methodology used is the collection of diabetes data obtained from Kaggle as many as 768 data with 8 input attributes and 1 output attribute as a class, pre-processing data is used to balance the dataset with Smote-TomekLink, classification using the random forest method, and performance evaluation based on accuracy, sensitivity, precision, and F1-score. Based on the tests carried out by dividing data using 10-fold cross-validation, the Random forest algorithm with Smote-TomekLink gets the highest accuracy, sensitivity, precision, and F1-score compared to Random Forest with Smote. The Random Forest algorithm with Smote-TomekLink has 86.4% accuracy, 88.2% sensitivity, 82.3% precision, and 85.1% F1-score. Thus, using Smote-TomekLink can improve the performance of the random forest method based on accuracy, sensitivity, precision, and F1-score.

**Keywords**— Class Imbalance; Smote-TomekLink; Random Forest Method; Diabetest Disease.

*Manuscript received dd mm yyyy; revised dd mm yyyy; accepted dd mm yyyy. Date of publication dd mm yyyy.  
International Journal on Informatics Visualization is licensed under a Creative Commons Attribution-Share Alike 4.0 International License.*



## I. INTRODUCTION

Most of the Health data contained unbalanced data that affected the performance of the classification method. Unbalanced data causes the classification method to more easily classify the majority data and ignore the minority class. One of the health data that has unbalanced data is Pima Indian Diabetes. Diabetes is a deadly disease caused by increased blood sugar in the body. Diabetes is caused by the body's inability to produce enough insulin. Complications of diabetes can cause heart attacks and strokes. One way to improve the performance of the classification method is to handle balanced data by adding minority data so that the number is equal to the majority class. The diabetes dataset has 768 instances of data. However, the problem is that there is an imbalance of data in the dataset, namely the negative diabetes class with 500 data (majority class), while the positive diabetes class with 268 data (minority class). Data imbalance is the amount of data in one class more than in the other class. The problem of data imbalance causes the classification method to be more

dominant in classifying the majority class than the minority class, or in other words, the classification method ignores the minority class. The problem of unbalanced data can be handled with a data sampling approach.

Several data sampling methods that can be used to solve the problem of data imbalance are oversampling [1][2], [3][4], undersampling [5][6], and Hybrid Sampling[6],[7]. Oversampling works by adding the minority class, while Undersampling works by removing the majority class so as to produce balanced data. However, both methods have their respective weaknesses. The weakness of the oversampling method is that there are too many repetitions of samples that can cause overfitting of the classification method, while the weakness of undersampling is that it will lose information from most of the samples in the dataset and cannot take full advantage of the available information[9].

To avoid overfitting the oversampling method, the Smote method was developed to overcome these weaknesses. Smote is an oversampling method to generate new synthesis training data by linear interpolation on minority classes[10]. However,

the Smote method has a weakness, namely overgeneralization, and the addition of a minority class randomly can generate noise data, because it does not differentiate between classes [11]. Therefore, the undersampling method is used to improve the performance of the oversampling method by cleaning the noise data in the majority class. The noise data is the majority class instance which is closest to the minority class instance. Usually, noise data reduces the level of accuracy for classification methods [5]. One method to remove noise data in the majority class is Tomeklink [12]. Tomeklink is an undersampling method that cleans noise data from the majority class which has similar characteristics and overlapping. However, Tomeklink only deletes instances defined as “Tomek Links” so that the analyzed data cannot be balanced and in its implementation the method is combined with other methods. Combining Tomeklink and Smote oversampling can improve accuracy better than individual performance [13].

Several previous studies that have discussed the classification of diabetes, namely Research [14] predicts diabetes using the k-NN method with an accuracy of 83%. The weakness of the research is that it does not address the problem of data imbalance. Research [15] classifying diabetes using the C4.5 method with an accuracy of 75.65%. The weakness of the research is that the accuracy obtained is low so that it can be improved, and also does not address the problem of data imbalance. Research [16] Using XGBoost to predict diabetes with 74% accuracy. The weakness of the research is that the accuracy obtained is low so that it can be improved, and also does not address the problem of data imbalance.

Research [17] using the Correlated Naïve Bayes method with correlation-based feature selection to predict diabetes with an accuracy of 69.51%. The weakness of the research is that the accuracy obtained is low so that it can be improved, and also does not address the problem of data imbalance. Research [18] using the C4.5 method for diabetes detection with an accuracy of 68%.

Research [19] used logistic regression and smote methods to detect diabetes with 82% accuracy, 81% precision, 79% recall, and 80% F1-score. The weakness of the research is that the accuracy is good but can be improved using Tomeklink to clean noise data in the majority class. Research [20] using the C4.5 and Smote methods to predict diabetes with 82% accuracy, 80% precision, and 86% sensitivity. Research [21] used logistic and Smote-ENN methods to predict kidney disease with 75.2% accuracy, 70.6% recall, 4.9% precision, and 30% F1-score. The weakness of the research is the low accuracy so that it can be improved using Tomeklink to clean noise data in the majority class. Research [22] SME-XGBoost with Smote-ENN for heart disease prediction with 80% AUC.

Several previous studies have applied various approaches to improve diabetes classification methods such as the oversampling approach with SMOTE. However, there are weaknesses in previous studies, namely the accuracy of the proposed method still ranges from 82% to 83% so that there is a gap to improve its accuracy. So, this study proposes the Smote-Tomeklink hybrid sampling method to overcome the imbalance in diabetes data, so as to improve the accuracy of the classification method.

Smote-Tomeklink is a good way to avoid the drawbacks of SMOTE and Tomeklink techniques [9]. The classification method used in this research is Random Forest. The Random

Forest method was chosen because it has several advantages, namely high accuracy [23], the ability to handle noise data, fast performance in training data, overfitting control, and easy to implement [24].

This study aims to apply the Smote-Tomeklink hybrid sampling method to balance the data on diabetes data so as to improve the performance of the Random forest classification method. Measurement of the performance of the random forest method based on accuracy, sensitivity (recall), precision, and F1-score

## II. MATERIALS AND METHOD

This research consists of several stages as shown in Figure 1.

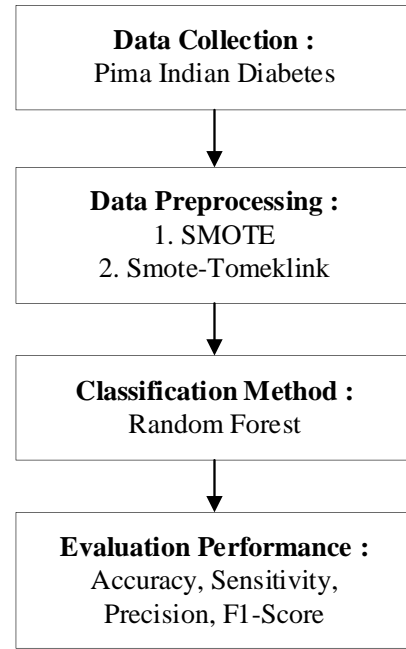


Fig 1. Research Stages

### A. Data Collection

The dataset used in this study is a diabetes dataset obtained from Kaggle, which consists of 768 instances and 9 attributes. The description of the attributes and the sample data used are shown respectively in Table I and Table II.

TABLE I  
DESCRIPTION ATRIBUT DATASET

No	Atribute	Description	Label
1	Pregnancies	Number of Pregnancy	X1
2	Glucose	Glucose level 2 hours after eating	X2
3	Blood Pressure	Blood Pressure	X3
4	Skin Thickness	Skin Thickness	X4
5	Insulin	Insulin	X5
6	BMI	Body Massa Index	X6
7	Diabetes Pedigree Function	Diabetes Pedigree Function	X7
8	Age	Age	X8
9	Outcome	Diabetes Status ( 1 = Positive Diabetes, 2 = Negative Diabetes	Y



TABLE III  
SAMPLE DATASET

No	X1	X2	X3	X4	X5	X6	X7	X8	Y
1	6	148	72	35	0	33.6	0.627	50	1
2	1	85	66	29	0	26.6	0.351	31	0
3	8	183	64	0	0	23.3	0.672	32	1
4	1	89	66	23	94	28.1	0.167	21	0
5	0	137	40	35	168	43.1	2.288	33	1
6	5	116	74	0	0	25.6	0.201	30	1
7	3	78	50	32	88	31	0.248	26	1
8	10	115	0	0	0	35.3	0.134	29	0
9	2	197	70	45	543	30.5	0.158	53	1
10	8	125	96	0	0	0	0.232	54	1
11	4	110	92	0	0	37.6	0.191	30	0
12	10	168	74	0	0	38	0.537	34	1
13	10	139	80	0	0	27.1	1.441	57	0
14	1	189	60	23	846	30.1	0.398	59	1
15	5	166	72	19	175	25.8	0.587	51	1
..	..	...	..	..	...	..	.....	..	..
754	0	181	88	44	510	43.3	0.222	26	1
755	8	154	78	32	0	32.4	0.443	45	1
756	1	128	88	39	110	36.5	1.057	37	1
757	7	137	90	41	0	32	0.391	39	0
758	0	123	72	0	0	36.3	0.258	52	1
759	1	106	76	0	0	37.5	0.197	26	0
760	6	190	92	0	0	35.5	0.278	66	1
761	2	88	58	26	16	28.4	0.766	22	0
762	9	170	74	31	0	44	0.403	43	1
763	9	89	62	0	0	22.5	0.142	33	0
764	10	101	76	48	180	32.9	0.171	63	0
765	2	122	70	27	0	36.8	0.34	27	0
766	5	121	72	23	112	26.2	0.245	30	0
767	1	126	60	0	0	30.1	0.349	47	1
768	1	93	70	31	0	30.4	0.315	23	0

## B. Data Pre-processing

Data Pre-processing is one of the important stages in data mining to improve the quality of datasets. This study focuses on dealing with unbalanced data contained in the diabetes dataset. The dataset used has 268 instances of negative diabetes and 500 instances of Positive Diabetes. The algorithms used to handle unbalanced data in the dataset are SMOTE (Synthetic Minority Oversampling Technique) and Smote-Tomeklink.

SMOTE is one of the most commonly used oversampling methods to solve the problem of data distribution imbalance in machine learning modeling. SMOTE aims to balance the distribution of classes by increasing the number of minority classes randomly by creating synthetic data for oversampling purposes [10]. Creating new data on the minority class using the equation (1).

$$Y' = Y^i + (Y^j - Y^i) * \gamma \quad (1)$$

$Y'$  is the representation of the addition of the minority class.  $Y^i$  is the representation of minority class,  $Y^j$  is a value chosen at random from the k-nearest neighbors of the minority class on  $Y^i$ , and  $\gamma$  is a value in a randomly selected vector with a range of 0 to 1 [2].

SMOTE generates new synthesis training data by linear interpolation for the minority class. Synthesis training data is generated by randomly selecting one or more of the k-nearest neighbors for each sample in the minority class as shown in Figure 2.

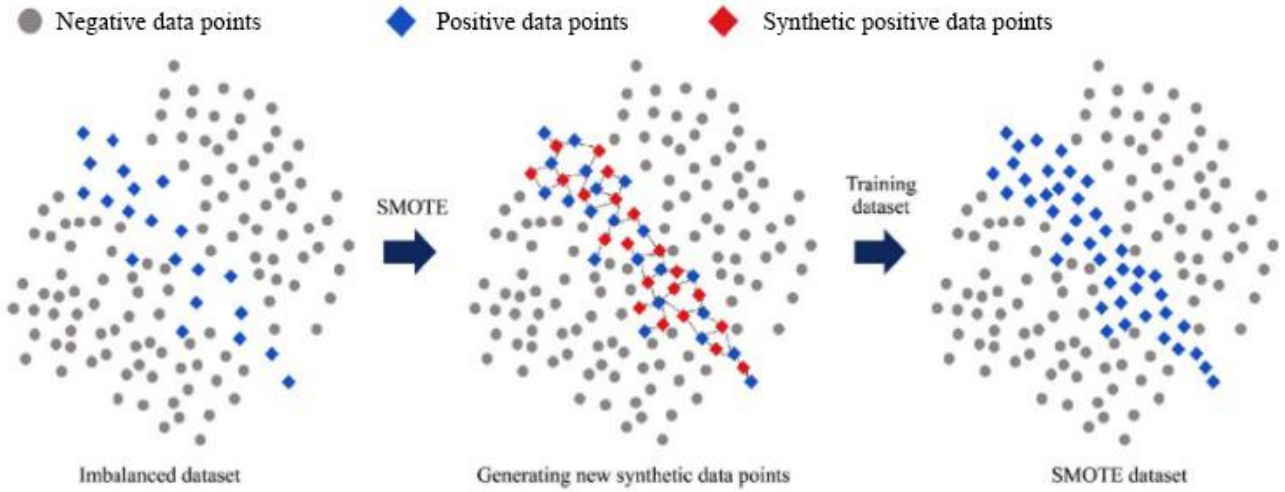


Fig 2. Process of Synthetic Minority Oversampling Technique (SMOTE) Algorithm [25]

Tomeklink is an undersampling method that cleans noise data from the majority class that has similar characteristics and overlapping [12]. Tomeklink works by eliminating the majority class instances that are closer to the minority class by applying the nearest neighbor rule to select instances. The combination of Tomeklink and Smote oversampling can improve accuracy better than individual performance [13].

## C. Random Forest Method

Random Forest is a decision tree-based ensemble learning method [26]. The Random Forest method has the advantages of high accuracy, the ability to handle noise data, fast performance in training data, overfitting control, and easy to implement [24]. The working process of the Random Forest method in classifying a data is shown in Figure 3.

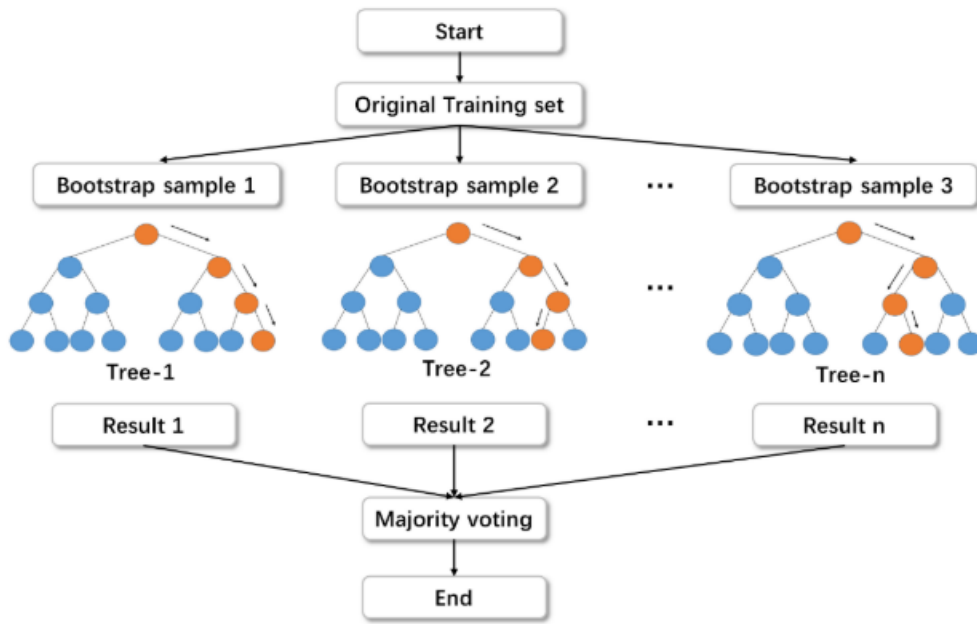


Fig 1. Working Process of Random Forest Method[24]

Figure 3 shows how the Random Forest algorithm works by creating a set of decision trees from a randomly selected subset, getting predictions from each decision tree, voting for each predicted outcome, and choosing the best prediction result based on the most votes assigned as final prediction

#### D. Evaluation Performance

Performance testing uses a confusion matrix table. The confusion matrix is a table that is used to describe the performance of the classification method on a dataset whose true value is known. The confusion matrix can visualize the amount of data that is classified as true and false as shown in the Table III[27].

TABLE III  
CONFUSION MATRIX

Actual	Predicted	
	Negative	Positive
Negative	TN	FP
Positive	FN	TP

Formula used to calculate *Accuracy* (6), *Sensitivity* (7), *Precision* (8) [28] [29][30], and *F1-score* (5)[31].

$$Accuracy = \frac{TP + TN}{TP + FN + TN + FP} \quad (6)$$

$$Sensitivity = \frac{TP}{TP + FN} \quad (7)$$

$$Precision = \frac{TP}{TP + FP} \quad (8)$$

$$F1-score = \frac{2 \times precision \times recall}{precision + recall} \quad (9)$$

True Positive (TP) is a class of positive diabetes that is predicted correctly. False Positive (FP) is a diabetes negative class but is predicted to be diabetes positive. True Negative (TN) is a diabetes negative class that is predicted correctly. False Negative (FN) is a positive diabetes class but is predicted to be diabetes negative.

### III. RESULT AND DISCUSSION

This research starts from the stages of data collection, data pre-processing, classification, and performance testing. The data used in this study is diabetes data obtained from Kaggle. The pre-processing of this study used the Smote and Smote-Tomeklink algorithms to deal with class imbalances in diabetes data. The classification method of this research is Random Forest. The performance test is based on accuracy, sensitivity, precision, and F1-score. The results of the comparison of the original data with the data from Smote and the results of Smote-Tomeklink are shown in Figure 4.

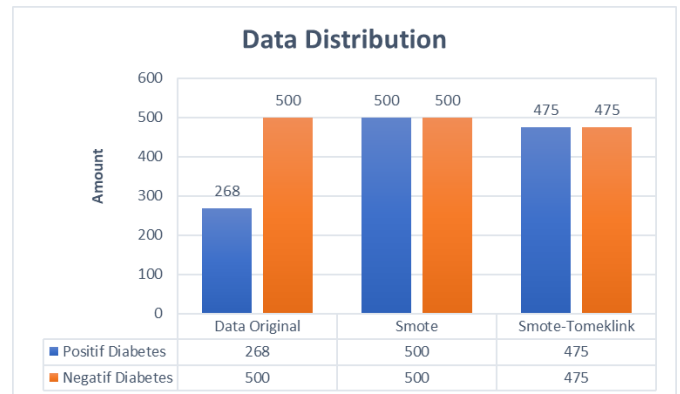


Fig 4. Data Distribution Result

The classification method of this research is Random Forest. Performance testing is based on accuracy, sensitivity, precision, and F1-score using a confusion matrix table. Based on testing the Random Forest method using 10-fold cross-validation, the

results obtained in the form of a confusion matrix table as shown in Table IV for the Random Forest method on the original data, Table V for the results of the Random Forest method with Smote, and Table VI for the results of the Random Forest method with Smote-Tomeklink. The results of the comparison of the performance of the Random Forest method as a whole are shown in Figure 5.

TABLE IV  
RESULT CONFUSION MATRIX OF RANDOM FOREST

Actual	Predicted	
	Negative	Positive
Negative	429	71
Positive	113	155

TABLE V  
RESULT CONFUSION MATRIX OF RANDOM FOREST AND SMOTE

Actual	Predicted	
	Negative	Positive
Negative	390	110
Positive	71	429

TABLE VI  
RESULT CONFUSION MATRIX OF RANDOM FOREST AND SMOTE-TOMEKLINK

Actual	Predicted	
	Negative	Positive
Negative	385	90
Positive	56	419

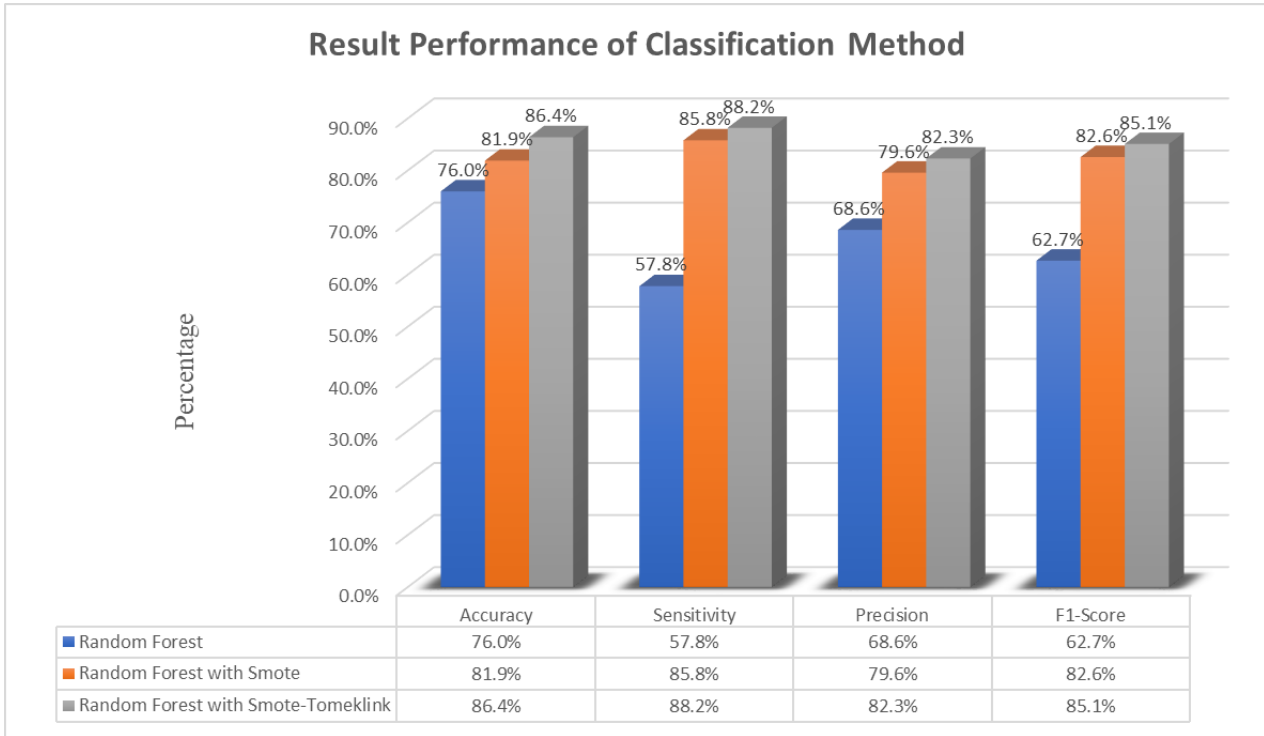


Fig 5. Result Performance of Classification Method

In Table IV, the Random Forest method succeeded in correctly classifying the negative class (TN) as many as 429 instances and the negative class classified incorrectly (FP) as many as 17 instances. While the correctly classified positive class (TP) is 155 instances and the incorrectly classified positive class is 113 instances.

In Table V, the Random Forest method with Smote succeeded in correctly classifying the negative class (TN) as many as 390 instances and the negative class classified incorrectly (FP) as many as 110 instances. While the positive class that is classified correctly (TP) is 429 instances and the positive class that is classified incorrectly is 71 instances.

In Table VI, the Random Forest method with Smote-Tomeklink succeeded in correctly classifying the negative class (TN) as many as 385 instances and the negative class classified incorrectly (FP) as 90 instances. While the positive class that is classified correctly (TP) is 419 instances and the positive class that is classified incorrectly is 56 instances.

Based on Figure 4, there was an increase in the performance of the Random Forest method with Smote-Tomeklink based on accuracy, sensitivity, precision, and F1-score. In the original

dataset, the Random Forest method has 76% accuracy, 57.8% sensitivity, 68.6% precision, and 62.7% F1-score. The Random Forest method with Smote has an accuracy of 81.9%, sensitivity of 85.8%, precision of 79.6%, and F1-score of 82.6%. Meanwhile, the use of the Random Forest method with Smote-Tomeklink resulted in an accuracy of 86.4%, a sensitivity of 88.2%, a precision of 83.3%, and F1-score of 85.1%.

Sensitivity has a very important role to improve the accuracy and F1-score performance of the Random Forest method with Smote-Tomeklink. The Random Forest method with Smote-Tomeklink gives higher accuracy, sensitivity, precision, and F1-score results than smote and without sampling.

Random Forest method with Smote an increase in performance indicators accuracy, sensitivity, precision, and F1-score. The increase in accuracy scores is 5.9%, Sensitivity is 28%, precision is 11%, and F1-score is 19.9%. The Random Forest method with Smote-Tomeklink showed an increase in the indicators of accuracy by 10.4%, Sensitivity by 30.4%, precision by 13.7%, and F1-score by 22.4%. Therefore, the use of the Smote-tomeklink method can increase accuracy,

sensitivity, precision, and F1-score in the Random Forest method [11][32][33]. The comparison of the proposed method is better than previous studies, which can be shown in Table VII.

TABLE VII  
COMPARISON OF THE PROPOSED MODEL PERFORMANCE WITH PREVIOUS STUDIES

No	Author (Year)	Dataset	Method	Accuracy
1	[14]	Pima Indian Diabetes	KNN	83%
2	[15]	Pima Indian Diabetes	Decision Tree C.45	75.65%
3	[11]	Pima Indian Diabetes	SVM + K-Means Smote	82%
4	[19]	Pima Indian Diabetes	Logistic Regression + Smote	82%
5	[20]	Pima Indian Diabetes	C4.5 Method + Smote	82%
6	The Proposed Method	Pima Indian Diabetes	Random Forest + Smote Tomek links	86%

#### IV. CONCLUSION

This study applies the Smote-Tomeklink algorithm to the Random Forest method for the classification of diabetes. The application of Smote-Tomeklink can improve the performance of accuracy, sensitivity, precision, and F1-score in the Random Forest method. The combination of Random Forest and Smote-Tomeklink got the best accuracy, sensitivity, and precision compared to Smote and without sampling for the classification of diabetes. Where, there was an increase in performance indicators of 10.4% accuracy, 30.4% sensitivity, 13.7% precision, and 22.4 F1-score. Further research can apply Smote-Tomeklink to deal with the problem of data imbalance in multiclass data.

#### REFERENCES

[1] O. Heranova, "Synthetic Minority Oversampling Technique pada Averaged One Dependence Estimators untuk Klasifikasi Credit Scoring," *Jurnal RESTI (Rekayasa Sistem dan Teknologi Informatika)*, vol. 3, no. 3, pp. 443–450, 2019, doi: 10.29207/resti.v3i3.1275.

[2] T. Zhu, Y. Lin, and Y. Liu, "Synthetic minority oversampling technique for multiclass imbalance problems," *Pattern Recognition*, vol. 72, pp. 327–340, Dec. 2017, doi: 10.1016/j.patcog.2017.07.024.

[3] F. Last, G. Douzas, and F. Bacao, "Oversampling for Imbalanced Learning Based on K-Means and SMOTE," no. November, 2017, [Online]. Available: <http://arxiv.org/abs/1711.00837>.

[4] G. A. Pradipta, R. Wardoyo, A. Musdholifah, and I. N. H. Sanjaya, "Radius-SMOTE: A New Oversampling Technique of Minority Samples Based on Radius Distance for Learning from Imbalanced Data," *IEEE Access*, vol. 9, pp. 74763–74777, 2021, doi: 10.1109/ACCESS.2021.3080316.

[5] M. Kamaladevi, V. Venkataraman, and K. R. Sekar, "Tomek link Undersampling with Stacked Ensemble classifier for Imbalanced data classification," vol. 25, no. 4, pp. 2182–2190, 2021.

[6] W. C. Lin, C. F. Tsai, Y. H. Hu, and J. S. Jhang, "Clustering-based undersampling in class-imbalanced data," *Information Sciences*, vol. 409–410, pp. 17–26, 2017, doi: 10.1016/j.ins.2017.05.008.

[7] Z. Xu, D. Shen, T. Nie, and Y. Kou, "A hybrid sampling algorithm combining M-SMOTE and ENN based on Random Forest for medical imbalanced data," *Journal of Biomedical Informatics*, p. 103465, 2020, doi: 10.1016/j.jbi.2020.103465.

[8] E. AT, A. M, A.-M. F, and S. M, "Classification of Imbalance Data

using Tomek Link (T-Link) Combined with Random Under-sampling (RUS) as a Data Reduction Method," *Global Journal of Technology and Optimization*, vol. 01, no. S1, 2016, doi: 10.4172/2229-8711.s1111.

[9] Z. Wang, C. Wu, K. Zheng, X. Niu, and X. Wang, "SMOTETomek-Based Resampling for Personality Recognition," *IEEE Access*, vol. 7, pp. 129678–129689, 2019, doi: 10.1109/ACCESS.2019.2940061.

[10] N. V. Chawla, K. W. Bowyer, and L. O. Hall, "SMOTE : Synthetic Minority Over-sampling TEchnique," *Journal of Artificial Intelligence Research*, vol. 16, pp. 341–378, 2002.

[11] H. Hairani, K. E. Saputro, and S. Fadli, "K-means-SMOTE for handling class imbalance in the classification of diabetes with C4.5, SVM, and naive Bayes," *Jurnal Teknologi dan Sistem Komputer*, vol. 8, no. 2, pp. 89–93, 2020, doi: 10.14710/jtsiskom.8.2.2020.89-93.

[12] I. Tomek, "Tomek Link: Two Modifications of CNN," *IEEE Trans. Systems, Man and Cybernetics*, pp. 769–772, 1976, [Online]. Available: <https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&number=4309452>.

[13] E. F. Swana, W. Doorsamy, and P. Bokoro, "Tomek Link and SMOTE Approaches for Machine Fault Classification with an Imbalanced Dataset," *Sensors*, vol. 22, no. 9, 2022, doi: 10.3390/s22093246.

[14] R. Kaur, "Predicting diabetes by adopting classification approach in data mining," *International Journal on Informatics Visualization*, vol. 3, no. 2–2, pp. 218–221, 2019, doi: 10.30630/joiv.3.2-2.229.

[15] A. Azrar, M. Awais, Y. Ali, and K. Zabeer, "Data mining models comparison for diabetes prediction," *International Journal of Advanced Computer Science and Applications*, vol. 9, no. 8, pp. 320–323, 2018, doi: 10.14569/ijacsa.2018.090841.

[16] S. Barik, S. Mohanty, S. Mohanty, and D. Singh, "Analysis of prediction accuracy of diabetes using classifier and hybrid machine learning techniques," *Smart Innovation, Systems and Technologies*, vol. 153, no. January, pp. 399–409, 2021, doi: 10.1007/978-981-15-6202-0\_41.

[17] H. Hairani, M. Innuddin, and M. Rahardi, "Accuracy Enhancement of Correlated Naive Bayes Method by Using Correlation Feature Selection (CFS) for Health Data Classification," in *2020 3rd International Conference on Information and Communications Technology (ICOIACT)*, 2020, pp. 51–55, doi: 10.1109/ICOIACT50329.2020.9332021.

[18] C. Fiarni, E. M. Sipayung, and S. Maemunah, "Analysis and prediction of diabetes complication disease using data mining algorithm," *Procedia Computer Science*, vol. 161, pp. 449–457, 2019, doi: 10.1016/j.procs.2019.11.144.

[19] Erlin, Y. N. Marlim, Junadhi, L. Suryati, and N. Agustina, "Early Detection of Diabetes Using Machine Learning with Logistic Regression Algorithm," *Jurnal Nasional Teknik Elektro dan Teknologi Informasi*, vol. 11, no. 2, pp. 88–96, 2022.

[20] C. Azad, B. Bhushan, R. Sharma, A. Shankar, K. K. Singh, and A. Khamparia, "Prediction model using SMOTE, genetic algorithm and decision tree (PMSGD) for classification of diabetes mellitus," *Multimedia Systems*, vol. 28, no. 4, pp. 1289–1307, 2022, doi: 10.1007/s00530-021-00817-2.

[21] X. Shi, T. Qu, G. Van Pottelbergh, M. van den Akker, and B. De Moor, "A Resampling Method to Improve the Prognostic Model of End-Stage Kidney Disease: A Better Strategy for Imbalanced Data," *Frontiers in Medicine*, vol. 9, no. March, pp. 1–9, 2022, doi: 10.3389/fmed.2022.730748.

[22] K. Wang *et al.*, "Improving risk identification of adverse outcomes in chronic heart failure using smote+enn and machine learning," *Risk Management and Healthcare Policy*, vol. 14, no. May, pp. 2453–2463, 2021, doi: 10.2147/RMHP.S310295.

[23] H. Kaur, H. S. Pannu, and A. K. Malhi, "A systematic review on imbalanced data challenges in machine learning: Applications and solutions," *ACM Computing Surveys*, vol. 52, no. 4. Association for Computing Machinery, pp. 1–34, Aug. 01, 2019, doi: 10.1145/3343440.

[24] K. Guo, X. Wan, L. Liu, Z. Gao, and M. Yang, "Fault diagnosis of intelligent production line based on digital twin and improved random forest," *Applied Sciences (Switzerland)*, vol. 11, no. 16, pp. 1–18, 2021, doi: 10.3390/app11167733.

[25] J. Chen, H. Huang, A. G. Cohn, D. Zhang, and M. Zhou, "Machine learning-based classification of rock discontinuity trace: SMOTE oversampling integrated with GBT ensemble learning," *International Journal of Mining Science and Technology*, vol. 32, no. 2, pp. 309–322, 2021, doi: 10.1016/j.ijmst.2021.08.004.

- [26] Y. Sun, H. Zhang, T. Zhao, Z. Zou, B. Shen, and L. Yang, "A New Convolutional Neural Network with Random Forest Method for Hydrogen Sensor Fault Diagnosis," *IEEE Access*, vol. 8, pp. 85421–85430, 2020, doi: 10.1109/ACCESS.2020.2992231.
- [27] H. Hartono and E. Ongko, "Avoiding Overfitting dan Overlapping in Handling Class Imbalanced Using Hybrid Approach with Smoothed Bootstrap Resampling and Feature Selection," *International Journal on Informatics Visualization*, vol. 6, no. June, pp. 343–348, 2022.
- [28] H. Hairani, A. Anggrawan, A. I. Wathan, K. A. Latif, K. Marzuki, and M. Zulfikri, "The Abstract of Thesis Classifier by Using Naive Bayes Method," in *Proceedings - 2021 International Conference on Software Engineering and Computer Systems and 4th International Conference on Computational Science and Information Management, ICSECS-ICOCSIM 2021*, 2021, no. August, pp. 312–315, doi: 10.1109/ICSECS52883.2021.00063.
- [29] A. Luque, A. Carrasco, A. Martín, and A. de las Heras, "The impact of class imbalance in classification performance metrics based on the binary confusion matrix," *Pattern Recognition*, vol. 91, pp. 216–231, 2019, doi: 10.1016/j.patcog.2019.02.023.
- [30] H. Qteat and M. Awad, "Using Hybrid Model of Particle Swarm Optimization and Multi-Layer Perceptron Neural Networks for Classification of Diabetes," *International Journal of Intelligent Engineering and Systems*, vol. 14, no. 3, pp. 11–22, 2021, doi: 10.22266/ijies2021.0630.02.
- [31] H. Hanafi, A. H. Muhammad, I. Verawati, and R. Hardi, "An Intrusion Detection System Using SDAE to Enhance Dimensional Reduction in Machine Learning," *International Journal on Informatics Visualization*, vol. 6, no. June, pp. 306–316, 2022.
- [32] H. Hairani, A. S. Suweleh, and D. Susilowaty, "Penanganan Ketidak Seimbangan Kelas Menggunakan Pendekatan Level Data," *MATRIK : Jurnal Manajemen, Teknik Informatika dan Rekayasa Komputer*, vol. 20, no. 1, pp. 109–116, 2020, doi: 10.30812/matrik.v20i1.846.
- [33] M. Y. Thanoun, M. T. Yaseen, and A. M. Aleesa, "Development of Intelligent Parkinson Disease Detection System Based on Machine Learning Techniques Using Speech Signal," *International Journal on Advanced Science, Engineering and Information Technology*, vol. 11, no. 1, pp. 388–392, 2021.

# Improvement Performance of the Random Forest Method on Unbalanced Diabetes Data Classification Using Smote-Tomek Link

Hairani Hairani<sup>a\*</sup>, Anthony Anggrawan<sup>b</sup>, Dadang Priyanto<sup>c</sup>

<sup>abc</sup> Department of Computer Science, Universitas Bumigora, Mataram, 83127, Indonesia

Corresponding author: Hairani@universitasbumigora.ac.id

**Abstract**— Most of the health data contained unbalanced data that affected the performance of the classification method. Unbalanced data causes the classification method to more easily classify the majority data and ignore the minority class. One of the health data that has unbalanced data is Pima Indian Diabetes. Diabetes is a deadly disease caused by the body's inability to produce enough insulin. Complications of diabetes can cause heart attacks and strokes. Early diagnosis of diabetes is needed to minimize the occurrence of more severe complications. In the diabetes dataset used, there is an imbalance of data between positive and negative diabetes classes. Diabetes negative class data (500 data) is more than diabetes positive class (268) so that it can affect the performance of the classification method. Therefore, this study aims to apply the Smote-TomekLink and Random Forest methods in the classification of diabetes. The research methodology used is the collection of diabetes data obtained from Kaggle as many as 768 data with 8 input attributes and 1 output attribute as a class, pre-processing data is used to balance the dataset with Smote-TomekLink, classification using the random forest method, and performance evaluation based on accuracy, sensitivity, precision, and F1-score. Based on the tests carried out by dividing data using 10-fold cross-validation, the Random forest algorithm with Smote-TomekLink gets the highest accuracy, sensitivity, precision, and F1-score compared to Random Forest with Smote. The Random Forest algorithm with Smote-TomekLink has 86.4% accuracy, 88.2% sensitivity, 82.3% precision, and 85.1% F1-score. Thus, using Smote-TomekLink can improve the performance of the random forest method based on accuracy, sensitivity, precision, and F1-score.

**Keywords**— Class Imbalance; Smote-TomekLink; Random Forest Method; Diabetest Disease.

*Manuscript received dd mm yyyy; revised dd mm yyyy; accepted dd mm yyyy. Date of publication dd mm yyyy.  
International Journal on Informatics Visualization is licensed under a Creative Commons Attribution-Share Alike 4.0 International License.*



## I. INTRODUCTION

Most of the Health data contained unbalanced data that affected the performance of the classification method. Unbalanced data causes the classification method to more easily classify the majority data and ignore the minority class. One of the health data that has unbalanced data is Pima Indian Diabetes. Diabetes is a deadly disease caused by increased blood sugar in the body. Diabetes is caused by the body's inability to produce enough insulin. Complications of diabetes can cause heart attacks and strokes. One way to improve the performance of the classification method is to handle balanced data by adding minority data so that the number is equal to the majority class. The diabetes dataset has 768 instances of data. However, the problem is that there is an imbalance of data in the dataset, namely the negative diabetes class with 500 data (majority class), while the positive diabetes class with 268 data (minority class). Data imbalance is the amount of data in one class more than in the other class. The problem of data imbalance causes the classification method to be more

dominant in classifying the majority class than the minority class, or in other words, the classification method ignores the minority class. The problem of unbalanced data can be handled with a data sampling approach.

Several data sampling methods that can be used to solve the problem of data imbalance are oversampling [1][2], [3][4], undersampling [5][6], and Hybrid Sampling[6],[7]. Oversampling works by adding the minority class, while Undersampling works by removing the majority class so as to produce balanced data. However, both methods have their respective weaknesses. The weakness of the oversampling method is that there are too many repetitions of samples that can cause overfitting of the classification method, while the weakness of undersampling is that it will lose information from most of the samples in the dataset and cannot take full advantage of the available information[9].

To avoid overfitting the oversampling method, the Smote method was developed to overcome these weaknesses. Smote is an oversampling method to generate new synthesis training data by linear interpolation on minority classes[10]. However,

the Smote method has a weakness, namely overgeneralization, and the addition of a minority class randomly can generate noise data, because it does not differentiate between classes[11]. Therefore, the undersampling method is used to improve the performance of the oversampling method by cleaning the noise data in the majority class. The noise data is the majority class instance which is closest to the minority class instance. Usually, noise data reduces the level of accuracy for classification methods[5]. One method to remove noise data in the majority class is Tomeklink[12]. Tomeklink is an undersampling method that cleans noise data from the majority class which has similar characteristics and overlapping. However, Tomeklink only deletes instances defined as “Tomek Links” so that the analyzed data cannot be balanced and in its implementation the method is combined with other methods. Combining Tomeklink and Smote oversampling can improve accuracy better than individual performance[13].

Data mining research in Health plays an important role, especially in predicting various types of diseases using different techniques or methods[14]. Research [15] uses a statistical approach to analyze the success rate of students following subjects using online or face-to-face learning. The results show that online students have significantly higher average grades than face-to-face classes.

Several previous studies have focused on the classification of diabetes, namely Research [16] predicts diabetes using the k-NN method with an accuracy of 83%. The weakness of the research is that it does not address the problem of data imbalance. Research [17] classifying diabetes using the C4.5 method with an accuracy of 75.65%. The weakness of the research is that the accuracy obtained is low so that it can be improved, and also does not address the problem of data imbalance. Research [18] Using XGBoost to predict diabetes with 74% accuracy. The weakness of the research is that the accuracy obtained is low so that it can be improved, and also does not address the problem of data imbalance.

Research [19] using the Correlated Naïve Bayes method with correlation-based feature selection to predict diabetes with an accuracy of 69.51%. The weakness of the research is that the accuracy obtained is low so that it can be improved, and also does not address the problem of data imbalance. Research [20] using the C4.5 method for diabetes detection with an accuracy of 68%.

Research [21] used logistic regression and smote methods to detect diabetes with 82% accuracy, 81% precision, 79% recall, and 80% F1-score. The weakness of the research is that the accuracy is good but can be improved using Tomeklink to clean noise data in the majority class. Research [22] using the C4.5 and Smote methods to predict diabetes with 82% accuracy, 80% precision, and 86% sensitivity. Research [23] used logistic and Smote-ENN methods to predict kidney disease with 75.2% accuracy, 70.6% recall, 4.9% precision, and 30% F1-score. The weakness of the research is the low accuracy so that it can be improved using Tomeklink to clean noise data in the majority class. Research [24] SME-XGBoost with Smote-ENN for heart disease prediction with 80% AUC.

Several previous studies have applied various approaches to improve diabetes classification methods such as the oversampling approach with SMOTE. However, there are weaknesses in previous studies, namely the accuracy of the proposed method still ranges from 82% to 83% so that there is

a gap to improve its accuracy. So, this study proposes the Smote-Tomeklink hybrid sampling method to overcome the imbalance in diabetes data, so as to improve the accuracy of the classification method.

Smote-Tomeklink is a good way to avoid the drawbacks of SMOTE and Tomeklink techniques [9]. The classification method used in this research is Random Forest. The Random Forest method was chosen because it has several advantages, namely high accuracy [25], the ability to handle noise data, fast performance in training data, overfitting control, and easy to implement [26].

This study aims to apply the Smote-Tomeklink hybrid sampling method to balance the data on diabetes data so as to improve the performance of the Random forest classification method. Measurement of the performance of the random forest method based on accuracy, sensitivity (recall), precision, and F1-score

## II. MATERIALS AND METHOD

This research consists of several stages as shown in Figure 1.

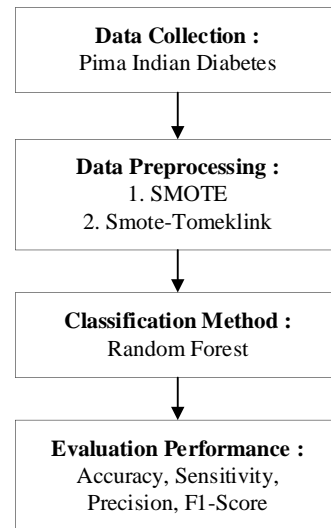


Fig 1. Research Stages

### A. Data Collection

The dataset used in this study is a diabetes dataset obtained from Kaggle, which consists of 768 instances and 9 attributes. The description of the attributes and the sample data used are shown respectively in Table I and Table II.

TABLE I  
DESCRIPTION ATRIBUT DATASET

No	Atribute	Description	Label
1	Pregnancies	Number of Pregnancy	X1
2	Glucose	Glucose level 2 hours after eating	X2
3	Blood Pressure	Blood Pressure	X3
4	Skin Thickness	Skin Thickness	X4
5	Insulin	Insulin	X5
6	BMI	Body Massa Index	X6
7	Diabetes Pedigree Function	Diabetes Pedigree Function	X7
8	Age	Age	X8

9	Outcome	Diabetes Status ( 1 = Positive Diabetes, 2 = Negative Diabetes	Y	767	1	126	60	0	0	30.1	0.349	47	1
				768	1	93	70	31	0	30.4	0.315	23	0

TABLE III  
SAMPLE DATASET

No	X1	X2	X3	X4	X5	X6	X7	X8	Y
1	6	148	72	35	0	33.6	0.627	50	1
2	1	85	66	29	0	26.6	0.351	31	0
3	8	183	64	0	0	23.3	0.672	32	1
4	1	89	66	23	94	28.1	0.167	21	0
5	0	137	40	35	168	43.1	2.288	33	1
6	5	116	74	0	0	25.6	0.201	30	1
7	3	78	50	32	88	31	0.248	26	1
8	10	115	0	0	0	35.3	0.134	29	0
9	2	197	70	45	543	30.5	0.158	53	1
10	8	125	96	0	0	0	0.232	54	1
11	4	110	92	0	0	37.6	0.191	30	0
12	10	168	74	0	0	38	0.537	34	1
13	10	139	80	0	0	27.1	1.441	57	0
14	1	189	60	23	846	30.1	0.398	59	1
15	5	166	72	19	175	25.8	0.587	51	1
..	..	...	..	..	...	..	.....	..	..
754	0	181	88	44	510	43.3	0.222	26	1
755	8	154	78	32	0	32.4	0.443	45	1
756	1	128	88	39	110	36.5	1.057	37	1
757	7	137	90	41	0	32	0.391	39	0
758	0	123	72	0	0	36.3	0.258	52	1
759	1	106	76	0	0	37.5	0.197	26	0
760	6	190	92	0	0	35.5	0.278	66	1
761	2	88	58	26	16	28.4	0.766	22	0
762	9	170	74	31	0	44	0.403	43	1
763	9	89	62	0	0	22.5	0.142	33	0
764	10	101	76	48	180	32.9	0.171	63	0
765	2	122	70	27	0	36.8	0.34	27	0
766	5	121	72	23	112	26.2	0.245	30	0

### B. Data Pre-processing

Data Pre-processing is one of the important stages in data mining to improve the quality of datasets. This study focuses on dealing with unbalanced data contained in the diabetes dataset. The dataset used has 268 instances of negative diabetes and 500 instances of Positive Diabetes. The algorithms used to handle unbalanced data in the dataset are SMOTE (Synthetic Minority Oversampling Technique) and Smote-Tomeklink.

SMOTE is one of the most commonly used oversampling methods to solve the problem of data distribution imbalance in machine learning modeling. SMOTE aims to balance the distribution of classes by increasing the number of minority classes randomly by creating synthetic data for oversampling purposes [10]. Creating new data on the minority class using the equation (1).

$$Y' = Y^i + (Y^j - Y^i) * \gamma \quad (1)$$

$Y'$  is the representation of the addition of the minority class.  $Y^i$  is the representation of minority class,  $Y^j$  is a value chosen at random from the k-nearest neighbors of the minority class on  $Y^i$ , and  $\gamma$  is a value in a randomly selected vector with a range of 0 to 1 [2].

SMOTE generates new synthesis training data by linear interpolation for the minority class. Synthesis training data is generated by randomly selecting one or more of the k-nearest neighbors for each sample in the minority class as shown in Figure 2.

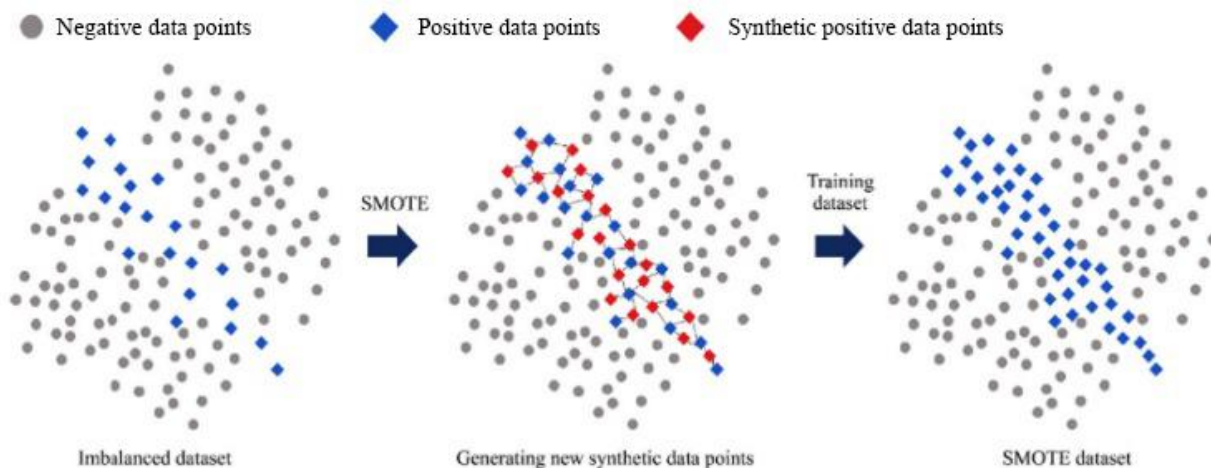


Fig 2. Process of Synthetic Minority Oversampling Technique (SMOTE) Algorithm [27]

Tomeklink is an undersampling method that cleans noise data from the majority class that has similar characteristics and overlapping[12]. Tomeklink works by eliminating the majority class instances that are closer to the minority class by applying the nearest neighbor rule to select instances. The combination of Tomeklink and Smote oversampling can improve accuracy better than individual performance [13].

### C. Random Forest Method

Random Forest is a decision tree-based ensemble learning method [28]. The Random Forest method has the advantages of high accuracy, the ability to handle noise data, fast performance in training data, overfitting control, and easy to implement [26]. The working process of the Random Forest method in classifying a data is shown in Figure 3.



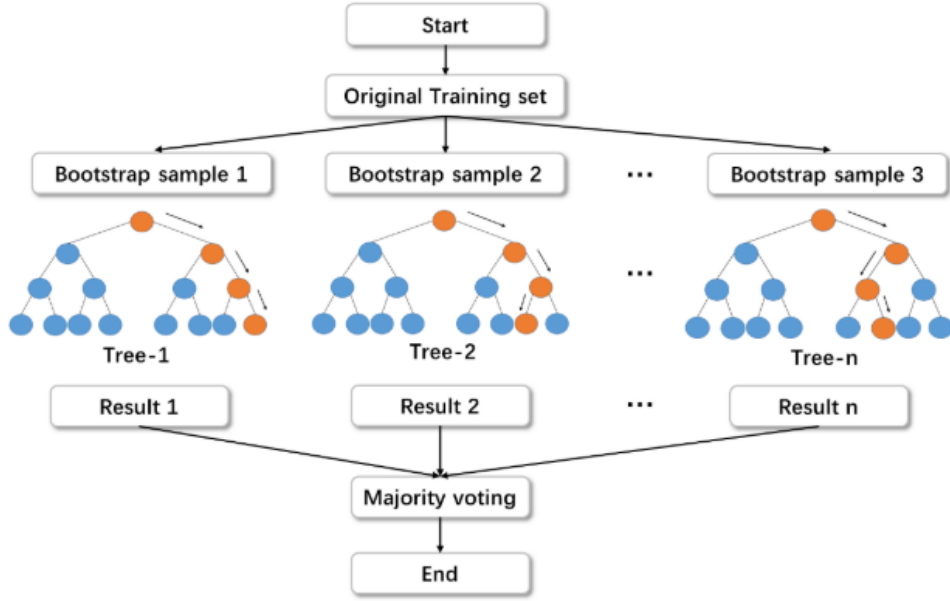


Fig 1. Working Process of Random Forest Method[26]

Figure 3 shows how the Random Forest algorithm works by creating a set of decision trees from a randomly selected subset, getting predictions from each decision tree, voting for each predicted outcome, and choosing the best prediction result based on the most votes assigned as final prediction

#### D. Evaluation Performance

Performance testing uses a confusion matrix table. The confusion matrix is a table that is used to describe the performance of the classification method on a dataset whose true value is known. The confusion matrix can visualize the amount of data that is classified as true and false as shown in the Table III[29].

TABLE III  
CONFUSION MATRIX

Actual	Predicted	
	Negative	Positive
Negative	TN	FP
Positive	FN	TP

Formula used to calculate *Accuracy* (6), *Sensitivity* (7), *Precision* (8) [30] [31][32], and *F1-score* (5)[33].

$$Accuracy = \frac{TP + TN}{TP + FN + TN + FP} \quad (6)$$

$$Sensitivity = \frac{TP}{TP + FN} \quad (7)$$

$$Precision = \frac{TP}{TP + FP} \quad (8)$$

$$F1-score = \frac{2 \times precision \times recall}{precision + recall} \quad (9)$$

True Positive (TP) is a class of positive diabetes that is predicted correctly. False Positive (FP) is a diabetes negative

class but is predicted to be diabetes positive. True Negative (TN) is a diabetes negative class that is predicted correctly. False Negative (FN) is a positive diabetes class but is predicted to be diabetes negative.

### III. RESULT AND DISCUSSION

This research starts from the stages of data collection, data pre-processing, classification, and performance testing. The data used in this study is diabetes data obtained from Kaggle. The pre-processing of this study used the Smote and Smote-Tomeklink algorithms to deal with class imbalances in diabetes data. The classification method of this research is Random Forest. The performance test is based on accuracy, sensitivity, precision, and F1-score. The results of the comparison of the original data with the data from Smote and the results of Smote-Tomeklink are shown in Figure 4.

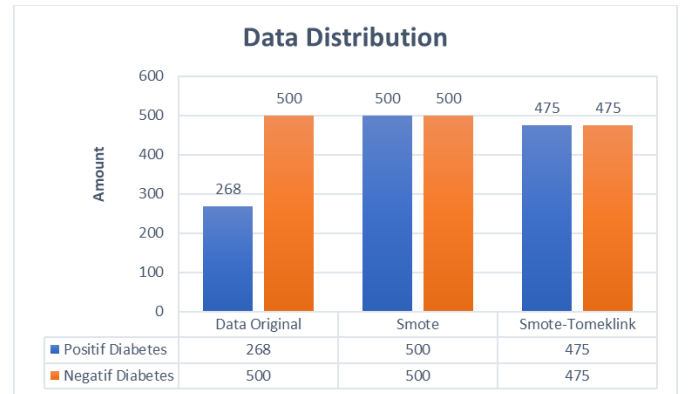


Fig 4. Data Distribution Result

The classification method of this research is Random Forest. Performance testing is based on accuracy, sensitivity, precision, and F1-score using a confusion matrix table. Based on testing the Random Forest method using 10-fold cross-validation, the results obtained in the form of a confusion matrix table as shown in Table IV for the Random Forest method on the original data, Table V for the results of the Random Forest

method with Smote, and Table VI for the results of the Random Forest method with Smote-Tomeklink. The results of the comparison of the performance of the Random Forest method as a whole are shown in Figure 5.

TABLE IV  
RESULT CONFUSION MATRIX OF RANDOM FOREST

Actual	Predicted	
	Negative	Positive
Negative	429	71
Positive	113	155

TABLE V  
RESULT CONFUSION MATRIX OF RANDOM FOREST AND SMOTE

Actual	Predicted	
	Negative	Positive
Negative	390	110
Positive	71	429

TABLE VI  
RESULT CONFUSION MATRIX OF RANDOM FOREST AND SMOTE-TOMEKLINK

Actual	Predicted	
	Negative	Positive
Negative	385	90
Positive	56	419

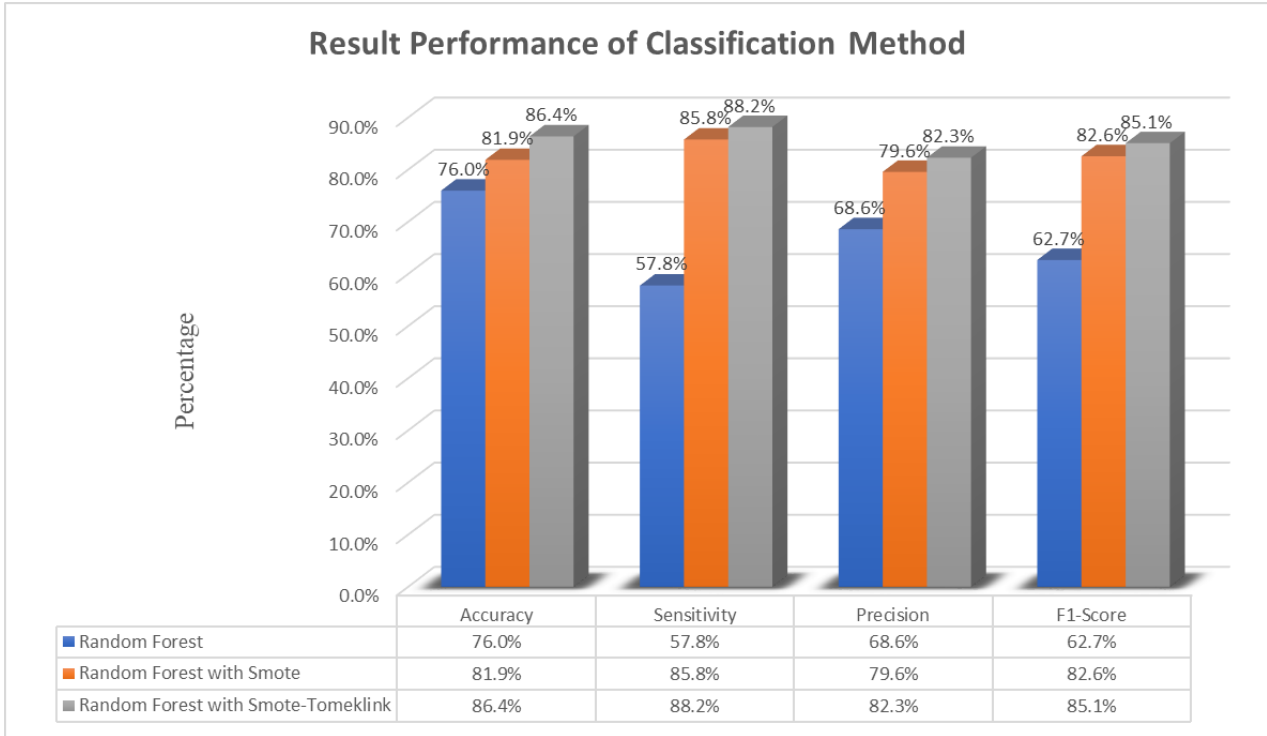


Fig 5. Result Performance of Classification Method

In Table IV, the Random Forest method succeeded in correctly classifying the negative class (TN) as many as 429 instances and the negative class classified incorrectly (FP) as many as 17 instances. While the correctly classified positive class (TP) is 155 instances and the incorrectly classified positive class is 113 instances.

In Table V, the Random Forest method with Smote succeeded in correctly classifying the negative class (TN) as many as 390 instances and the negative class classified incorrectly (FP) as many as 110 instances. While the positive class that is classified correctly (TP) is 429 instances and the positive class that is classified incorrectly is 71 instances.

In Table VI, the Random Forest method with Smote-Tomeklink succeeded in correctly classifying the negative class (TN) as many as 385 instances and the negative class classified incorrectly (FP) as 90 instances. While the positive class that is classified correctly (TP) is 419 instances and the positive class that is classified incorrectly is 56 instances.

Based on Figure 4, there was an increase in the performance of the Random Forest method with Smote-Tomeklink based on accuracy, sensitivity, precision, and F1-score. In the original dataset, the Random Forest method has 76% accuracy, 57.8%

sensitivity, 68.6% precision, and 62.7% F1-score. The Random Forest method with Smote has an accuracy of 81.9%, sensitivity of 85.8%, precision of 79.6%, and F1-score of 82.6%. Meanwhile, the use of the Random Forest method with Smote-Tomeklink resulted in an accuracy of 86.4%, a sensitivity of 88.2%, a precision of 83.3%, and F1-score of 85.1%.

Sensitivity has a very important role to improve the accuracy and F1-score performance of the Random Forest method with Smote-Tomeklink. The Random Forest method with Smote-Tomeklink gives higher accuracy, sensitivity, precision, and F1-score results than smote and without sampling.

Random Forest method with Smote an increase in performance indicators accuracy, sensitivity, precision, and F1-score. The increase in accuracy scores is 5.9%, Sensitivity is 28%, precision is 11%, and F1-score is 19.9%. The Random Forest method with Smote-Tomeklink showed an increase in the indicators of accuracy by 10.4%, Sensitivity by 30.4%, precision by 13.7%, and F1-score by 22.4%. Therefore, the use of the Smote-tomeklink method can increase accuracy, sensitivity, precision, and F1-score in the Random Forest method [11][34][35]. The comparison of the proposed method

is better than previous studies, which can be shown in Table VII.

TABLE VII  
COMPARISON OF THE PROPOSED MODEL PERFORMANCE WITH PREVIOUS STUDIES

No	Author (Year)	Dataset	Method	Accuracy
1	[16]	Pima Indian Diabetes	KNN	83%
2	[17]	Pima Indian Diabetes	Decision Tree C.45	75.65%
3	[11]	Pima Indian Diabetes	SVM + K-Means Smote	82%
4	[21]	Pima Indian Diabetes	Logistic Regression + Smote	82%
5	[22]	Pima Indian Diabetes	C4.5 Method + Smote	82%
6	<b>The Proposed Method</b>	<b>Pima Indian Diabetes</b>	<b>Random Forest + SMOTE Tomek links</b>	<b>86%</b>

#### IV. CONCLUSION

This study applies the Smote-Tomeklink algorithm to the Random Forest method for the classification of diabetes. The implementation of Smote-Tomeklink can improve the performance of accuracy, sensitivity, precision, and F1-score in the Random Forest method. The combination of Random Forest and Smote-Tomeklink got the best accuracy, sensitivity, and precision compared to Smote and without sampling for the classification of diabetes. Where, there was an increase in performance indicators of 10.4% accuracy, 30.4% sensitivity, 13.7% precision, and 22.4 F1-score. Further research can apply Smote-Tomeklink to deal with the problem of data imbalance in multiclass data.

#### REFERENCES

- [1] O. Heranova, "Synthetic Minority Oversampling Technique pada Averaged One Dependence Estimators untuk Klasifikasi Credit Scoring," *Jurnal RESTI (Rekayasa Sistem dan Teknologi Informasi)*, vol. 3, no. 3, pp. 443–450, 2019, doi: 10.29207/resti.v3i3.1275.
- [2] T. Zhu, Y. Lin, and Y. Liu, "Synthetic minority oversampling technique for multiclass imbalance problems," *Pattern Recognition*, vol. 72, pp. 327–340, Dec. 2017, doi: 10.1016/j.patcog.2017.07.024.
- [3] F. Last, G. Douzas, and F. Bacao, "Oversampling for Imbalanced Learning Based on K-Means and SMOTE," no. November, 2017, [Online]. Available: <http://arxiv.org/abs/1711.00837>.
- [4] G. A. Pradipta, R. Wardoyo, A. Musdholifah, and I. N. H. Sanjaya, "Radius-SMOTE: A New Oversampling Technique of Minority Samples Based on Radius Distance for Learning from Imbalanced Data," *IEEE Access*, vol. 9, pp. 74763–74777, 2021, doi: 10.1109/ACCESS.2021.3080316.
- [5] M. Kamaladevi, V. Venkataraman, and K. R. Sekar, "Tomek link Undersampling with Stacked Ensemble classifier for Imbalanced data classification," vol. 25, no. 4, pp. 2182–2190, 2021.
- [6] W. C. Lin, C. F. Tsai, Y. H. Hu, and J. S. Jhang, "Clustering-based undersampling in class-imbalanced data," *Information Sciences*, vol. 409–410, pp. 17–26, 2017, doi: 10.1016/j.ins.2017.05.008.
- [7] Z. Xu, D. Shen, T. Nie, and Y. Kou, "A hybrid sampling algorithm combining M-SMOTE and ENN based on Random Forest for medical imbalanced data," *Journal of Biomedical Informatics*, p. 103465, 2020, doi: 10.1016/j.jbi.2020.103465.
- [8] E. AT, A. M. A.-M. F., and S. M., "Classification of Imbalance Data using Tomek Link (T-Link) Combined with Random Undersampling (RUS) as a Data Reduction Method," *Global Journal of Technology and Optimization*, vol. 01, no. S1, 2016, doi: 10.4172/2229-8711.s1111.
- [9] Z. Wang, C. Wu, K. Zheng, X. Niu, and X. Wang, "SMOTETomek-Based Resampling for Personality Recognition," *IEEE Access*, vol. 7, pp. 129678–129689, 2019, doi: 10.1109/ACCESS.2019.2940061.
- [10] N. V. Chawla, K. W. Bowyer, and L. O. Hall, "SMOTE: Synthetic Minority Over-sampling TEchnique," *Journal of Artificial Intelligence Research*, vol. 16, pp. 341–378, 2002.
- [11] H. Hairani, K. E. Saputro, and S. Fadli, "K-means-SMOTE for handling class imbalance in the classification of diabetes with C4.5, SVM, and naive Bayes," *Jurnal Teknologi dan Sistem Komputer*, vol. 8, no. 2, pp. 89–93, 2020, doi: 10.14710/jtsiskom.8.2.2020.89-93.
- [12] I. Tomek, "Tomek Link: Two Modifications of CNN," *IEEE Trans. Systems, Man and Cybernetics*, pp. 769–772, 1976, [Online]. Available: <https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&number=4309452>.
- [13] E. F. Swana, W. Doorsamy, and P. Bokoro, "Tomek Link and SMOTE Approaches for Machine Fault Classification with an Imbalanced Dataset," *Sensors*, vol. 22, no. 9, 2022, doi: 10.3390/s22093246.
- [14] A. Alzahrani and A. Safhi, "The role of data mining techniques and tools in big data management in healthcare field," *Sustainable Engineering and Innovation*, vol. 4, no. 1, pp. 58–65, 2022, doi: 10.37868/sei.v4i1.id128.
- [15] S. Sarač and B. Duraković, "Analysis of student performances in online and face-to-face learning: A case study from a Bosnian public university," *Heritage and Sustainable Development*, vol. 4, no. 2, pp. 87–94, 2022, doi: 10.37868/HSD.V4I2.91.
- [16] R. Kaur, "Predicting diabetes by adopting classification approach in data mining," *International Journal on Informatics Visualization*, vol. 3, no. 2–2, pp. 218–221, 2019, doi: 10.30630/ijov.3.2-2.229.
- [17] A. Azrar, M. Awais, Y. Ali, and K. Zaheer, "Data mining models comparison for diabetes prediction," *International Journal of Advanced Computer Science and Applications*, vol. 9, no. 8, pp. 320–323, 2018, doi: 10.14569/ijacsa.2018.090841.
- [18] S. Barik, S. Mohanty, S. Mohanty, and D. Singh, "Analysis of prediction accuracy of diabetes using classifier and hybrid machine learning techniques," *Smart Innovation, Systems and Technologies*, vol. 153, no. January, pp. 399–409, 2021, doi: 10.1007/978-981-15-6202-0\_41.
- [19] H. Hairani, M. Innuddin, and M. Rahardi, "Accuracy Enhancement of Correlated Naive Bayes Method by Using Correlation Feature Selection (CFS) for Health Data Classification," in *2020 3rd International Conference on Information and Communications Technology (ICOIACT)*, 2020, pp. 51–55, doi: 10.1109/ICOIACT50329.2020.9332021.
- [20] C. Fiarni, E. M. Sipayung, and S. Maemunah, "Analysis and prediction of diabetes complication disease using data mining algorithm," *Procedia Computer Science*, vol. 161, pp. 449–457, 2019, doi: 10.1016/j.procs.2019.11.144.
- [21] Erlin, Y. N. Marlim, Junadhi, L. Suryati, and N. Agustina, "Early Detection of Diabetes Using Machine Learning with Logistic Regression Algorithm," *Jurnal Nasional Teknik Elektro dan Teknologi Informasi*, vol. 11, no. 2, pp. 88–96, 2022.
- [22] C. Azad, B. Bhushan, R. Sharma, A. Shankar, K. K. Singh, and A. Khamparia, "Prediction model using SMOTE, genetic algorithm and decision tree (PMSGD) for classification of diabetes mellitus," *Multimedia Systems*, vol. 28, no. 4, pp. 1289–1307, 2022, doi: 10.1007/s00530-021-00817-2.
- [23] X. Shi, T. Qu, G. Van Pottelbergh, M. van den Akker, and B. De Moor, "A Resampling Method to Improve the Prognostic Model of End-Stage Kidney Disease: A Better Strategy for Imbalanced Data," *Frontiers in Medicine*, vol. 9, no. March, pp. 1–9, 2022, doi: 10.3389/fmed.2022.730748.
- [24] K. Wang *et al.*, "Improving risk identification of adverse outcomes in chronic heart failure using smote+enn and machine learning," *Risk Management and Healthcare Policy*, vol. 14, no. May, pp. 2453–2463, 2021, doi: 10.2147/RMHP.S310295.
- [25] H. Kaur, H. S. Pannu, and A. K. Malhi, "A systematic review on imbalanced data challenges in machine learning: Applications and solutions," *ACM Computing Surveys*, vol. 52, no. 4. Association for Computing Machinery, pp. 1–34, Aug. 01, 2019, doi: 10.1145/3343440.
- [26] K. Guo, X. Wan, L. Liu, Z. Gao, and M. Yang, "Fault diagnosis of intelligent production line based on digital twin and improved random forest," *Applied Sciences (Switzerland)*, vol. 11, no. 16, pp. 1–18, 2021, doi: 10.3390/app11167733.
- [27] J. Chen, H. Huang, A. G. Cohn, D. Zhang, and M. Zhou, "Machine

- learning-based classification of rock discontinuity trace: SMOTE oversampling integrated with GBT ensemble learning,” *International Journal of Mining Science and Technology*, vol. 32, no. 2, pp. 309–322, 2021, doi: 10.1016/j.ijmst.2021.08.004.
- [28] Y. Sun, H. Zhang, T. Zhao, Z. Zou, B. Shen, and L. Yang, “A New Convolutional Neural Network with Random Forest Method for Hydrogen Sensor Fault Diagnosis,” *IEEE Access*, vol. 8, pp. 85421–85430, 2020, doi: 10.1109/ACCESS.2020.2992231.
- [29] H. Hartono and E. Ongko, “Avoiding Overfitting dan Overlapping in Handling Class Imbalanced Using Hybrid Approach with Smoothed Bootstrap Resampling and Feature Selection,” *International Journal on Informatics Visualization*, vol. 6, no. June, pp. 343–348, 2022.
- [30] H. Hairani, A. Anggrawan, A. I. Wathan, K. A. Latif, K. Marzuki, and M. Zulfikri, “The Abstract of Thesis Classifier by Using Naive Bayes Method,” in *Proceedings - 2021 International Conference on Software Engineering and Computer Systems and 4th International Conference on Computational Science and Information Management, ICSECS-ICOCSIM 2021*, 2021, no. August, pp. 312–315, doi: 10.1109/ICSECS52883.2021.00063.
- [31] A. Luque, A. Carrasco, A. Martín, and A. de las Heras, “The impact of class imbalance in classification performance metrics based on the binary confusion matrix,” *Pattern Recognition*, vol. 91, pp. 216–231, 2019, doi: 10.1016/j.patcog.2019.02.023.
- [32] H. Qteat and M. Awad, “Using Hybrid Model of Particle Swarm Optimization and Multi-Layer Perceptron Neural Networks for Classification of Diabetes,” *International Journal of Intelligent Engineering and Systems*, vol. 14, no. 3, pp. 11–22, 2021, doi: 10.22266/ijies2021.0630.02.
- [33] H. Hanafi, A. H. Muhammad, I. Verawati, and R. Hardi, “An Intrusion Detection System Using SDAE to Enhance Dimensional Reduction in Machine Learning,” *International Journal on Informatics Visualization*, vol. 6, no. June, pp. 306–316, 2022.
- [34] H. Hairani, A. S. Suweleh, and D. Susilowaty, “Penanganan Ketidak Seimbangan Kelas Menggunakan Pendekatan Level Data,” *MATRIK : Jurnal Manajemen, Teknik Informatika dan Rekayasa Komputer*, vol. 20, no. 1, pp. 109–116, 2020, doi: 10.30812/matrik.v20i1.846.
- [35] M. Y. Thanoun, M. T. Yaseen, and A. M. Aleesa, “Development of Intelligent Parkinson Disease Detection System Based on Machine Learning Techniques Using Speech Signal,” *International Journal on Advanced Science, Engineering and Information Technology*, vol. 11, no. 1, pp. 388–392, 2021.

# Improvement Performance of the Random Forest Method on Unbalanced Diabetes Data Classification Using Smote-Tomek Link

Hairani Hairani<sup>a\*</sup>, Anthony Anggrawan<sup>b</sup>, Dadang Priyanto<sup>c</sup>

<sup>abc</sup> Department of Computer Science, Universitas Bumigora, Mataram, 83127, Indonesia

Corresponding author: Hairani@universitasbumigora.ac.id

---

**Abstract**— Most of the health data contained unbalanced data that affected the performance of the classification method. Unbalanced data causes the classification method to more easily classify the majority data and ignore the minority class. One of the health data that has unbalanced data is Pima Indian Diabetes. Diabetes is a deadly disease caused by the body's inability to produce enough insulin. Complications of diabetes can cause heart attacks and strokes. Early diagnosis of diabetes is needed to minimize the occurrence of more severe complications. In the diabetes dataset used, there is an imbalance of data between positive and negative diabetes classes. Diabetes negative class data (500 data) is more than diabetes positive class (268) so that it can affect the performance of the classification method. Therefore, this study aims to apply the Smote-Tomeklink and Random Forest methods in the classification of diabetes. The research methodology used is the collection of diabetes data obtained from Kaggle as many as 768 data with 8 input attributes and 1 output attribute as a class, pre-processing data is used to balance the dataset with Smote-Tomeklink, classification using the random forest method, and performance evaluation based on accuracy, sensitivity, precision, and F1-score. Based on the tests carried out by dividing data using 10-fold cross-validation, the Random forest algorithm with Smote-TomekLink gets the highest accuracy, sensitivity, precision, and F1-score compared to Random Forest with Smote. The Random Forest algorithm with Smote-Tomeklink has 86.4% accuracy, 88.2% sensitivity, 82.3% precision, and 85.1% F1-score. Thus, using Smote-Tomeklink can improve the performance of the random forest method based on accuracy, sensitivity, precision, and F1-score.

**Keywords**— Class Imbalance; Smote-Tomeklink; Random Forest Method; Diabetest Disease.

*Manuscript received dd mmm. yyyy; revised dd mmm. yyyy; accepted dd mmm. yyyy. Date of publication dd mmm. yyyy.*

*International Journal on Informatics Visualization is licensed under a Creative Commons Attribution-Share Alike 4.0 International License.*



## I. INTRODUCTION

Most of the Health data contained unbalanced data that affected the performance of the classification method. Unbalanced data causes the classification method to more easily classify the majority data and ignore the minority class. One of the health data that has unbalanced data is Pima Indian Diabetes. Diabetes is a deadly disease caused by increased blood sugar in the body. Diabetes is caused by the body's inability to produce enough insulin. Complications of diabetes can cause heart attacks and strokes. One way to improve the performance of the classification method is to handle balanced data by adding minority data so that the number is equal to the majority class. The diabetes dataset has 768 instances of data. However, the problem is that there is an imbalance of data in the dataset, namely the negative diabetes class with 500 data (majority class), while the positive diabetes class with 268 data (minority class). Data imbalance is the amount of data in one class more than in the other class. The problem of data imbalance causes the classification method to be more

dominant in classifying the majority class than the minority class, or in other words, the classification method ignores the minority class. The problem of unbalanced data can be handled with a data sampling approach.

Several data sampling methods that can be used to solve the problem of data imbalance are oversampling [1][2], [3][4], undersampling [5][6], and Hybrid Sampling[6],[7]. Oversampling works by adding the minority class, while Undersampling works by removing the majority class so as to produce balanced data. However, both methods have their respective weaknesses. The weakness of the oversampling method is that there are too many repetitions of samples that can cause overfitting of the classification method, while the weakness of undersampling is that it will lose information from most of the samples in the dataset and cannot take full advantage of the available information[9].

To avoid overfitting the oversampling method, the Smote method was developed to overcome these weaknesses. Smote is an oversampling method to generate new synthesis training data by linear interpolation on minority classes[10]. However,

the Smote method has a weakness, namely overgeneralization, and the addition of a minority class randomly can generate noise data, because it does not differentiate between classes[11]. Therefore, the undersampling method is used to improve the performance of the oversampling method by cleaning the noise data in the majority class. The noise data is the majority class instance which is closest to the minority class instance. Usually, noise data reduces the level of accuracy for classification methods[5]. One method to remove noise data in the majority class is Tomeklink[12]. Tomeklink is an undersampling method that cleans noise data from the majority class which has similar characteristics and overlapping. However, Tomeklink only deletes instances defined as “Tomek Links” so that the analyzed data cannot be balanced and in its implementation the method is combined with other methods. Combining Tomeklink and Smote oversampling can improve accuracy better than individual performance[13].

Data mining research in Health plays an important role, especially in predicting various types of diseases using different techniques or methods[14]. Research [15] uses a statistical approach to analyze the success rate of students following subjects using online or face-to-face learning. The results show that online students have significantly higher average grades than face-to-face classes.

Several previous studies have focused on the classification of diabetes, namely Research [16] predicts diabetes using the k-NN method with an accuracy of 83%. The weakness of the research is that it does not address the problem of data imbalance. Research [17] classifying diabetes using the C4.5 method with an accuracy of 75.65%. The weakness of the research is that the accuracy obtained is low so that it can be improved, and also does not address the problem of data imbalance. Research [18] Using XGBoost to predict diabetes with 74% accuracy. The weakness of the research is that the accuracy obtained is low so that it can be improved, and also does not address the problem of data imbalance.

Research [19] using the Correlated Naïve Bayes method with correlation-based feature selection to predict diabetes with an accuracy of 69.51%. The weakness of the research is that the accuracy obtained is low so that it can be improved, and also does not address the problem of data imbalance. Research [20] using the C4.5 method for diabetes detection with an accuracy of 68%.

Research [21] used logistic regression and smote methods to detect diabetes with 82% accuracy, 81% precision, 79% recall, and 80% F1-score. The weakness of the research is that the accuracy is good but can be improved using Tomeklink to clean noise data in the majority class. Research [22] using the C4.5 and Smote methods to predict diabetes with 82% accuracy, 80% precision, and 86% sensitivity. Research [23] used logistic and Smote-ENN methods to predict kidney disease with 75.2% accuracy, 70.6% recall, 4.9% precision, and 30% F1-score. The weakness of the research is the low accuracy so that it can be improved using Tomeklink to clean noise data in the majority class. Research [24] SME-XGBoost with Smote-ENN for heart disease prediction with 80% AUC.

Several previous studies have applied various approaches to improve diabetes classification methods such as the oversampling approach with SMOTE. However, there are weaknesses in previous studies, namely the accuracy of the proposed method still ranges from 82% to 83% so that there is

a gap to improve its accuracy. So, this study proposes the Smote-Tomeklink hybrid sampling method to overcome the imbalance in diabetes data, so as to improve the accuracy of the classification method.

Smote-Tomeklink is a good way to avoid the drawbacks of SMOTE and Tomeklink techniques [9]. The classification method used in this research is Random Forest. The Random Forest method was chosen because it has several advantages, namely high accuracy [25], the ability to handle noise data, fast performance in training data, overfitting control, and easy to implement [26].

This study aims to apply the Smote-Tomeklink hybrid sampling method to balance the data on diabetes data so as to improve the performance of the Random forest classification method. Measurement of the performance of the random forest method based on accuracy, sensitivity (recall), precision, and F1-score

## II. MATERIALS AND METHOD

This research consists of several stages as shown in Figure 1.

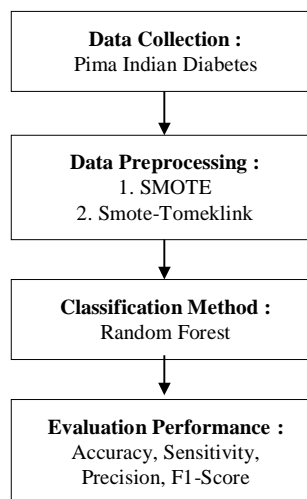


Fig 1. Research Stages

### A. Data Collection

The dataset used in this study is a diabetes dataset obtained from Kaggle, which consists of 768 instances and 9 attributes. The description of the attributes and the sample data used are shown respectively in Table I and Table II.

TABLE I  
DESCRIPTION ATRIBUT DATASET

No	Atribute	Description	Label
1	Pregnancies	Number of Pregnancy	X1
2	Glucose	Glucose level 2 hours after eating	X2
3	Blood Pressure	Blood Pressure	X3
4	Skin Thickness	Skin Thickness	X4
5	Insulin	Insulin	X5
6	BMI	Body Massa Index	X6
7	Diabetes Pedigree Function	Diabetes Pedigree Function	X7
8	Age	Age	X8

9	Outcome	Diabetes Status ( 1 = Positive Diabetes, 2 = Negative Diabetes	Y
---	---------	--	---

767	1	126	60	0	0	30.1	0.349	47	1
768	1	93	70	31	0	30.4	0.315	23	0

TABLE III  
SAMPLE DATASET

No	X1	X2	X3	X4	X5	X6	X7	X8	Y
1	6	148	72	35	0	33.6	0.627	50	1
2	1	85	66	29	0	26.6	0.351	31	0
3	8	183	64	0	0	23.3	0.672	32	1
4	1	89	66	23	94	28.1	0.167	21	0
5	0	137	40	35	168	43.1	2.288	33	1
6	5	116	74	0	0	25.6	0.201	30	1
7	3	78	50	32	88	31	0.248	26	1
8	10	115	0	0	0	35.3	0.134	29	0
9	2	197	70	45	543	30.5	0.158	53	1
10	8	125	96	0	0	0	0.232	54	1
11	4	110	92	0	0	37.6	0.191	30	0
12	10	168	74	0	0	38	0.537	34	1
13	10	139	80	0	0	27.1	1.441	57	0
14	1	189	60	23	846	30.1	0.398	59	1
15	5	166	72	19	175	25.8	0.587	51	1
..	..	..	..	..	..	..	..	..	..
754	0	181	88	44	510	43.3	0.222	26	1
755	8	154	78	32	0	32.4	0.443	45	1
756	1	128	88	39	110	36.5	1.057	37	1
757	7	137	90	41	0	32	0.391	39	0
758	0	123	72	0	0	36.3	0.258	52	1
759	1	106	76	0	0	37.5	0.197	26	0
760	6	190	92	0	0	35.5	0.278	66	1
761	2	88	58	26	16	28.4	0.766	22	0
762	9	170	74	31	0	44	0.403	43	1
763	9	89	62	0	0	22.5	0.142	33	0
764	10	101	76	48	180	32.9	0.171	63	0
765	2	122	70	27	0	36.8	0.34	27	0
766	5	121	72	23	112	26.2	0.245	30	0

### B. Data Pre-processing

Data Pre-processing is one of the important stages in data mining to improve the quality of datasets. This study focuses on dealing with unbalanced data contained in the diabetes dataset. The dataset used has 268 instances of negative diabetes and 500 instances of Positive Diabetes. The algorithms used to handle unbalanced data in the dataset are SMOTE (Synthetic Minority Oversampling Technique) and Smote-Tomeklink.

SMOTE is one of the most commonly used oversampling methods to solve the problem of data distribution imbalance in machine learning modeling. SMOTE aims to balance the distribution of classes by increasing the number of minority classes randomly by creating synthetic data for oversampling purposes [10]. Creating new data on the minority class using the equation (1).

$$Y' = Y^i + (Y^j - Y^i) * \gamma \quad (1)$$

$Y'$  is the representation of the addition of the minority class.  $Y^i$  is the representation of minority class,  $Y^j$  is a value chosen at random from the k-nearest neighbors of the minority class on  $Y^i$ , and  $\gamma$  is a value in a randomly selected vector with a range of 0 to 1 [2].

SMOTE generates new synthesis training data by linear interpolation for the minority class. Synthesis training data is generated by randomly selecting one or more of the k-nearest neighbors for each sample in the minority class as shown in Figure 2.

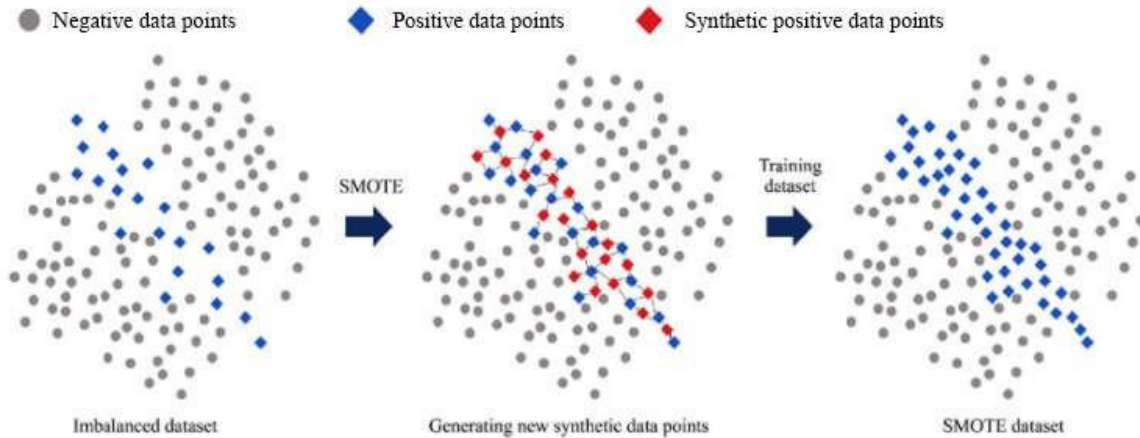


Fig 2. Process of Synthetic Minority Oversampling Technique (SMOTE) Algorithm [27]

Tomeklink is an undersampling method that cleans noise data from the majority class that has similar characteristics and overlapping [12]. Tomeklink works by eliminating the majority class instances that are closer to the minority class by applying the nearest neighbor rule to select instances. The combination of Tomeklink and Smote oversampling can improve accuracy better than individual performance [13].

### C. Random Forest Method

Random Forest is a decision tree-based ensemble learning method [28]. The Random Forest method has the advantages of high accuracy, the ability to handle noise data, fast performance in training data, overfitting control, and easy to implement [26]. The working process of the Random Forest method in classifying a data is shown in Figure 3.

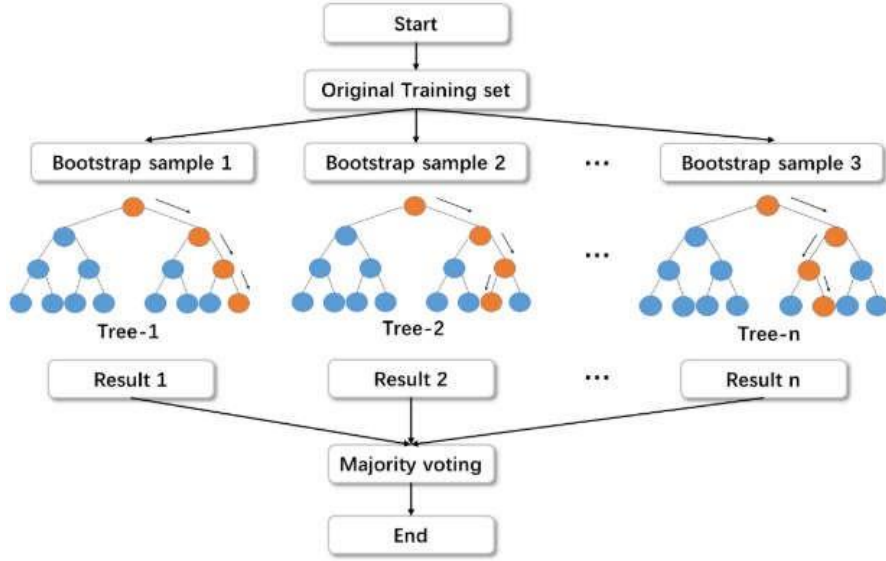


Fig 1. Working Process of Random Forest Method[26]

Figure 3 shows how the Random Forest algorithm works by creating a set of decision trees from a randomly selected subset, getting predictions from each decision tree, voting for each predicted outcome, and choosing the best prediction result based on the most votes assigned as final prediction

#### D. Evaluation Performance

Performance testing uses a confusion matrix table. The confusion matrix is a table that is used to describe the performance of the classification method on a dataset whose true value is known. The confusion matrix can visualize the amount of data that is classified as true and false as shown in the Table III[29].

TABLE III  
CONFUSION MATRIX

Actual	Predicted	
	Negative	Positive
Negative	TN	FP
Positive	FN	TP

Formula used to calculate *Accuracy* (6), *Sensitivity* (7), *Precision* (8) [30] [31][32], and *F1-score* (5)[33].

$$Accuracy = \frac{TP + TN}{TP + FN + TN + FP} \quad (6)$$

$$Sensitivity = \frac{TP}{TP + FN} \quad (7)$$

$$Precision = \frac{TP}{TP + FP} \quad (8)$$

$$F1 - score = \frac{2 \times precision \times recall}{precision + recall} \quad (9)$$

True Positive (TP) is a class of positive diabetes that is predicted correctly. False Positive (FP) is a diabetes negative

class but is predicted to be diabetes positive. True Negative (TN) is a diabetes negative class that is predicted correctly. False Negative (FN) is a positive diabetes class but is predicted to be diabetes negative.

### III. RESULT AND DISCUSSION

This research starts from the stages of data collection, data pre-processing, classification, and performance testing. The data used in this study is diabetes data obtained from Kaggle. The pre-processing of this study used the Smote and Smote-Tomeklink algorithms to deal with class imbalances in diabetes data. The classification method of this research is Random Forest. The performance test is based on accuracy, sensitivity, precision, and F1-score. The results of the comparison of the original data with the data from Smote and the results of Smote-Tomeklink are shown in Figure 4.



Fig 4. Data Distribution Result

The classification method of this research is Random Forest. Performance testing is based on accuracy, sensitivity, precision, and F1-score using a confusion matrix table. Based on testing the Random Forest method using 10-fold cross-validation, the results obtained in the form of a confusion matrix table as shown in Table IV for the Random Forest method on the original data, Table V for the results of the Random Forest



method with Smote, and Table VI for the results of the Random Forest method with Smote-Tomeklink. The results of the comparison of the performance of the Random Forest method as a whole are shown in Figure 5.

TABLE IV  
RESULT CONFUSION MATRIX OF RANDOM FOREST

Actual	Predicted	
	Negative	Positive
Negative	429	71
Positive	113	155

TABLE V  
RESULT CONFUSION MATRIX OF RANDOM FOREST AND SMOTE

Actual	Predicted	
	Negative	Positive
Negative	390	110
Positive	71	429

TABLE VI  
RESULT CONFUSION MATRIX OF RANDOM FOREST AND SMOTE-TOMEKLINK

Actual	Predicted	
	Negative	Positive
Negative	385	90
Positive	56	419

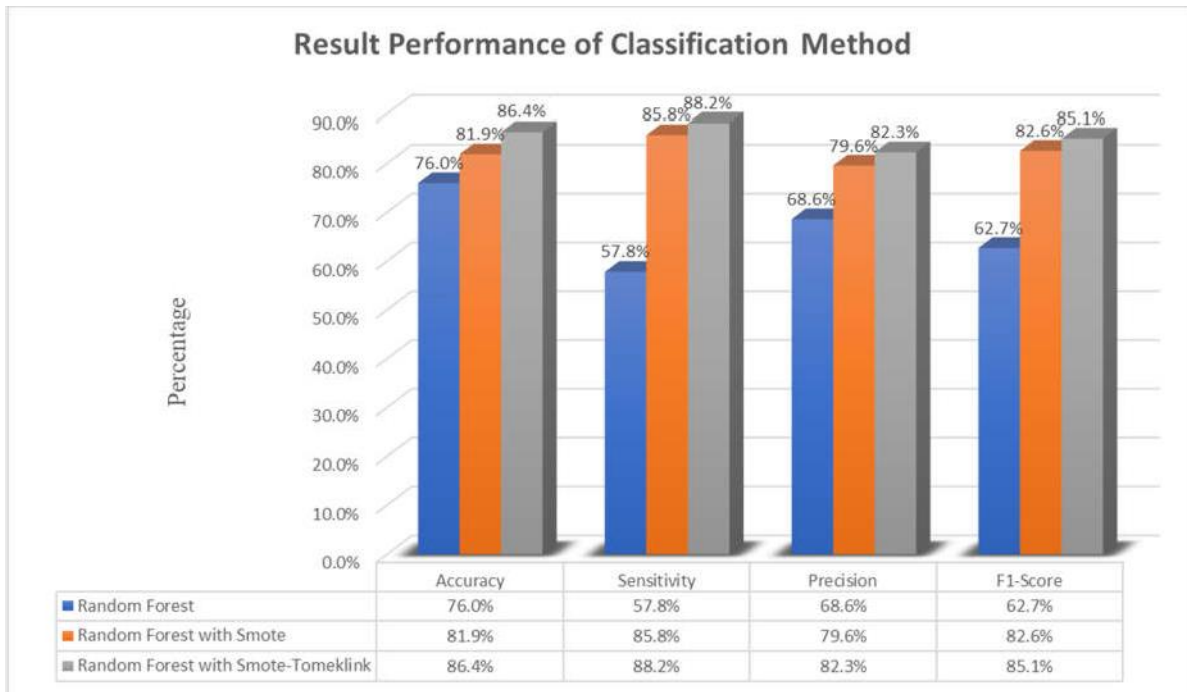


Fig 5. Result Performance of Classification Method

In Table IV, the Random Forest method succeeded in correctly classifying the negative class (TN) as many as 429 instances and the negative class classified incorrectly (FP) as many as 17 instances. While the correctly classified positive class (TP) is 155 instances and the incorrectly classified positive class is 113 instances.

In Table V, the Random Forest method with Smote succeeded in correctly classifying the negative class (TN) as many as 390 instances and the negative class classified incorrectly (FP) as many as 110 instances. While the positive class that is classified correctly (TP) is 429 instances and the positive class that is classified incorrectly is 71 instances.

In Table VI, the Random Forest method with Smote-Tomeklink succeeded in correctly classifying the negative class (TN) as many as 385 instances and the negative class classified incorrectly (FP) as 90 instances. While the positive class that is classified correctly (TP) is 419 instances and the positive class that is classified incorrectly is 56 instances.

Based on Figure 4, there was an increase in the performance of the Random Forest method with Smote-Tomeklink based on accuracy, sensitivity, precision, and F1-score. In the original dataset, the Random Forest method has 76% accuracy, 57.8%

sensitivity, 68.6% precision, and 62.7% F1-score. The Random Forest method with Smote has an accuracy of 81.9%, sensitivity of 85.8%, precision of 79.6%, and F1-score of 82.6%. Meanwhile, the use of the Random Forest method with Smote-Tomeklink resulted in an accuracy of 86.4%, a sensitivity of 88.2%, a precision of 83.3%, and F1-score of 85.1%.

Sensitivity has a very important role to improve the accuracy and F1-score performance of the Random Forest method with Smote-Tomeklink. The Random Forest method with Smote-Tomeklink gives higher accuracy, sensitivity, precision, and F1-score results than smote and without sampling.

Random Forest method with Smote an increase in performance indicators accuracy, sensitivity, precision, and F1-score. The increase in accuracy scores is 5.9%, Sensitivity is 28%, precision is 11%, and F1-score is 19.9%. The Random Forest method with Smote-Tomeklink showed an increase in the indicators of accuracy by 10.4%, Sensitivity by 30.4%, precision by 13.7%, and F1-score by 22.4%. Therefore, the use of the Smote-tomeklink method can increase accuracy, sensitivity, precision, and F1-score in the Random Forest method [11][34][35]. The comparison of the proposed method

is better than previous studies, which can be shown in Table VII.

TABLE VII  
COMPARISON OF THE PROPOSED MODEL PERFORMANCE WITH PREVIOUS STUDIES

No	Author (Year)	Dataset	Method	Accuracy
1	[16]	Pima Indian Diabetes	KNN	83%
2	[17]	Pima Indian Diabetes	Decision Tree C.45	75.65%
3	[11]	Pima Indian Diabetes	SVM + K-Means Smote	82%
4	[21]	Pima Indian Diabetes	Logistic Regression + Smote	82%
5	[22]	Pima Indian Diabetes	C4.5 Method + Smote	82%
6	<b>The Proposed Method</b>	<b>Pima Indian Diabetes</b>	<b>Random Forest + SMOTE</b>	<b>86%</b>

**Tomek links**

#### IV. CONCLUSION

This study applies the Smote-Tomeklink algorithm to the Random Forest method for the classification of diabetes. The implementation of Smote-Tomeklink can improve the performance of accuracy, sensitivity, precision, and F1-score in the Random Forest method. The combination of Random Forest and Smote-Tomeklink got the best accuracy, sensitivity, and precision compared to Smote and without sampling for the classification of diabetes. Where, there was an increase in performance indicators of 10.4% accuracy, 30.4% sensitivity, 13.7% precision, and 22.4 F1-score. Further research can apply Smote-Tomeklink to deal with the problem of data imbalance in multiclass data.

#### REFERENCES

- [1] O. Heranova, "Synthetic Minority Oversampling Technique pada Averaged One Dependence Estimators untuk Klasifikasi Credit Scoring," *Jurnal RESTI (Rekayasa Sistem dan Teknologi Informasi)*, vol. 3, no. 3, pp. 443–450, 2019, doi: 10.29207/resti.v3i3.1275.
- [2] T. Zhu, Y. Lin, and Y. Liu, "Synthetic minority oversampling technique for multiclass imbalance problems," *Pattern Recognition*, vol. 72, pp. 327–340, Dec. 2017, doi: 10.1016/j.patcog.2017.07.024.
- [3] F. Last, G. Douzas, and F. Bacao, "Oversampling for Imbalanced Learning Based on K-Means and SMOTE," no. November, 2017, [Online]. Available: <http://arxiv.org/abs/1711.00837>.
- [4] G. A. Pradipta, R. Wardoyo, A. Musdholifah, and I. N. H. Sanjaya, "Radius-SMOTE: A New Oversampling Technique of Minority Samples Based on Radius Distance for Learning from Imbalanced Data," *IEEE Access*, vol. 9, pp. 74763–74777, 2021, doi: 10.1109/ACCESS.2021.3080316.
- [5] M. Kamaladevi, V. Venkataraman, and K. R. Sekar, "Tomek link Undersampling with Stacked Ensemble classifier for Imbalanced data classification," vol. 25, no. 4, pp. 2182–2190, 2021.
- [6] W. C. Lin, C. F. Tsai, Y. H. Hu, and J. S. Jhang, "Clustering-based undersampling in class-imbalanced data," *Information Sciences*, vol. 409–410, pp. 17–26, 2017, doi: 10.1016/j.ins.2017.05.008.
- [7] Z. Xu, D. Shen, T. Nie, and Y. Kou, "A hybrid sampling algorithm combining M-SMOTE and ENN based on Random Forest for medical imbalanced data," *Journal of Biomedical Informatics*, p. 103465, 2020, doi: 10.1016/j.jbi.2020.103465.
- [8] E. AT, A. M, A.-M. F, and S. M, "Classification of Imbalance Data using Tomek Link (T-Link) Combined with Random Undersampling (RUS) as a Data Reduction Method," *Global Journal of Technology and Optimization*, vol. 01, no. S1, 2016, doi: 10.4172/2229-8711.s1111.
- [9] Z. Wang, C. Wu, K. Zheng, X. Niu, and X. Wang, "SMOTETomek-Based Resampling for Personality Recognition," *IEEE Access*, vol. 7, pp. 129678–129689, 2019, doi: 10.1109/ACCESS.2019.2940061.
- [10] N. V. Chawla, K. W. Bowyer, and L. O. Hall, "SMOTE: Synthetic Minority Over-sampling TEchnique," *Journal of Artificial Intelligence Research*, vol. 16, pp. 341–378, 2002.
- [11] H. Hairani, K. E. Saputro, and S. Fadli, "K-means-SMOTE for handling class imbalance in the classification of diabetes with C4.5, SVM, and naive Bayes," *Jurnal Teknologi dan Sistem Komputer*, vol. 8, no. 2, pp. 89–93, 2020, doi: 10.14710/jtsiskom.8.2.2020.89-93.
- [12] I. Tomek, "Tomek Link: Two Modifications of CNN," *IEEE Trans. Systems, Man and Cybernetics*, pp. 769–772, 1976, [Online]. Available: <https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&number=4309452>.
- [13] E. F. Swana, W. Doorsamy, and P. Bokoro, "Tomek Link and SMOTE Approaches for Machine Fault Classification with an Imbalanced Dataset," *Sensors*, vol. 22, no. 9, 2022, doi: 10.3390/s22093246.
- [14] A. Alzahrani and A. Safhi, "The role of data mining techniques and tools in big data management in healthcare field," *Sustainable Engineering and Innovation*, vol. 4, no. 1, pp. 58–65, 2022, doi: 10.37868/sei.v4i1.id128.
- [15] S. Sarać and B. Duraković, "Analysis of student performances in online and face-to-face learning: A case study from a Bosnian public university," *Heritage and Sustainable Development*, vol. 4, no. 2, pp. 87–94, 2022, doi: 10.37868/HSD.V4I2.91.
- [16] R. Kaur, "Predicting diabetes by adopting classification approach in data mining," *International Journal on Informatics Visualization*, vol. 3, no. 2–2, pp. 218–221, 2019, doi: 10.30630/ijov.3.2-2.229.
- [17] A. Azrar, M. Awais, Y. Ali, and K. Zaheer, "Data mining models comparison for diabetes prediction," *International Journal of Advanced Computer Science and Applications*, vol. 9, no. 8, pp. 320–323, 2018, doi: 10.14569/ijacsa.2018.090841.
- [18] S. Barik, S. Mohanty, S. Mohanty, and D. Singh, "Analysis of prediction accuracy of diabetes using classifier and hybrid machine learning techniques," *Smart Innovation, Systems and Technologies*, vol. 153, no. January, pp. 399–409, 2021, doi: 10.1007/978-981-15-6202-0\_41.
- [19] H. Hairani, M. Innuddin, and M. Rahardi, "Accuracy Enhancement of Correlated Naive Bayes Method by Using Correlation Feature Selection (CFS) for Health Data Classification," in *2020 3rd International Conference on Information and Communications Technology (ICOIACT)*, 2020, pp. 51–55, doi: 10.1109/ICOIACT50329.2020.9332021.
- [20] C. Fiarri, E. M. Sipayung, and S. Maemunah, "Analysis and prediction of diabetes complication disease using data mining algorithm," *Procedia Computer Science*, vol. 161, pp. 449–457, 2019, doi: 10.1016/j.procs.2019.11.144.
- [21] Erlin, Y. N. Marlim, Junadhi, L. Suryati, and N. Agustina, "Early Detection of Diabetes Using Machine Learning with Logistic Regression Algorithm," *Jurnal Nasional Teknik Elektro dan Teknologi Informasi*, vol. 11, no. 2, pp. 88–96, 2022.
- [22] C. Azad, B. Bhushan, R. Sharma, A. Shankar, K. K. Singh, and A. Khamparia, "Prediction model using SMOTE, genetic algorithm and decision tree (PMSGD) for classification of diabetes mellitus," *Multimedia Systems*, vol. 28, no. 4, pp. 1289–1307, 2022, doi: 10.1007/s00530-021-00817-2.
- [23] X. Shi, T. Qu, G. Van Pottelbergh, M. van den Akker, and B. De Moor, "A Resampling Method to Improve the Prognostic Model of End-Stage Kidney Disease: A Better Strategy for Imbalanced Data," *Frontiers in Medicine*, vol. 9, no. March, pp. 1–9, 2022, doi: 10.3389/fmed.2022.730748.
- [24] K. Wang *et al.*, "Improving risk identification of adverse outcomes in chronic heart failure using smote +enn and machine learning," *Risk Management and Healthcare Policy*, vol. 14, no. May, pp. 2453–2463, 2021, doi: 10.2147/RMHP.S310295.
- [25] H. Kaur, H. S. Pannu, and A. K. Malhi, "A systematic review on imbalanced data challenges in machine learning: Applications and solutions," *ACM Computing Surveys*, vol. 52, no. 4. Association for Computing Machinery, pp. 1–34, Aug. 01, 2019, doi: 10.1145/3343440.
- [26] K. Guo, X. Wan, L. Liu, Z. Gao, and M. Yang, "Fault diagnosis of intelligent production line based on digital twin and improved random forest," *Applied Sciences (Switzerland)*, vol. 11, no. 16, pp. 1–18, 2021, doi: 10.3390/app11167733.
- [27] J. Chen, H. Huang, A. G. Cohn, D. Zhang, and M. Zhou, "Machine

learning-based classification of rock discontinuity trace: SMOTE oversampling integrated with GBT ensemble learning," *International Journal of Mining Science and Technology*, vol. 32, no. 2, pp. 309–322, 2021, doi: 10.1016/j.ijmst.2021.08.004.

- [28] Y. Sun, H. Zhang, T. Zhao, Z. Zou, B. Shen, and L. Yang, "A New Convolutional Neural Network with Random Forest Method for Hydrogen Sensor Fault Diagnosis," *IEEE Access*, vol. 8, pp. 85421–85430, 2020, doi: 10.1109/ACCESS.2020.2992231.
- [29] H. Hartono and E. Ongko, "Avoiding Overfitting dan Overlapping in Handling Class Imbalanced Using Hybrid Approach with Smoothed Bootstrap Resampling and Feature Selection," *International Journal on Informatics Visualization*, vol. 6, no. June, pp. 343–348, 2022.
- [30] H. Hairani, A. Anggrawan, A. I. Wathan, K. A. Latif, K. Marzuki, and M. Zulfikri, "The Abstract of Thesis Classifier by Using Naive Bayes Method," in *Proceedings - 2021 International Conference on Software Engineering and Computer Systems and 4th International Conference on Computational Science and Information Management, ICSECS-ICOCSIM 2021*, 2021, no. August, pp. 312–315, doi: 10.1109/ICSECS52883.2021.00063.
- [31] A. Luque, A. Carrasco, A. Martin, and A. de las Heras, "The impact of class imbalance in classification performance metrics based on the binary confusion matrix," *Pattern Recognition*, vol. 91, pp. 216–231, 2019, doi: 10.1016/j.patcog.2019.02.023.
- [32] H. Qteat and M. Awad, "Using Hybrid Model of Particle Swarm Optimization and Multi-Layer Perceptron Neural Networks for Classification of Diabetes," *International Journal of Intelligent Engineering and Systems*, vol. 14, no. 3, pp. 11–22, 2021, doi: 10.22266/ijies2021.0630.02.
- [33] H. Hanafi, A. H. Muhammad, I. Verawati, and R. Hardi, "An Intrusion Detection System Using SDAE to Enhance Dimensional Reduction in Machine Learning," *International Journal on Informatics Visualization*, vol. 6, no. June, pp. 306–316, 2022.
- [34] H. Hairani, A. S. Suweleh, and D. Susilowaty, "Penanganan Ketidak Seimbangan Kelas Menggunakan Pendekatan Level Data," *MATRIK : Jurnal Manajemen, Teknik Informatika dan Rekayasa Komputer*, vol. 20, no. 1, pp. 109–116, 2020, doi: 10.30812/matrik.v20i1.846.
- [35] M. Y. Thanoun, M. T. Yaseen, and A. M. Alceesa, "Development of Intelligent Parkinson Disease Detection System Based on Machine Learning Techniques Using Speech Signal," *International Journal on Advanced Science, Engineering and Information Technology*, vol. 11, no. 1, pp. 388–392, 2021.

20 December 2022

Dear Hairani

Email: [hairani@universitasbumigora.ac.id](mailto:hairani@universitasbumigora.ac.id)

RE: JOURNAL ACCEPTANCE LETTER

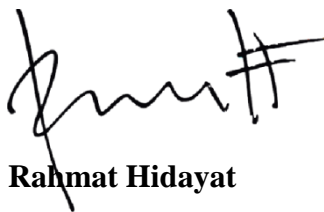
We are happy to inform you that the International Journal on Informatics Visualization (JOIV) has been indexed in Scopus. The Scientific committee of JOIV agrees that the following manuscript is **accepted** for publication in **JOIV**.

Title	Improvement Performance of the Random Forest Method on Unbalanced Diabetes Data Classification Using Smote-Tomek Link
-------	---

Thank you for your contribution the International Journal on Informatics Visualization (JOIV) and we look forward to receiving further submissions from you.

Sincerely

Regards,



**Rahmat Hidayat**

Editor in Chief  
International Journal on Informatics  
Visualization  
<http://joiv.org>

# Improvement Performance of the Random Forest Method on Unbalanced Diabetes Data Classification Using Smote-Tomek Link

*By Anthony Anggrawan*

# Improvement Performance of the Random Forest Method on Unbalanced Diabetes Data Classification Using Smote-Tomek Link

Hairani Hairani<sup>a\*</sup>, Anthony Anggrawan<sup>b</sup>, Dadang Priyanto<sup>c</sup>

<sup>abc</sup> Department of Computer Science, Universitas Bumigora, Mataram, 83127, Indonesia

Corresponding author: Hairani@universitasbumigora.ac.id

**Abstract**— Most of the health data contained unbalanced data that affected the performance of the classification method. Unbalanced data causes the classification method to more easily classify the majority data and ignore the minority class. One of the health data that has unbalanced data is Pima Indian Diabetes. Diabetes is a deadly disease caused by the body's inability to produce enough insulin. Complications of diabetes can cause heart attacks and strokes. Early diagnosis of diabetes is needed to minimize the occurrence of more severe complications. In the diabetes dataset used, there is an imbalance of data between positive and negative diabetes classes. Diabetes negative class data (500 data) is more than diabetes positive class (268) so that it can affect the performance of the classification method. Therefore, this study aims to apply the Smote-TomekLink and Random Forest methods in the classification of diabetes. The research methodology used is the collection of diabetes data obtained from Kaggle as many as 768 data with 8 input attributes and 1 output attribute as a class, pre-processing data is used to balance the dataset with Smote-TomekLink, classification using the random forest method, and performance evaluation based on accuracy, sensitivity, precision, and F1-score. Based on the tests carried out by dividing data using 10-fold cross-validation, the Random forest algorithm with Smote-TomekLink gets the highest accuracy, sensitivity, precision, and F1-score compared to Random Forest with Smote. The Random Forest algorithm with Smote-TomekLink has 86.4% accuracy, 88.2% sensitivity, 82.3% precision, and 85.1% F1-score. Thus, using Smote-TomekLink can improve the performance of the random forest method based on accuracy, sensitivity, precision, and F1-score.

**Keywords**— Class Imbalance; Smote-TomekLink; Random Forest Method; Diabetest Disease.

Manuscript received 15 Oct 2020; revised 29 Jan. 2021; accepted 2 Feb. 2021. Date of publication 17 Feb. 2021.  
International Journal on Informatics Visualization is licensed under a Creative Commons Attribution-Share Alike 4.0 International License.



## I. INTRODUCTION

Most of the Health data contained unbalanced data that affected the performance of the classification method. Unbalanced data causes the classification method to more easily classify the majority data and ignore the minority class. One of the health data that has unbalanced data is Pima Indian Diabetes. Diabetes is a deadly disease caused by increased blood sugar in the body. Diabetes is caused by the body's inability to produce enough insulin. Complications of diabetes can cause heart attacks and strokes. One way to improve the performance of the classification method is to handle balanced data by adding minority data so that the number is equal to the majority class. The diabetes dataset has 768 instances of data. However, the problem is that there is an imbalance of data in the dataset, namely the negative diabetes class with 500 data (majority class), while the positive diabetes class with 268 data (minority class). Data imbalance is the amount of data in one class more than in the other class. The problem of data imbalance causes the classification method to be more

dominant in classifying the majority class than the minority class, or in other words, the classification method ignores the minority class. The problem of unbalanced data can be handled with a data sampling approach.

Several data sampling methods that can be used to solve the problem of data imbalance are oversampling [1][2], [3][4], undersampling [5][6], and Hybrid Sampling[6],[7]. Oversampling works by adding the minority class, while Undersampling works by removing the majority class so as to produce balanced data. However, both methods have their respective weaknesses. The weakness of the oversampling method is that there are too many repetitions of samples that can cause overfitting of the classification method, while the weakness of undersampling is that it will lose information from most of the samples in the dataset and cannot take full advantage of the available information[9].

To avoid overfitting the oversampling method, the Smote method was developed to overcome these weaknesses. Smote is an oversampling method to generate new synthesis training data by linear interpolation on minority classes[10]. However,

the Smote method has a weakness, namely overgeneralization, and the addition of a minority class randomly can generate noise data, because it does not differentiate between classes [11]. Therefore, the undersampling method is used to improve the performance of the oversampling method by cleaning the noise data in the majority class. The noise data is the majority class instance which is closest to the minority class instance. Usually, noise data reduces the level of accuracy for classification methods [5]. One method to remove noise data in the majority class is Tomeklink [12]. Tomeklink is an undersampling method that cleans noise data from the majority class which has similar characteristics and overlapping. However, Tomeklink only deletes instances defined as "Tomek Links" so that the analyzed data cannot be balanced and in its implementation the method is combined with other methods. Combining Tomeklink and Smote oversampling can improve accuracy better than individual performance [13].

Data mining research in Health plays an important role, especially in predicting various types of diseases using different techniques or methods [14]. Research [15] uses a statistical approach to analyze the success rate of students following subjects using online or face-to-face learning. The results show that online students have significantly higher average grades than face-to-face classes.

Several previous studies have focused on the classification of diabetes, namely Research [16] predicts diabetes using the k-NN method with an accuracy of 83%. The weakness of the research is that it does not address the problem of data imbalance. Research [17] classifying diabetes using the C4.5 method with an accuracy of 75.65%. The weakness of the research is that the accuracy obtained is low so that it can be improved, and also does not address the problem of data imbalance. Research [18] Using XGBoost to predict diabetes with 74% accuracy. The weakness of the research is that the accuracy obtained is low so that it can be improved, and also does not address the problem of data imbalance.

Research [19] using the Correlated Naïve Bayes method with correlation-based feature selection to predict diabetes with an accuracy of 69.51%. The weakness of the research is that the accuracy gained is low so that it can be improved, and also does not address the problem of data imbalance. Research [20] using the C4.5 method for diabetes detection with an accuracy of 68%.

Research [21] used logistic regression and smote methods to detect diabetes with 82% accuracy, 81% precision, 79% recall, and 80% F1-score. The weakness of the research is that the accuracy is good but can be improved using Tomeklink to clean noise data in the majority class. Research [22] using the C4.5 and Smote methods to predict diabetes with 82% accuracy, 80% precision, and 86% sensitivity. Research [23] used logistic and Smote-ENN methods to predict kidney disease with 75.2% accuracy, 70.6% recall, 4.9% precision, and 30% F1-score. The weakness of the research is the low accuracy so that it can be improved using Tomeklink to clean noise data in the majority class. Research [24] SME-XGBoost with Smote-ENN for heart disease prediction with 80% AUC.

Several previous studies have applied various approaches to improve diabetes classification methods such as the oversampling approach with SMOTE. However, there are weaknesses in previous studies, namely the accuracy of the proposed method still ranges from 82% to 83% so that there is

a gap to improve its accuracy. So, this study proposes the Smote-Tomeklink hybrid sampling method to overcome the imbalance in diabetes data, so as to improve the accuracy of the classification method.

Smote-Tomeklink is a good way to avoid the drawbacks of SMOTE and Tomeklink techniques [9]. The classification method used in this research is Random Forest. The Random Forest method was chosen because it has several advantages, namely high accuracy [25], the ability to handle noise data, fast performance in training data, overfitting control, and easy to implement [26].

This study aims to apply the Smote-Tomeklink hybrid sampling method to balance the data on diabetes data so as to improve the performance of the Random forest classification method. Measurement of the performance of the random forest method based on accuracy, sensitivity (recall), precision, and F1-score

## II. MATERIALS AND METHOD

This research consists of several stages as shown in Figure 1.

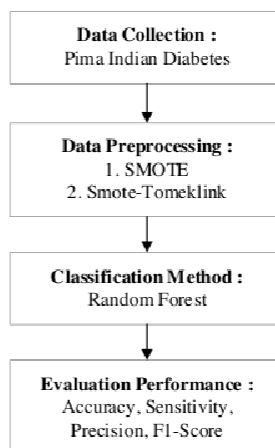


Fig 1. Research Stages

### A. Data Collection

The dataset used in this study is a diabetes dataset obtained from Kaggle, which consists of 768 instances and 9 attributes. The description of the attributes and the sample data used are shown respectively in Table I and Table II.

TABLE I  
DESCRIPTION ATRIBUT DATASET

No	Atribute	Description	Label
1	Pregnancies	Number of Pregnancy	X1
2	Glucose	Glucose level 2 hours after eating	X2
3	Blood Pressure	Blood Pressure	X3
4	Skin Thickness	Skin Thickness	X4
5	Insulin	Insulin	X5
6	BMI	Body Massa Index	X6
7	Diabetes Pedigree Function	Diabetes Pedigree Function	X7
8	Age	Age	X8

9	Outcome	Diabetes Status ( 1 = Positive Diabetes, 2 = Negative Diabetes	Y	767	1	126	60	0	0	30.1	0.349	47	1
				768	1	93	70	31	0	30.4	0.315	23	0

14

TABLE III  
SAMPLE DATASET

No	X1	X2	X3	X4	X5	X6	X7	X8	Y
1	6	148	72	35	0	33.6	0.627	50	1
2	1	85	66	29	0	26.6	0.351	31	0
3	8	183	64	0	0	23.3	0.672	32	1
4	1	89	66	23	94	28.1	0.167	21	0
5	0	137	40	35	168	43.1	2.288	33	1
6	5	116	74	0	0	25.6	0.201	30	1
7	3	78	50	32	88	31	0.248	26	1
8	10	115	0	0	0	35.3	0.134	29	0
9	2	197	70	45	543	30.5	0.158	53	1
10	8	125	96	0	0	0	0.232	54	1
11	4	110	92	0	0	37.6	0.191	30	0
12	10	168	74	0	0	38	0.537	34	1
13	10	139	80	0	0	27.1	1.441	57	0
14	1	189	60	23	846	30.1	0.398	59	1
15	5	166	72	19	175	25.8	0.587	51	1
..	..	..	..	..	..	..	..	..	..
754	0	181	88	44	510	43.3	0.222	26	1
755	8	154	78	32	0	32.4	0.443	45	1
756	1	128	88	39	110	36.5	1.057	37	1
757	7	137	90	41	0	32	0.391	39	0
758	0	123	72	0	0	36.3	0.258	52	1
759	1	106	76	0	0	37.5	0.197	26	0
760	6	190	92	0	0	35.5	0.278	66	1
761	2	88	58	26	16	28.4	0.766	22	0
762	9	170	74	31	0	44	0.403	43	1
763	9	89	62	0	0	22.5	0.142	33	0
764	10	101	76	48	180	32.9	0.171	63	0
765	2	122	70	27	0	36.8	0.34	27	0
766	5	121	72	23	112	26.2	0.245	30	0

12  
B. Data Pre-processing

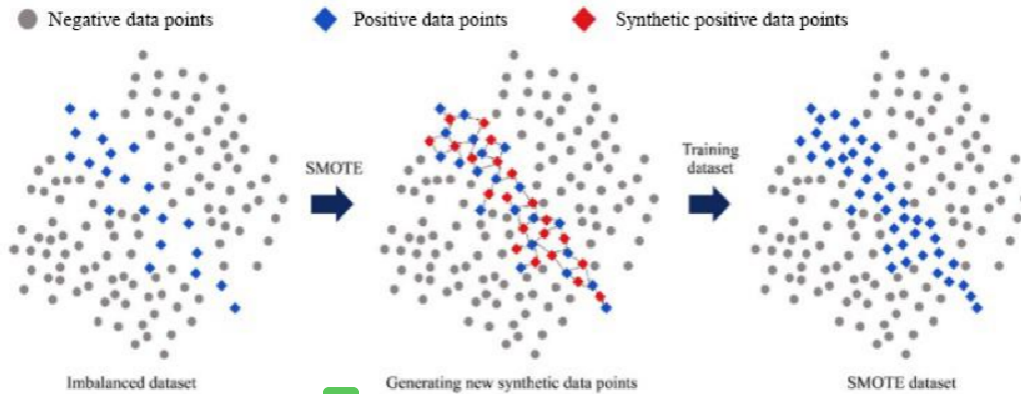
Data Pre-processing is one of the important stages in data mining to improve the quality of datasets. This study focuses on dealing with unbalanced data contained in the diabetes dataset. The dataset used has 268 instances of negative diabetes and 500 instances of Positive Diabetes. The algorithm used to handle unbalanced data in the dataset are SMOTE (Synthetic Minority Oversampling Technique) and Smote-Tomeklink.

SMOTE is one of the most commonly used oversampling methods to solve the problem of data distribution imbalance machine learning modeling. SMOTE aims to balance the distribution of classes by increasing the number of minority classes randomly by creating synthetic data for oversampling purposes [10]. Creating new data on the minority class using the equation (1).

$$Y' = Y^i + (Y^j - Y^i) * \gamma \quad (1)$$

36  
Y' is the representation of the addition of the minority class. Y<sup>i</sup> is the representation of minority class, Y<sup>j</sup> is a value chosen at random from the k-nearest neighbors of the minority class on Y<sup>i</sup>, and γ is a value in a randomly selected vector with a range of 0 to 1 [2].

SMOTE generates new synthesis training data by linear interpolation for the minority class. Synthesis training data is generated by randomly selecting 0.38 or more of the k-nearest neighbors for each sample in the minority class as shown in Figure 2.



29  
Fig 2. Process of Synthetic Minority Oversampling Technique (SMOTE) Algorithm [27]

Tomeklink is an undersampling method that cleans noise data from the majority class that has similar characteristics and overlapping [12]. Tomeklink works by eliminating the majority class instances that are closer to the minority class by applying the nearest neighbor rule to select instances. The combination of Tomeklink and Smote oversampling can improve accuracy better than individual performance [13].

8  
C. Random Forest Method

Random Forest is a decision tree-based ensemble learning method [28]. The Random Forest method has the advantages of high accuracy, the ability to handle noise data, fast performance training data, overfitting control, and easy to implement [26]. The working process of the Random Forest method in classifying a data is shown in Figure 3.



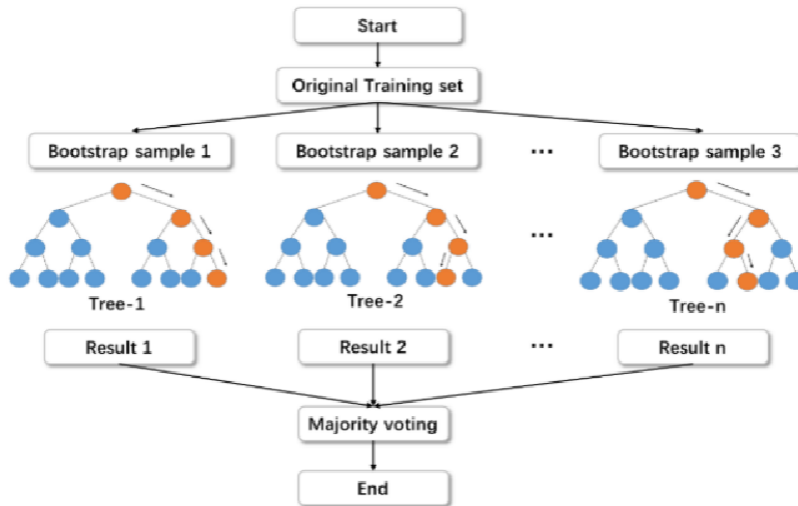


Fig 1. Working Process of Random Forest Method[26]

Figure 3 shows how the Random Forest algorithm works by creating a set of decision trees from a randomly selected subset, getting predictions from each decision tree, voting for each predicted outcome, and choosing the best prediction result based on the most votes assigned as final prediction

#### D. Evaluation Performance

Performance testing uses a confusion matrix table. The confusion matrix is a table that is used to describe the performance of the classification method on a dataset whose true value is known. The confusion matrix can visualize the amount of data that is classified as true and false as shown in the Table III[29].

TABLE 20  
CONFUSION MATRIX

Actual	Predicted	
	Negative	Positive
Negative	TN	FP
Positive	FN	TP

Formula used to calculate Accuracy (6), Sensitivity (7), Precision (8) [30] [31][32], and F1-score (5)[33].

$$Accuracy = \frac{TP+TN}{TP+FN+TN+FP} \quad (6)$$

$$Sensitivity = \frac{TP}{TP+FN} \quad (7)$$

$$Precision = \frac{TP}{TP+FP} \quad (8)$$

$$F1-score = \frac{2 \times precision \times recall}{precision + recall} \quad (9)$$

True Positive (TP) is a class of positive diabetes that is predicted correctly. False Positive (FP) is a diabetes negative

class but is predicted to be diabetes positive. True Negative (TN) is a diabetes negative class that is predicted correctly. False Negative (FN) is a positive diabetes class but is predicted to be diabetes negative.

### III. RESULT AND DISCUSSION

This research starts from the stages of data collection, data pre-processing, classification, and performance testing. The data used in this study is diabetes data obtained from Kaggle. The pre-processing of this study used the Smote and Smote-Tomeklink algorithms to deal with class imbalances in diabetes data. The classification method of this research is Random Forest. The performance is based on accuracy, sensitivity, precision, and F1-score. The results of the comparison of the original data with the data from Smote and the results of Smote-Tomeklink are shown in Figure 4.



Fig 4. Data Distribution Result

The classification method of this research is Random Forest. Performance testing is based on accuracy, sensitivity, precision, and F1-score using a confusion matrix table. Based on testing the Random Forest method using 10-fold cross-validation, the results obtained in the form of a confusion matrix table as shown in Table IV for the Random Forest method on the original data, Table V for the results of the Random Forest

method with Smote, and Table VI for the results of the Random Forest method with Smote-Tomeklink. The results of the comparison of the performance of the Random Forest method as a whole are shown in Figure 5.

TABLE IV  
RESULT CONFUSION MATRIX OF RANDOM FOREST

Actual	Predicted	
	Negative	Positive
Negative	429	71
Positive	113	155

TABLE V  
RESULT CONFUSION MATRIX OF RANDOM FOREST AND SMOTE

Actual	Predicted	
	Negative	Positive
Negative	390	110
Positive	71	429

TABLE VI  
RESULT CONFUSION MATRIX OF RANDOM FOREST AND SMOTE-TOMEKLINK

Actual	Predicted	
	Negative	Positive
Negative	385	90
Positive	56	419

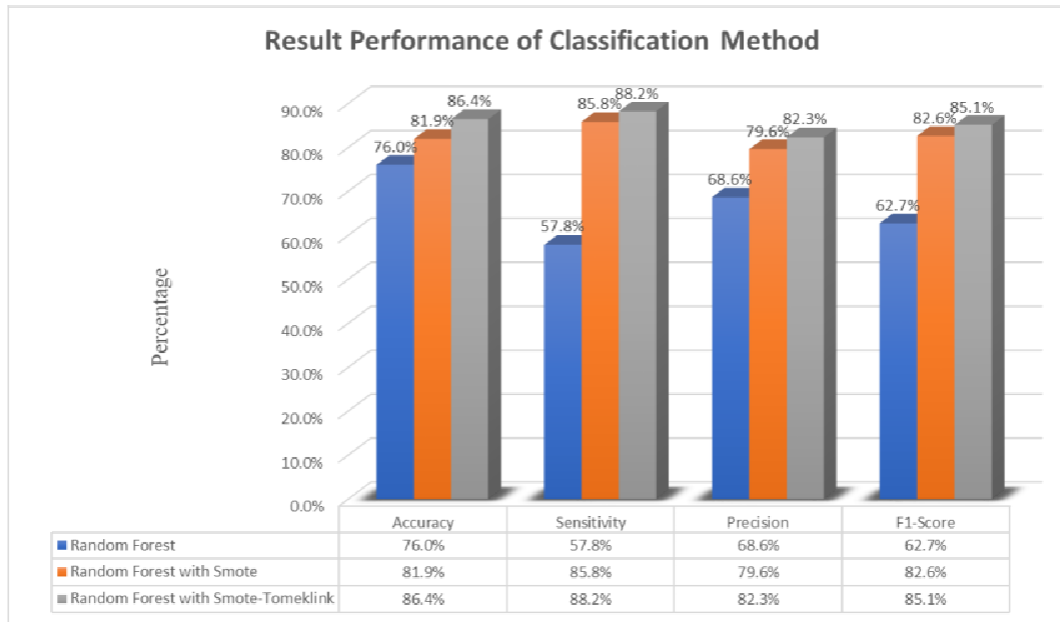


Fig 5. Result Performance of Classification Method

In Table IV, the Random Forest method succeeded in correctly classifying the negative class (TN) as many as 429 instances and the negative class classified incorrectly (FP) as many as 17 instances. While the correctly classified positive class (TP) is 155 instances and the incorrectly classified positive class is 113 instances.

In Table V, the Random Forest method with Smote succeeded in correctly classifying the negative class (TN) as many as 390 instances and the negative class classified incorrectly (FP) as many as 110 instances. While the positive class that is classified correctly (TP) is 429 instances and the positive class that is classified incorrectly is 71 instances.

In Table VI, the Random Forest method with Smote-Tomeklink succeeded in correctly classifying the negative class (TN) as many as 385 instances and the negative class classified incorrectly (FP) as 90 instances. While the positive class that is classified correctly (TP) is 419 instances and the positive class that is classified incorrectly is 56 instances.

Based on Figure 4, there was an increase in the performance of the Random Forest method with Smote-Tomeklink based on accuracy, sensitivity, precision, and F1-score. In the original dataset, the Random Forest method has 76% accuracy, 57.8%

sensitivity, 68.6% precision, and 62.7% F1-score. The Random Forest method with Smote has an accuracy of 81.9%, sensitivity of 85.8%, precision of 79.6%, and F1-score of 82.6%. Meanwhile, the use of the Random Forest method with Smote-Tomeklink resulted in an accuracy of 86.4%, a sensitivity of 88.2%, a precision of 83.3%, and F1-score of 85.1%.

Sensitivity has a very important role to improve the accuracy and F1-score performance of the Random Forest method with Smote-Tomeklink. The Random Forest method with Smote-Tomeklink gives higher accuracy, sensitivity, precision, and F1-score results than smote and without sampling.

Random Forest method with Smote an increase in performance indicators accuracy, sensitivity, precision, and F1-score. The increase in accuracy scores is 5.9%, Sensitivity is 28%, precision is 11%, and F1-score is 19.9%. The Random Forest method with Smote-Tomeklink showed an increase in the indicators of accuracy by 10.4%, Sensitivity by 30.4%, precision by 13.7%, and F1-score by 22.4%. Therefore, the use of the Smote-tomeklink method can increase accuracy, sensitivity, precision, and F1-score in the Random Forest method [11][34][35]. The comparison of the proposed method

is better than previous studies, which can be shown in Table VII.

TABLE VII  
COMPARISON OF THE PROPOSED MODEL PERFORMANCE WITH PREVIOUS STUDIES

No	Author (Year)	Dataset	Method	Accuracy
1	[16]	Pima Indian Diabetes	KNN	83%
2	[17]	Pima Indian Diabetes	Decision Tree C.45	75.65%
3	[11]	Pima Indian Diabetes	SVM + K-Means Smote	82%
4	[21]	Pima Indian Diabetes	Logistic Regression + Smote	82%
5	[22]	Pima Indian Diabetes	C4.5 Method + Smote	82%
6	The Proposed Method	Pima Indian Diabetes	Random Forest + SMOTE Tomek links	86%

#### IV. CONCLUSION

This study applies the Smote-Tomeklink algorithm to the Random Forest method for the classification of diabetes. The implementation of Smote-Tomeklink can improve the performance of accuracy, sensitivity, precision, and F1-score in the Random Forest method. The combination of Random Forest and Smote-Tomeklink got the best accuracy, sensitivity, and precision compared to Smote and without sampling for the classification of diabetes. Where, there was an increase in performance indicators of 10.4% accuracy, 30.4% sensitivity, 13.7% precision, and 2.4 F1-score. Further research can apply Smote-Tomeklink to deal with the problem of data imbalance in multiclass data.

#### REFERENCES

- [1] O. Heranova, "Synthetic Minority Oversampling Technique pada Averaged One Dependence Estimators untuk Klasifikasi Credit Scoring," *Jurnal RESTI (Rekayasa Sistem dan Teknologi Informatika)*, vol. 3, no. 3, pp. 443–450, 2019, doi: 10.29207/resti.v3i3.1275.
- [2] T. Zhu, Y. Lin, and Y. Liu, "Synthetic minority oversampling technique for multiclass imbalance problems," *Pattern Recognition*, vol. 72, pp. 327–340, Dec. 2017, doi: 10.1016/j.patcog.2017.07.024.
- [3] F. Last, G. Douzas, and F. Bacao, "Oversampling for Imbalanced Learning Based on K-Means and SMOTE," no. November, 2017, [Online]. Available: <http://arxiv.org/abs/1711.00837>.
- [4] G. A. Pradipta, R. Wardoyo, A. Musdholifah, and I. N. H. Sanjaya, "Radius-SMOTE: A New Oversampling Technique of Minority Samples Based on Radius Distance for Learning from Imbalanced Data," *IEEE Access*, vol. 9, pp. 74763–74777, 2021, doi: 10.1109/ACCESS.2021.3080316.
- [5] M. Kamaladevi, V. Venkataraman, and K. R. Sekar, "Tomek link Undersampling with Stacked Ensemble classifier for Imbalanced data classification," vol. 25, no. 4, pp. 2182–2190, 2021.
- [6] W. C. Lin, C. F. Tsai, Y. H. Hu, and J. S. Jhang, "Clustering-based undersampling in class-imbalanced data," *Information Sciences*, vol. 409–410, pp. 17–26, 2017, doi: 10.1016/j.ins.2017.05.008.
- [7] Z. Xu, D. Shen, T. Nie, and Y. Kou, "A hybrid sampling algorithm combining M-SMOTE and ENN based on Random Forest for medical imbalanced data," *Journal of Biomedical Informatics*, p. 103465, 2020, doi: 10.1016/j.jbi.2020.103465.
- [8] E. AT, A. M. A.-M. F., and S. M., "Classification of Imbalance Data using Tomek Link (T-Link) Combined with Random Under-sampling (RUS) as a Data Reduction Method," *Global Journal of Technology and Optimization*, vol. 01, no. S1, 2016, doi: 10.4172/2229-8711.s1111.
- [9] Z. Wang, C. Wu, K. Zheng, X. Niu, and X. Wang, "SMOTETomek-Based Resampling for Personality Recognition," *IEEE Access*, vol. 7, pp. 129678–129689, 2019, doi: 10.1109/ACCESS.2019.2940061.
- [10] N. V. Chawla, K. W. Bowyer, and L. O. Hall, "SMOTE: Synthetic Minority Over-sampling Technique," *Journal of Artificial Intelligence Research*, vol. 16, pp. 341–378, 2002.
- [11] H. Hairani, K. E. Saputro, and S. Fadli, "K-means-SMOTE for handling class imbalance in the classification of diabetes with C4.5, SVM, and naive Bayes," *Jurnal Teknologi dan Sistem Komputer*, vol. 8, no. 2, pp. 89–93, 2020, doi: 10.14710/jtsiskom.8.2.2020.89-93.
- [12] I. Tomek, "Tomek Link: Two Modifications of CNN," *IEEE Trans. Systems, Man and Cybernetics*, pp. 769–772, 1976, [Online]. Available: <https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&number=4309452>.
- [13] E. F. Swana, W. Doorsamy, and P. Bokoro, "Tomek Link and SMOTE Approaches for Machine Fault Classification with an Imbalanced Dataset," *Sensors*, vol. 22, no. 9, 2022, doi: 10.3390/s22093246.
- [14] A. Alzahrani and A. Safhi, "The role of data mining techniques and tools in big data management in healthcare field," *Sustainable Engineering and Innovation*, vol. 4, no. 1, pp. 58–65, 2022, doi: 10.37868/sei.v4i1.id128.
- [15] S. Sarač and B. Duraković, "Analysis of student performances in online and face-to-face learning: A case study from a Bosnian public university," *Heritage and Sustainable Development*, vol. 4, no. 2, pp. 87–94, 2022, doi: 10.37868/HSD.V4I2.91.
- [16] R. Kaur, "Predicting diabetes by adopting classification approach in data mining," *International Journal on Informatics Visualization*, vol. 3, no. 2–2, pp. 218–221, 2019, doi: 10.30630/ijov.3.2.2.229.
- [17] A. Azrar, M. Awais, Y. Ali, and K. Zaheer, "Data mining models comparison for diabetes prediction," *International Journal of Advanced Computer Science and Applications*, vol. 9, no. 8, pp. 320–323, 2018, doi: 10.14569/ijcsa.2018.090841.
- [18] S. Barik, S. Mohanty, S. Mohanty, and D. Singh, "Analysis of prediction accuracy of diabetes using classifier and hybrid machine learning techniques," *Smart Innovation, Systems and Technologies*, vol. 153, no. January, pp. 399–409, 2021, doi: 10.1007/978-981-15-6202-0\_41.
- [19] H. Hairani, M. Innuddin, and M. Rahardi, "Accuracy Enhancement of Correlated Naive Bayes Method by Using Correlation Feature Selection (CFS) for Health Data Classification," in *2020 3rd International Conference on Information and Communications Technology (ICOIACT)*, 2020, pp. 51–55, doi: 10.1109/ICOIACT50329.2020.9332021.
- [20] C. Fiarni, E. M. Sipayung, and S. Maemunah, "Analysis and prediction of diabetes complication disease using data mining algorithm," *Procedia Computer Science*, vol. 161, pp. 449–457, 2019, doi: 10.1016/j.procs.2019.11.144.
- [21] Erlin, Y. N. Marlum, Junadhi, L. Suryati, and N. Agustina, "Early Detection of Diabetes Using Machine Learning with Logistic Regression Algorithm," *Jurnal Nasional Teknik Elektro dan Teknologi Informatika*, vol. 11, no. 2, pp. 88–96, 2022.
- [22] C. Azad, B. Bhushan, R. Sharma, A. Shankar, K. K. Singh, and A. Khamparia, "Prediction model using SMOTE, genetic algorithm and decision tree (PMSGD) for classification of diabetes mellitus," *Multimedia Systems*, vol. 28, no. 4, pp. 1289–1307, 2022, doi: 10.1007/s00530-021-00817-2.
- [23] X. Shi, T. Qu, G. Van Pottelbergh, M. van den Akker, and B. De Moor, "A Resampling Method to Improve the Prognostic Model of End-Stage Kidney Disease: A Better Strategy for Imbalanced Data," *Frontiers in Medicine*, vol. 9, no. March, pp. 1–9, 2022, doi: 10.3389/fmed.2022.730748.
- [24] K. Wang *et al.*, "Improving risk identification of adverse outcomes in chronic heart failure using smote+enn and machine learning," *Risk Management and Healthcare Policy*, vol. 14, no. May, pp. 2453–2463, 2021, doi: 10.2147/RMHP.S310295.
- [25] H. Kaur, H. S. Pannu, and A. K. Malhi, "A systematic review on imbalanced data challenges in machine learning: Applications and solutions," *ACM Computing Surveys*, vol. 52, no. 4, Association for Computing Machinery, pp. 1–34, Aug. 01, 2019, doi: 10.1145/3343440.
- [26] K. Guo, X. Wan, L. Liu, Z. Gao, and M. Yang, "Fault diagnosis of intelligent production line based on digital twin and improved random forest," *Applied Sciences (Switzerland)*, vol. 11, no. 16, pp. 1–18, 2021, doi: 10.3390/app11167733.
- [27] J. Chen, H. Huang, A. G. Cohn, D. Zhang, and M. Zhou, "Machine

- learning-based classification of rock discontinuity trace: SMOTE oversampling integrated with GBT ensemble learning," *International Journal of Mining Science and Technology*, vol. 32, no. 2, pp. 309–322, 2021, doi: 10.1016/j.ijmst.2021.08.004.
- [28] Y. Sun, H. Zhang, T. Zhao, Z. Zou, B. Shen, and L. Yang, "A New Convolutional Neural Network with Random Forest Method for Hydrogen Sensor Fault Diagnosis," *IEEE Access*, vol. 8, pp. 85421–85430, 2020, doi: 10.1109/ACCESS.2020.2992231.
- [29] H. Hartono and E. Ongko, "Avoiding Overfitting dan Overlapping in Handling Class Imbalanced Using Hybrid Approach with Smoothed Bootstrap Resampling and Feature Selection," *International Journal on Informatics Visualization*, vol. 6, no. June, pp. 343–348, 2022.
- [30] H. Hairani, A. Anggrawan, A. I. Wathan, K. A. Latif, K. Marzuki, and M. Zulfikri, "The Abstract of Thesis Classifier by Using Naive Bayes Method," in *Proceedings - 2021 International Conference on Software Engineering and Computer Systems and 4th International Conference on Computational Science and Information Management, ICSECS-ICOCSIM 2021*, 2021, no. August, pp. 312–315, doi: 10.1109/ICSECS52883.2021.00063.
- [31] A. Luque, A. Carrasco, A. Martin, and A. de las Heras, "The impact of class imbalance in classification performance metrics based on the binary confusion matrix," *Pattern Recognition*, vol. 91, pp. 216–231, 2019, doi: 10.1016/j.patcog.2019.02.023.
- [32] H. Qteat and M. Awad, "Using Hybrid Model of Particle Swarm Optimization and Multi-Layer Perceptron Neural Networks for Classification of Diabetes," *International Journal of Intelligent Engineering and Systems*, vol. 14, no. 3, pp. 11–22, 2021, doi: 10.22266/ijies2021.0630.02.
- [33] H. Hanafi, A. H. Muhammad, I. Verawati, and R. Hardi, "An Intrusion Detection System Using SDAE to Enhance Dimensional Reduction in Machine Learning," *International Journal on Informatics Visualization*, vol. 6, no. June, pp. 306–316, 2022.
- [34] H. Hairani, A. S. Suweleh, and D. Susilowaty, "Penanganan Ketidak Seimbangan Kelas Menggunakan Pendekatan Level Data," *MATRIK : Jurnal Manajemen, Teknik Informatika dan Rekayasa Komputer*, vol. 20, no. 1, pp. 109–116, 2020, doi: 10.30812/matrik.v20i1.846.
- [35] M. Y. Thanoun, M. T. Yaseen, and A. M. Aleesa, "Development of Intelligent Parkinson Disease Detection System Based on Machine Learning Techniques Using Speech Signal," *International Journal on Advanced Science, Engineering and Information Technology*, vol. 11, no. 1, pp. 388–392, 2021.

# Improvement Performance of the Random Forest Method on Unbalanced Diabetes Data Classification Using Smote-Tomek Link

---

ORIGINALITY REPORT

---

# 13%

SIMILARITY INDEX

---

PRIMARY SOURCES

---

- |   |   |                 |
|---|---|-----------------|
| 1 | <a href="https://link.springer.com">link.springer.com</a><br>Internet   | 17 words – < 1% |
| 2 | <a href="https://oaji.net">oaji.net</a><br>Internet   | 17 words – < 1% |
| 3 | <a href="https://www.researchgate.net">www.researchgate.net</a><br>Internet   | 17 words – < 1% |
| 4 | Shuai Zhang, Sheng Zhang, Zhuzhong Qian, Jie Wu, Yibo Jin, Sanglu Lu. "DeepSlicing: Collaborative and Adaptive CNN Inference With Low Latency", IEEE Transactions on Parallel and Distributed Systems, 2021<br>Crossref | 16 words – < 1% |
| 5 | Ya-Qi Chen, Jianjun Zhang, Wing W. Y. Ng. "Loan Default Prediction Using Diversified Sensitivity Undersampling", 2018 International Conference on Machine Learning and Cybernetics (ICMLC), 2018<br>Crossref            | 16 words – < 1% |
| 6 | <a href="https://www.semanticscholar.org">www.semanticscholar.org</a><br>Internet   | 16 words – < 1% |
| 7 | Annesha Ahsan, Nazmun Nessa Moon, Shayla Sharmin, Mohammad Monirul Islam, Refath Ara  | 14 words – < 1% |

Hossain, Samia Nawshin. "Machine Learning Approach to Predict Traffic Accident Occurrence in Bangladesh", 2021 IEEE International Conference on Biomedical Engineering, Computer and Information Technology for Health (BECITHCON), 2021

Crossref

8 Boucekine, M., L. Boyer, K. Baumstarck, A. Millier, B. Ghattas, P. Auquier, and M. Toumi. "Exploring the Response Shift Effect on the Quality of Life of Patients with Schizophrenia: An Application of the Random Forest Method", Medical Decision Making, 2014. 14 words – < 1%

Crossref

9 Mengfei Wu, Ximing Li. "Unbalanced Data Classification Algorithm Based on Hybrid Sampling and Ensemble Learning", 2021 16th International Conference on Intelligent Systems and Knowledge Engineering (ISKE), 2021 14 words – < 1%

Crossref

10 Irfan Pratama, Yoga Pristyanto, Putri Taqwa Prasetyaningrum. "Imbalanced Class handling and Classification on Educational Dataset", 2021 4th International Conference on Information and Communications Technology (ICOIACT), 2021 13 words – < 1%

Crossref

11 Lusi Li, Haibo He, Jie Li, Weijun Li. "EDOS: Entropy Difference-based Oversampling Approach for Imbalanced Learning", 2018 International Joint Conference on Neural Networks (IJCNN), 2018 13 words – < 1%

Crossref

12 [ijesrt.com](http://ijesrt.com) 13 words – < 1%

Internet

13 [pdfs.semanticscholar.org](https://pdfs.semanticscholar.org)

Internet

12 words – < 1%

14 repository.universitasbumigora.ac.id

Internet

12 words – < 1%

15 www.ijain.org

Internet

12 words – < 1%

16 Zainab S. Alharthi, Abdullah Alsaeedi, Wael M.S. Yafooz. "Software Defect Prediction Approaches: A Review", 2021 4th International Conference on Bio-Engineering for Smart Technologies (BioSMART), 2021

Crossref

11 words – < 1%

17 Iman Almomani, Raneem Qaddoura, Maria Habib, Samah Alsoghyer, Alaa Al Khayer, Ibrahim Aljarah, Hossam Faris. "Android Ransomware Detection Based on a Hybrid Evolutionary Approach in the Context of Highly Imbalanced Data", IEEE Access, 2021

Crossref

10 words – < 1%

18 Toshita Sharma, Manan Shah. "A comprehensive review of machine learning techniques on diabetes detection", Visual Computing for Industry, Biomedicine, and Art, 2021

Crossref

10 words – < 1%

19 jiki.cs.ui.ac.id

Internet

10 words – < 1%

20 www.ijsr.net

Internet

10 words – < 1%

21 Lei Wang, ZhiQiang Zhao, YanHong Luo, HongMei Yu, ShuQing Wu, XiaoLu Ren, ChuChu Zheng, XueQian Huang. "Classifying 2-year recurrence in patients with

9 words – < 1%

DLBCL using clinical variables with imbalanced data and machine learning methods", Computer Methods and Programs in Biomedicine, 2020

Crossref

**22** Miao Wang, JiWen Wang, YaTing Hu, BinBin Guo, Hong Tang. "Detection of pulmonary hypertension with six training strategies based on deep learning technology", Computational Intelligence, 2022

9 words – < 1%

Crossref

**23** Truong Ho-Quang, Arif Nurwidyantoro, Satrio Adi Rukmono, Michel R.V. Chaudron, Fabian Fröding, Duy Nguyen Ngoc. "Role stereotypes in software designs and their evolution", Journal of Systems and Software, 2022

9 words – < 1%

Crossref

**24** [hsd.ardascience.com](http://hsd.ardascience.com)

Internet

9 words – < 1%

**25** [www.cs.bham.ac.uk](http://www.cs.bham.ac.uk)

Internet

9 words – < 1%

**26** [www.ncbi.nlm.nih.gov](http://www.ncbi.nlm.nih.gov)

Internet

9 words – < 1%

**27** H. Aljawazneh, A.M. Mora, P. Garcia-Sanchez, P.A. Castillo. "Comparing the Performance of Deep Learning Methods to Predict Companies' Financial Failure", IEEE Access, 2021

8 words – < 1%

Crossref

**28** Jun Guan, Xu Jiang, Baolei Mao. "A Method for Class-Imbalance Learning in Android Malware Detection", Electronics, 2021

8 words – < 1%

Crossref



---

**29** M. F. Mridha, Akibur Rahman Prodeep, A. S. M. Morshedul Hoque, Md. Rashedul Islam et al. "A Comprehensive Survey on the Progress, Process, and Challenges of Lung Cancer Detection and Classification", *Journal of Healthcare Engineering*, 2022

Crossref

8 words – < 1%

---

**30** Pin Lyu, Hanbin Zhang, Wenbing Yu, Chao Liu. "A novel model-independent data augmentation method for fault diagnosis in smart manufacturing", *Procedia CIRP*, 2022

Crossref

8 words – < 1%

---

**31** Rahma Fadhila Moenggah, Donni Richasdy, Mahendra Dwifabri Purbolaksono. "Telkom University Slogan Analysis on YouTube Using Naïve Bayes", *2022 International Conference on Data Science and Its Applications (ICoDSA)*, 2022

Crossref

8 words – < 1%

---

**32** [arxiv.org](https://arxiv.org)

Internet

8 words – < 1%

---

**33** [docplayer.net](https://docplayer.net)

Internet

8 words – < 1%

---

**34** [etd.aau.edu.et](https://etd.aau.edu.et)

Internet

8 words – < 1%

---

**35** [repository.futminna.edu.ng:8080](https://repository.futminna.edu.ng:8080)

Internet

8 words – < 1%

---

**36** "Computer Vision - ECCV 2020 Workshops", Springer Science and Business Media LLC, 2020

Crossref

7 words – < 1%

---

**37** Fanny Ramadhani, Al-Khowarizmi, Indah Purnama Sari. "Improving the Performance of Naïve Bayes Algorithm by Reducing the Attributes of Dataset Using Gain Ratio and Adaboost", 2021 International Conference on Computer Science and Engineering (IC2SE), 2021

Crossref

7 words – < 1%

---

**38** Na Liu, Xiaomei Li, Ershi Qi, Man Xu, Ling Li, Bo Gao. "A Novel Ensemble Learning Paradigm for Medical Diagnosis With Imbalanced Data", IEEE Access, 2020

Crossref

7 words – < 1%

---

**39** Hui Fern Soon, Amiza Amir, Saidatul Norlyana Azemi. "An Analysis of Multiclass Imbalanced Data Problem in Machine Learning for Network Attack Detections", Journal of Physics: Conference Series, 2021

Crossref

6 words – < 1%

---

**40** Matloob Khushi, Kamran Shaukat, Talha Mahboob Alam, Ibrahim A. Hameed et al. "A Comparative Analysis of Data Resampling Methods on Imbalance Medical Data", IEEE Access, 2021

Crossref

6 words – < 1%

---

**41** [jtsiskom.undip.ac.id](http://jtsiskom.undip.ac.id)

Internet

6 words – < 1%

---

EXCLUDE QUOTES ON

EXCLUDE SOURCES OFF

EXCLUDE BIBLIOGRAPHY ON

EXCLUDE MATCHES OFF