A New Approach of Hybrid Sampling SMOTE and ENN to the Accuracy of Machine Learning Methods on Unbalanced Diabetes Disease Data

Hairani Hairani, Dadang Priyanto
Department Computer Science, Bumigora University, Mataram, Indonesia

***Abstract*—The performance of machine learning methods in disease classification is affected by the quality of the dataset, one of which is unbalanced data. One example of health data that has unbalanced data is diabetes disease data. If unbalanced data is not addressed, it can affect the performance of the classification method. Therefore, this research proposed the SMOTE-ENN approach to improving the performance of the Support Vector Machine (SVM) and Random Forest classification methods for diabetes disease prediction. The methods used in this research were SVM and Random Forest classification methods with SMOTE-ENN. The SMOTE-ENN method was used to balance the diabetes data and remove noise data adjacent to the majority and minority classes. Data that has been balanced was predicted using SVM and Random Forest methods based on the division of training and testing data with 10-fold cross-validation. The results of this study were Random Forest method with SMOTE-ENN got the best performance compared to the SVM method, such as accuracy of 95.8%, sensitivity of 98.3%, and specificity of 92.5%. In addition, the proposed method approach (Random Forest with SMOTE-ENN) also obtained the best accuracy compared to previous studies referenced. Thus, the proposed method can be adopted to predict diabetes in a health application.**

***Keywords*—SMOTE-ENN; data imbalance; SVM; random forest; health dataset**

I. INTRODUCTION

Machine learning methods on health data, especially disease classification, have been widely practiced. The problem is that the dataset's quality influences the performance of machine learning methods in disease classification. In general, most health data, especially disease data, have data imbalance problems, such as diabetes [1], heart [2][3], and breast cancer [4]. If the problem of unbalanced data in health datasets is not addressed, it can affect the performance of classification methods, making the prediction results biased. With balanced data, classification methods can easily predict the majority class more accurately than the minority class. Therefore, this research seeks a method approach for handling unbalanced data on health data, especially diabetes disease data, so that classification methods achieve optimal accuracy.

Some previous studies have predicted diseases using various approaches, such as research [5] using the logistic regression machine learning method with SMOTE for predicting diabetes with an accuracy of 77%, precision of 75%, recall of 77%, and F1-score 76%. Research [6] uses forward chaining and certainty factor methods to diagnose types of rheumatic diseases with an accuracy of 80%. Research [7] uses the SMOTE method approach with machine learning algorithms such as Xgboost, Random Forest, KNN, Logistic regression, Decision Tree, Naive Bayes, and XGBoost for liver disease prediction with an accuracy of 80%. Based on the results of their research, the XGBoost method with SMOTE produces better performance than other methods, with accuracy of 93%, Recall of 97%, Precision of 92%, and F1-Score of 94%.

Research [4] uses a hybrid sampling method (SMOTE and SpreadSupsample) with several machine learning methods such as Naive Bayes, Decision Tree C4.5, and Random Forest for breast cancer disease prediction. His research shows that the use of hybrid sampling can improve the performance of the machine learning methods used, such as accuracy, ROC, Recall, and Precision. Research [8] uses hybrid sampling (SMOTE-ENN) with the Artificial Neural Network (ANN) method for the identification of Marburg virus inhibitors. The results show that using hybrid sampling (SMOTE-ENN) can effectively increase the ANN method's accuracy. Research [9] uses a hybrid sampling approach (M-SMOTE-ENN) with the Random Forest calcification method to solve unbalanced data problems in health data. The results show that using hybrid sampling (M-SMOTE and ENN) can improve the performance of the Random Forest method better than oversampling SMOTE and ENN individually without being combined.

Research [10] uses machine learning methods such as KNN, Decision Tree, Naïve Bayes, Random Forest, SVM, and histogram-based gradient boosting (HBGB) for diabetes prediction. The results show that the HBGB method performs better than other methods, with an accuracy of 92%. Research [11] compares several classification methods in machine learning for diabetes detection. The results show that the XGBoost method has better accuracy than other models, which is 94%. Research [12] uses a combination feature selection approach with several machine learning classification methods for diabetes detection. The results show that feature selection methods can improve the classification methods' accuracy. Random Forest is the method that gets the best accuracy, with a feature selection of 80%.

Research [13] uses a hybrid sampling approach (SMOTE-Tomek Link) with the Random Forest method for predicting diabetes. At the same time, the results show that the hybrid sampling method (SMOTE-Tomek Link) increases the

accuracy of the Random Forest method compared to SMOTE and Tomek Link separately. Research [14] predicts the risk of diabetes using a hybrid sampling approach (SMOTE-Tomek Link) with the ANN method. The results show that using hybrid sampling SMOTE-Tomek Link is better than SMOTE alone, with an accuracy of 92%.

Based on previous research, a gap can be improved; that is, the accuracy obtained in predicting diabetes is not optimal, so it can still be increased. Based on research [13][14], the highest accuracy is 92% using a hybrid sampling SMOTE-Tomek link with ANN. Therefore, this study proposes a hybrid sampling SMOTE-ENN approach to improving performance, such as accuracy, sensitivity, and specificity in SVM and Random Forest classification methods. This research adopts the use of the SMOTE-ENN hybrid sampling method, as it performs better than SMOTE-Tomek Links [15][16].

The purpose of this study is the implementation of hybrid sampling SMOTE-ENN to increase the accuracy of the machine learning method in predicting unbalanced diabetes data. This study consists of an introduction structure, research method, results and discussion, and conclusion.

II. Research Method

This study has several stages shown in Fig. 1. The first stage is the collection of diabetes disease datasets obtained from the Uci Repository with a total dataset of 768 instances and ten attributes. Attributes owned by the Pima Indian Diabetes contain datasets such as Pregnancies, Glucose, Blood Pressure, Skin Thickness, Insulin, BMI, Diabetes Pedigree Function, Age, and Outcome (Class). The dataset used has a total of two class categories, namely positive diabetes and negative diabetes.

The second stage is data preprocessing which is useful for improving the quality of the dataset used, such as removing missing values, outliers, and unbalanced data in the data preprocessing phase using data sampling to balance the data in the diabetes class, where there is a smaller number of positive classes (minority classes) compared to negative classes (majority classes) so that it can affect the performance of the classification method. If unbalanced diabetes data is not handled, the classification method will find it easier to classify the majority (negative) class than the minority (positive) class. In other words, the classification method makes biased prediction results.

This research uses several data sampling methods such as SMOTE, ENN, and hybrid sampling SMOTE-ENN. The SMOTE-ENN method combines SMOTE oversampling and ENN undersampling. The way the SMOTE-ENN method works is to add artificial data to the minority class by interpolating the original data using SMOTE so that the resulting artificial data is balanced. After the data is balanced, samples from the majority class adjacent to the minority class are removed by undersampling ENN. The use of the SMOTE-ENN method can reduce data overfitting and noise. The method of SMOTE-ENN can be shown in Fig. 2.

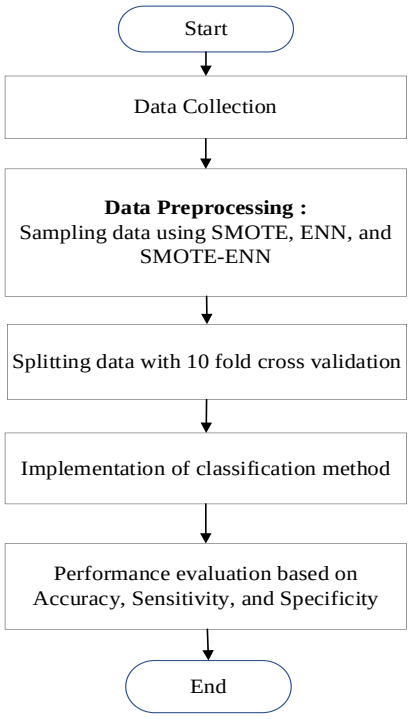

Fig. 1. Research stages.

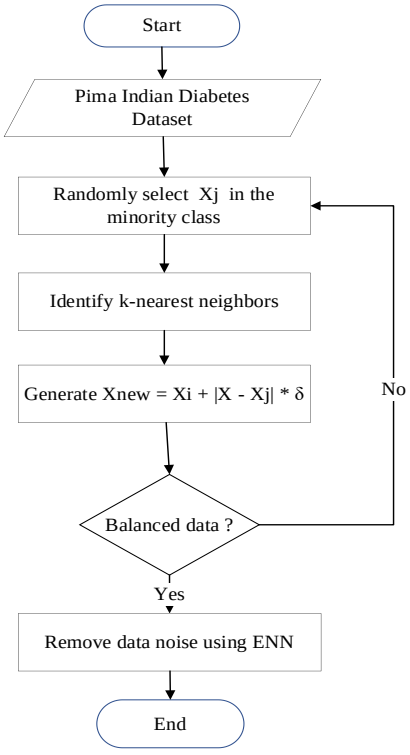

Fig. 2. SMOTE-ENN process.

After the data sampling, the next step is to divide the training and testing data using 10-fold cross-validation. 10-fold cross-validation works by dividing the data into ten groups, and each group can be used as training and testing data alternately. The illustration of how 10-fold cross-validation works is shown in Fig. 3.

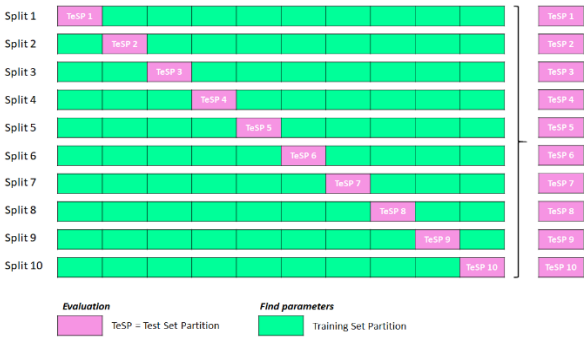

Fig. 3. Process of 10-Fold cross validation.

Data divided into 10 folds are then used to implement classification methods using SVM and Random Forest methods. The classification results of the SVM and Random Forest methods are tested for performance based on accuracy, sensitivity, and specificity using the confusion matrix table. The accuracy, sensitivity, and specificity formulas use Equations (1), (2), and (3), respectively [17] [13].

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

$$Sensitivity = \frac{TP}{TP+FN} \quad (2)$$

$$Specificity = \frac{TN}{TN+FP} \quad (3)$$

III. Result and Discussion

This section contains the results that have been achieved at each stage. The first stage is the collection of diabetes disease datasets obtained from the Uci Repository with a total of 768 instances and ten attributes. After data collection, the next step is data preprocessing. In the data preprocessing stage, it is used to improve data quality in diabetes disease data to optimize the classification method's performance.

There are unbalanced data in the data used so that it can reduce the performance of the classification method. The number of negative classes is 500 instances (majority class), and positive classes are 268 instances (minority class). This research proposes several sampling methods to balance the data: the SMOTE, ENN, and SMOTE-ENN hybrid. The amount of data generated by each sampling method is shown in Table I.

TABLE I. Data Distribution before and after Sampling

Sampling Method	*Positive Class*	*Negative Class*
Original Data	268	500
SMOTE	500	500
ENN	268	240
SMOTE-ENN	303	227

In Table I, the SMOTE method produces a balanced class by adding the minority class so that the number equals the majority class. However, the SMOTE method has the disadvantage of producing noise in the new data generated. The Edited Nearest Neighbor (ENN) method balances the data by removing the majority class (positive class) adjacent to the minority class so that it can reduce data noise in the dataset, while the SMOTE-ENN method makes the data balanced by combining the SMOTE and ENN methods. The SMOTE method is used to add new data to the minority class based on the nearest neighbor. After the SMOTE results are balanced, the removal of adjacent data between the majority and minority classes is carried out to minimize data noise.

The data balanced using the sampling method is then divided into training and testing data using 10-fold cross-validation. Diabetes data is divided into training and testing, then implementing Random Forest and SVM classification methods for diabetes prediction. The classification results of the SVM and Random Forest methods are tested for performance based on accuracy, sensitivity, and specificity using the confusion matrix table. The confusion matrix results are obtained using the SVM method with original data (see Fig. 4), SMOTE result data (see Fig. 5), ENN method result data (see Fig. 6), and SMOTE-ENN data results (see Fig. 7).

Based on Fig. 4, the SVM method correctly classifies negative classes in as many as 438 instances, correctly classifies positive classes in as many as 151 instances, incorrectly classifies negative classes in as many as 62 instances, and incorrectly classifies positive classes in as many as 117 instances. The performance of the SVM method with the original data obtained an accuracy of 76.7%, a sensitivity of 56.3%, and a specificity of 87.6%.

Based on Fig. 5, the SVM method with SMOTE can correctly classify negative classes with 387 instances, correctly classify positive classes with 354 instances, incorrectly classify negative classes with 113 instances and incorrectly classify positive classes with 146 instances. The performance of the SVM method with SMOTE has an accuracy of 74.1%, sensitivity of 70.8%, and specificity of 77.4%.

Based on Fig. 6, the SVM method with ENN can correctly classify negative classes in as many as 207 instances, correctly classify positive classes in as many as 229 instances, incorrectly classify negative classes in as many as 33 instances and incorrectly classify positive classes in as many as 39 instances. The performance of the SVM method with ENN has an accuracy of 85.8%, sensitivity of 85.4%, and specificity of 86.3%.

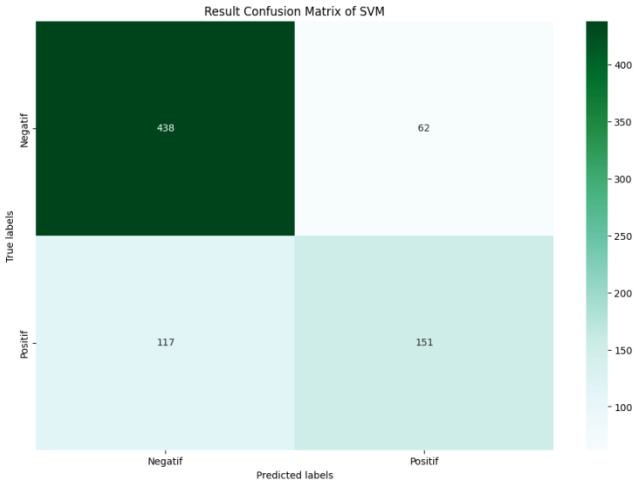

Fig. 4. SVM results with original data.

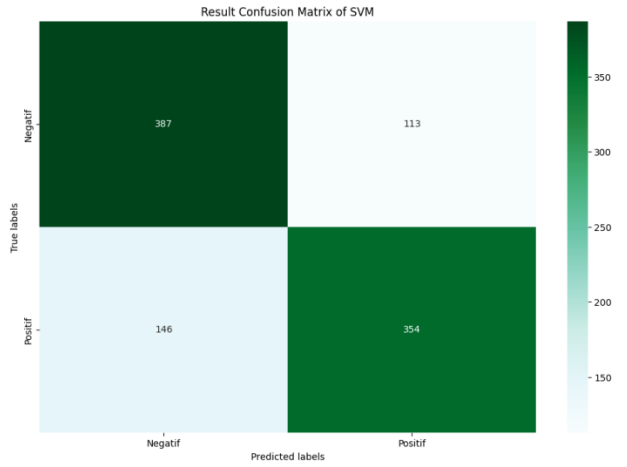

Fig. 5. SVM results with result data.

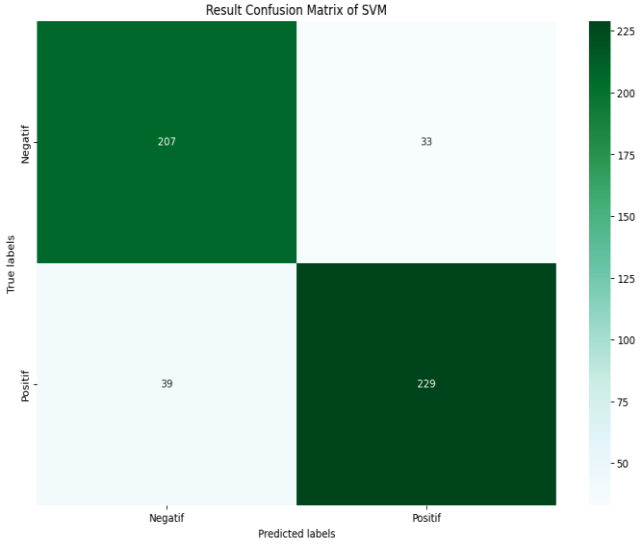

Fig. 6. SVM results with ENN result data.

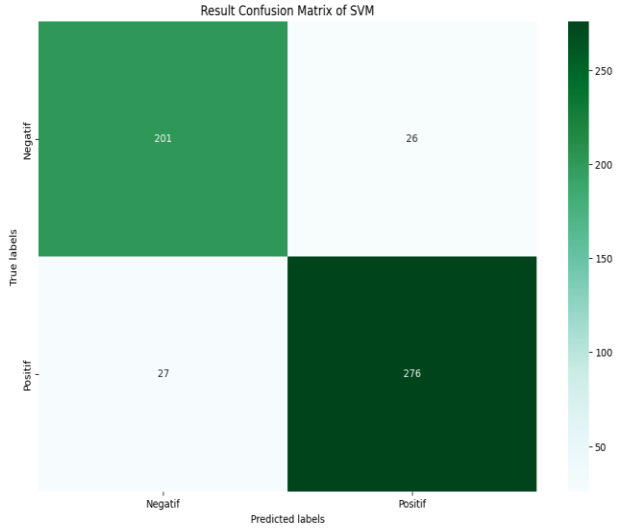

Fig. 7. SVM results with SMOTE-ENN result data.

Based on Fig. 7, the SVM method with SMOTE-ENN can classify negative classes correctly in as many as 201 instances, classify positive classes correctly in as many as 276 instances, classify negative classes incorrectly in as many as 26 instances, and classify positive classes incorrectly as many as 27 instances. The performance of the SVM method with SMOTE-ENN gets an accuracy of 90%, sensitivity of 91.1%, and specificity of 88.5%.

Then the confusion matrix results using the Random Forest method with original data (See Fig. 8), SMOTE data (See Fig. 9), ENN data (See Fig. 10), and SMOTE-ENN data (See Fig. 11).

Based on Fig. 8, the Random Forest method correctly classifies negative classes in as many as 429 instances, correctly classifies positive classes in as many as 156 instances, incorrectly classifies negative classes in as many as 72 instances, and incorrectly classifies positive classes in as many as 112 instances. The performance of the Random Forest method with the original data obtained an accuracy of 76.1%, sensitivity of 58.2%, and specificity of 85.8%.

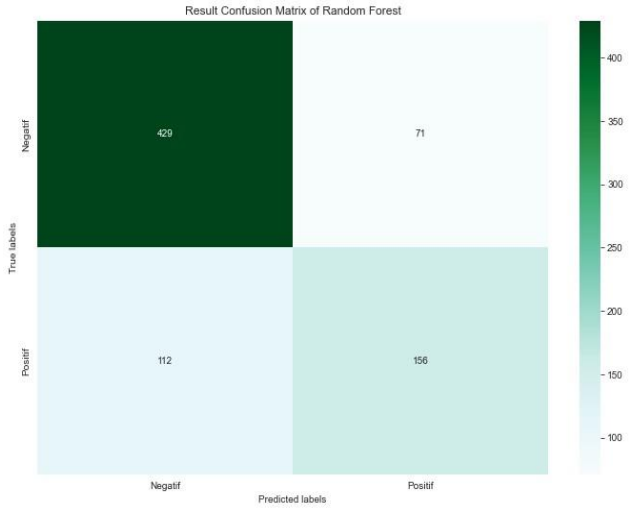

Fig. 8. Random forest results with original data.

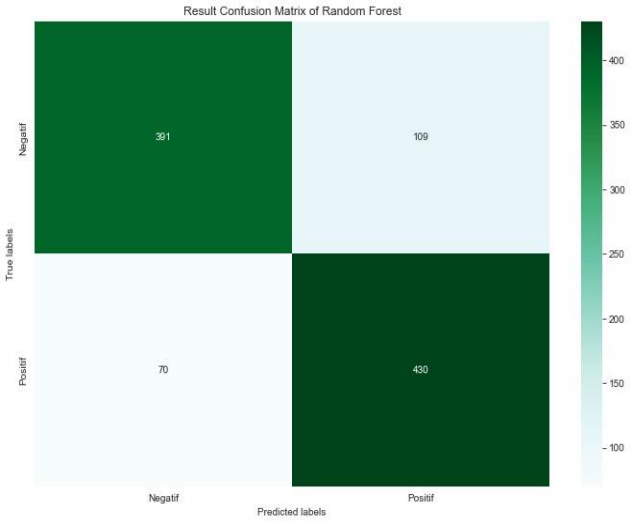

Fig. 9. Random forest results with SMOTE.

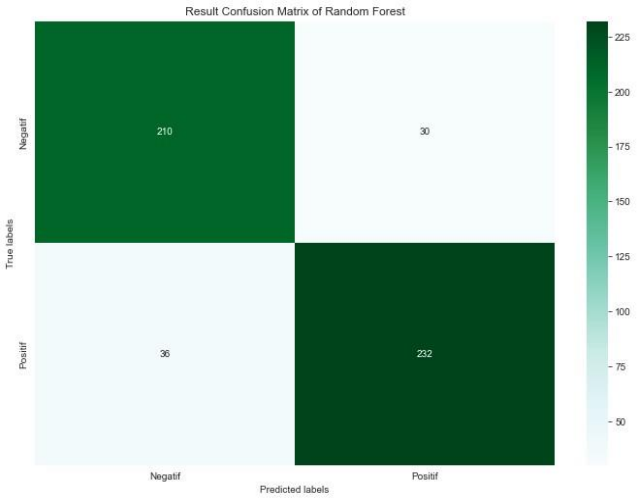

Fig. 10. Random forest results with ENN.

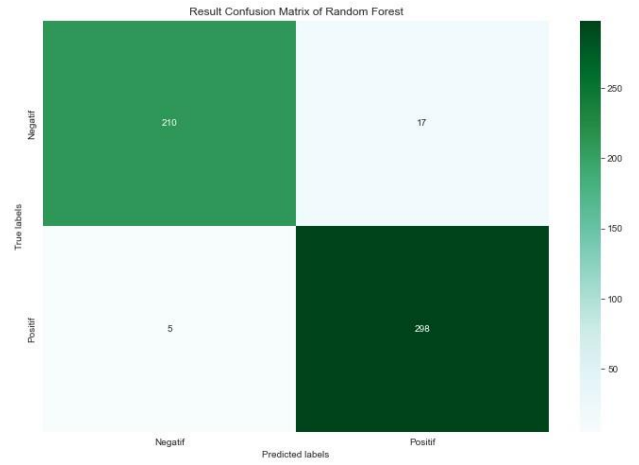

Fig. 11. Random forest results with SMOTE-ENN.

Based on Fig. 9, the Random Forest method with SMOTE can correctly classify negative classes in as many as 391 instances, correctly classify positive classes in as many as 430 instances, incorrectly classify negative classes in as many as 109 instances and incorrectly classify positive classes in as many as 70 instances. The performance of the Random Forest method with SMOTE gets an accuracy of 82.1%, sensitivity of 86%, and specificity of 78.2%.

Based on Fig. 10, the Random Forest method with ENN can correctly classify negative classes in as many as 210 instances, correctly classify positive classes in as many as 232 instances, incorrectly classify negative classes in as many as 30 instances and incorrectly classify positive classes in as many as 36 instances. The performance of the Random Forest method with ENN has an accuracy of 87%, sensitivity of 86.6%, and specificity of 87.5%.

Based on Fig. 11, the Random Forest method with SMOTE-ENN can correctly classify negative classes in as many as 210 instances, correctly classify positive classes in as many as 298 instances, incorrectly classify negative classes in as many as 17 instances and incorrectly classify positive classes as many as 5 instances. The performance of the Random Forest method with SMOTE-ENN gets an accuracy of 95.8%, sensitivity of 98.3%, and specificity of 92.5%.

The results can be seen in Table II to simplify the understanding of the research results achieved based on several experiments that have been carried out.

TABLE II. CLASSIFICATION METHOD PERFORMANCE RESULTS WITH DATA SAMPLING APPROACH

	Accuracy	***Sensitivity***	***Specificity***
SVM	76,7%	56,3%	87,6
SVM with SMOTE	74,1%	70,8%	77,4%
SVM with ENN	85,8%	85,4%	86,3%
SVM with SMOTE-ENN	90%	91,1%	88,5%
Random Forest	76,1%	58,2%	85,8%
Random Forest with SMOTE	82,1%	86%	78,2%
Random Forest with ENN	87%	86,6%	87,5%
Random Forest with SMOTE-ENN	**95,8%**	**98,3%**	**92,5%**

Based on Table II, the Random Forest method with SMOTE-ENN produces the highest performance compared to SVM, with an accuracy of 95.8%, sensitivity of 98.3%, and specificity of 92.5%. Furthermore, the approach using the SMOTE-ENN sampling method resulted in better average performance than the SMOTE and ENN methods separately, such as accuracy, sensitivity, and specificity. The SMOTE-ENN sampling method is better than SMOTE and ENN separately because it can minimize noise data in the artificial data produced. The noise data in this context is the minority class data that is close to the majority class, so the classification method makes biased predictions. Besides that, using hybrid sampling by combining oversampling and undersampling methods in solving unbalanced data performs better than oversampling without undersampling [18]. SMOTE-ENN hybrid sampling in this study can significantly improve the sensitivity performance [19][20]. In order to see that the method proposed in this study is better than some related previous studies, the following comparison of the results can be seen in Table III.

TABLE III. RESULTS COMPARISON WITH PREVIOUS RESEARCH

Previous Studies	**Methods**	**Scope of Study**	**Accuracy**
Hairani et al. [13]	Random Forest and SMOTE-Tomek Links	Diabetes Disease	86,4%
ElSeddawy, et al. [14]	ANN + Gridsearch + SMOTE		92%
Sabhita et al. [21]	SVM + RFE + SMOTE		82%
Abdullah, et al. [22]	Random Forest + SMOTE		83%
Ijaz et al. [23]	DBSCAN + SMOTE + Random Forest		83,6%
Butt et al. [24]	LSTM		87,3%
Proposed Method	Random Forest and SMOTE-ENN		**95,8%**

IV. CONCLUSION

Based on the research results obtained using the SVM and Random Forest methods combined with the SMOTE, ENN, and hybrid SMOTE-ENN sampling methods, the Random Forest method with SMOTE-ENN produces better performance than the SVM method based on an accuracy of 95.8%, sensitivity of 98.3%, and specificity of 92.5% in diabetes prediction. Moreover, it can also be concluded that the SMOTE-ENN sampling method produces better performance than the SMOTE method without ENN in the results of the classification method used. Future researchers are suggested to use the ensemble learning method to improve the performance of the classification method.

ACKNOWLEDGMENT

Thanks to DRTPM of the Ministry of Education, Culture, Research, and Technology for funding under the Regular Fundamental Research scheme in 2023.

REFERENCES

[1] H. Hairan, 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," *J. Teknol. dan Sist. Komput.*, vol. 8, no. 2, pp. 89–93, Apr. 2020, doi: 10.14710/jtsiskom.8.2.2020.89-93.

[2] E. Erlin, Y. Desnelita, N. Nasution, L. Suryati, and F. Zoromi, "Impact ofSMOTE on Random Forest Classifier Performance based on Imbalanced Data," *MATRIK J. Manajemen, Tek. Inform. dan Rekayasa Komput.*, vol. 21, no. 3, pp. 677–690, 2022, doi: 10.30812/matrik.v21i3.1726.

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

[4] K. Rajendran, M. Jayabalan, and V. Thiruchelvam, "Predicting breast cancer via supervised machine learning methods on class imbalanced data," *Int. J. Adv. Comput. Sci. Appl.*, vol. 11, no. 8, pp. 54–63, 2020, doi: 10.14569/IJACSA.2020.0110808.

[5] Erlin, Y. N. Marlim, Junadhi, L. Suryati, and N. Agustina, "Early Detection of Diabetes Using Machine Learning with Logistic Regression Algorithm," *J. Nas. Tek. Elektro dan Teknol. Inf.*, vol. 11, no. 2, pp. 88–96, 2022.

[6] Hairani, M. N. Abdillah, and M. Innuddin, "An Expert System for Diagnosis of Rheumatic Disease Types Using Forward Chaining Inference and Certainty Factor Method," in *2019 International Conference on Sustainable Information Engineering and Technology (SIET)*, 2019, pp. 104–109. doi: 10.1109/SIET48054.2019.8986035.

[7] J. Yang and J. Guan, "A Heart Disease Prediction Model Based on Feature Optimization and Smote-Xgboost Algorithm," *Information*, vol. 13, no. 10, pp. 1–15, Oct. 2022, doi: 10.3390/info13100475.

[8] M. Kumari and N. Subbarao, "A hybrid resampling algorithms SMOTE and ENN based deep learning models for identification of Marburg virus inhibitors," *Future Med. Chem.*, vol. 14, no. 10, pp. 701–715, Apr. 2022, doi: 10.4155/fmc-2021-0290.

[9] 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," *J. Biomed. Inform.*, vol. 107, no. June, p. 103465, 2020, doi: 10.1016/j.jbi.2020.103465.

[10] R. Islam, A. Sultana, M. N. Tuhin, M. S. H. Saikat, and M. R. Islam, "Clinical Decision Support System for Diabetic Patients by Predicting Type 2 Diabetes Using Machine Learning Algorithms," *J. Healthc. Eng.*, vol. 2023, pp. 1–11, May 2023, doi: 10.1155/2023/6992441.

[11] D. Sumathi, "Implementing a Model to Detect Diabetes Prediction using Machine Learning Implementing a Model to Detect Diabetes Prediction using Machine Learning Classifiers," *J. Algebr. Stat.*, vol. 13, no. 1, pp. 558–566, 2022.

[12] R. Saxena, S. K. Sharma, M. Gupta, and G. C. Sampada, "A Novel Approach for Feature Selection and Classification of Diabetes Mellitus: Machine Learning Methods," *Comput. Intell. Neurosci.*, vol. 2022, pp. 1–11, Apr. 2022, doi: 10.1155/2022/3820360.

[13] H. Hairani, A. Anggrawan, and D. Priyanto, "Improvement Performance of the Random Forest Method on Unbalanced Diabetes Data Classification Using Smote-Tomek Link," *Int. J. Informatics Vis.*, vol. 7, no. 1, pp. 258–264, 2023.

[14] A. I. ElSeddawy, F. K. Karim, A. M. Hussein, and D. S. Khafaga, "Predictive Analysis of Diabetes-Risk with Class Imbalance," *Comput. Intell. Neurosci.*, vol. 2022, pp. 1–16, Oct. 2022, doi: 10.1155/2022/3078025.

[15] U. Ependi, A. F. Rochim, and A. Wibowo, "A Hybrid Sampling Approach for Improving the Classification of Imbalanced Data Using ROS and NCL Methods," *Int. J. Intell. Eng. Syst.*, vol. 16, no. 3, pp. 345–361, 2023, doi: 10.22266/ijies2023.0630.28.

[16] T. Sasada, Z. Liu, T. Baba, K. Hatano, and Y. Kimura, "A resampling method for imbalanced datasets considering noise and overlap," in *Procedia Computer Science*, 2020, vol. 176, pp. 420–429. doi: 10.1016/j.procs.2020.08.043.

[17] 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.

[18] W. Nugraha, R. Maulana, Latifah, P. A. Rahayuningsih, and Nurmalasari, "Over-sampling strategies with data cleaning for handling imbalanced problems for diabetes prediction," *AIP Conf. Proc.*, vol. 2714, no. 1, p. 30017, May 2023, doi: 10.1063/5.0128407.

[19] S. Balasubramanian, R. Kashyap, S. T. CVN, and M. Anuradha, "Hybrid Prediction Model For Type-2 Diabetes With Class Imbalance," in *2020 IEEE International Conference on Machine Learning and Applied Network Technologies (ICMLANT)*, 2020, pp. 1–6. doi: 10.1109/ICMLANT50963.2020.9355975.

[20] A. Anggrawan, H. Hairani, and C. Satria, "Improving SVM Classification Performance on Unbalanced Student Graduation Time Data Using SMOTE," *Int. J. Inf. Educ. Technol.*, vol. 13, no. 2, pp. 289–295, 2023, doi: 10.18178/ijiet.2023.13.2.1806.

[21] E. Sabitha and M. Durgadevi, "Improving the Diabetes Diagnosis Prediction Rate Using Data Preprocessing, Data Augmentation and Recursive Feature Elimination Method," *Int. J. Adv. Comput. Sci. Appl.*, vol. 13, no. 9, pp. 921–930, 2022, doi: 10.14569/IJACSA.2022.01309107.

[22] M. N. Abdullah and Y. B. Wah, "Improving Diabetes Mellitus Prediction with MICE and SMOTE for Imbalanced Data," in *2022 3rd International Conference on Artificial Intelligence and Data Sciences (AiDAS)*, 2022, pp. 209–214. doi: 10.1109/AiDAS56890.2022.9918773.

[23] M. F. Ijaz, G. Alfian, M. Syafrudin, and J. Rhee, "Hybrid Prediction Model for type 2 diabetes and hypertension using DBSCAN-based outlier detection, Synthetic Minority Over Sampling Technique (SMOTE), and random forest," *Appl. Sci.*, vol. 8, no. 8, 2018, doi: 10.3390/app8081325.

[24] U. M. Butt, S. Letchmunan, M. Ali, F. H. Hassan, A. Baqir, and H. H. R. Sherazi, "Machine Learning Based Diabetes Classification and Prediction for Healthcare Applications," *J. Healthc. Eng.*, vol. 2021, pp. 1–17, Sep. 2021, doi: 10.1155/2021/9930985.