Scientific Journal of Informatics
Vol. 8, No. 2, November 2021
p-ISSN 2407-7658 http://journal.unnes.ac.id/nju/index.php/sji e-ISSN 2460-0040

Comparative Analysis of ADASYN-SVM and SMOTE-SVM Methods on the Detection of Type 2 Diabetes Mellitus

Nur Ghaniaviyanto Ramadhan[1*]
[*]Software Engineering Department, Faculty of Informatics, Institut Teknologi Telkom Purwokerto, Indonesia

Abstract. Most people with diabetes in the world are type 2. We can detect diabetes early to prevent things that are not desirable by checking sugar and insulin levels with the doctor. In addition to using this method, people with diabetes can also be grouped based on data from diabetes examination results. However, most of the data on health examination results have several parameters that are difficult for the public to understand. These problems can be done by means of automatic classification. In addition to these problems, there is another problem in the form of an unbalanced amount of data for diabetics and non-diabetics. This problem can be done by balancing the amount of data using the model to increase the ratio of the amount of data that is small or decrease the ratio of the amount of data that is too much.
Purpose: This study aims to detect type 2 diabetes mellitus using the SVM classification model and analyze the results of the comparison using the SMOTE and ADASYN data balancing technique which is the best.
Methods/Study design/approach: The research method starts from collecting the diabetes dataset, then the dataset cleaning process is carried out whether there is a null value or not. After applying two oversampling methods to analyze which method is the most appropriate. After the oversampling technique was carried out, data classification was carried out using a support vector machine model to see the accuracy results**.**
Result/Findings: The results obtained by the ADASYN-SVM method are superior to SMOTE-SVM. The ADASYN-SVM method has an accuracy of 87.3%, while the SMOTE-SVM has an accuracy of 85.4%.
Novelty/Originality/Value: The data used in this study came from the Karya Medika clinic, Indonesia which contains parameters related to type 2 diabetes.

Keywords: Diabetes Type 2, Classification, SVM, ADASYN, SMOTE

Received October 2021 / **Revised** October 2021 / **Accepted** November 2021

INTRODUCTION

Diabetes mellitus is a chronic metabolic disorder caused by the pancreas not producing enough insulin or the body cannot use insulin effectively [1]. Diabetes mellitus has several types, namely type 1, type 2, and gestational type [2]. Most people with diabetes in the world are type 2, which is the cause of an unhealthy lifestyle [3]. Diabetes is included in the 7 most deadly diseases [4]. Even according to [5] in 2000-2016 there was an increase of 5% of people with diabetes and in 2019 as many as 1.5 million people died from this disease. This disease if not prevented or treated can lead to complications of various other deadly diseases such as stroke and even kidney failure [1].

Diabetes can detect early to prevent unwanted things by checking sugar and insulin levels with the doctor [6]. In addition to using this method, people with diabetes can also be grouped based on data from diabetes examination results. The data certainly contains information that is very important for a person to find out whether he has diabetes or not. However, most of the data on health examination results have several parameters that are difficult for the public to understand. These problems can be done through automatic classification. In addition to these problems, there is another problem in the form of an unbalanced amount of data for diabetics and non-diabetics. This problem can be done by balancing the amount of data using the model by increasing the ratio of the amount of data that is small or decreasing the ratio of the amount of data that is too much.

[*]Corresponding author.
Email address: ghani@ittelkom-pwt.ac.id (Ramadhan)
DOI: 10.15294/sji.v8i2.32484

Several studies that have been conducted can address the problem of balancing the amount of data, classification, or detection of type 2 diabetes based on available diabetes data. For example, paper [6] detects type 2 diabetes mellitus using the Random Forest (RF) classification model and balances diabetes data with random oversampling techniques. Study [7] used datasets from Luzhou, China, and for decision tree models, random forests, and neural networks to predict diabetes mellitus with the highest accuracy obtained 80.8% using RF. In research [8] to predict diabetes based on support vector machine (SVM) and Naïve Bayes models using a dataset from Kosovo., the accuracy of the SVM model is 95%.

In the study [9] discussing a diabetes prediction model using SMOTE and a decision tree, the SMOTE model here serves to remove imbalanced datasets, the dataset used comes from a laboratory in Kashmir. Paper [10] tells about comparing data mining models (decision tree, nave Bayes, and KNN) for diabetes detection with the highest accuracy by 75.65% decision tree, the dataset used by Pima Indian. Research [11] discusses diabetes prediction using several classification models (logistic regression, SVM, J48, and KNN) with the dataset used by Pima Indian.

In paper [12] discusses the prediction algorithm for diabetes mellitus on imbalanced data and missing value problems, the model used by ADASYN for imbalanced data and random forest for classification, the dataset used by Pima Indian. Study [13] discusses the approach to the problem of imbalanced data for the machine learning classification process, one example of the dataset used by Pima Indian. One of the imbalanced models used is ADASYN, while the classification model used is SVM, KNN, and neural network (NN) [13]. In research [14] discussing the preprocessing technique of balancing data to improve the performance of the KNN classification model, the dataset used comes from the UCI repository.

Another study [15] discusses how to improve the predictive outcome of diabetes mellitus by creating a model that can be used for many datasets, one of which is Pima Indian with its k-means classification model and logistic regression. The SMOTE, Bagging, SVM, and MLP models were also used to predict type 2 diabetes mellitus, with the dataset used from Kashmir [16]. Comparison of the Support Vector Machine and Modified Balanced Random Forest models can also find the best classification model for diabetes cases [17]. Research [18] to compare the performance of algorithms used to predict diabetes using data mining techniques. in this paper, we compare machine learning classifiers (J48 Decision Tree, K-Nearest Neighbors, and Random Forest, Support Vector Machine) for classifying diabetes mellitus patients. Study [19] the main objectives are (i) Gaussian process classification (GPC), (ii) comparative classifier for diabetes data classification, (iii) data analysis using a cross-validation approach, (iv) interpretation of data analysis, and (v) comparing our method with others.

Based on the explanation of several previous studies above, there are still few that discuss the problem of imbalanced datasets. Only a few papers discuss these problems such as [6], [9], [12], and [16]. Thus, this study aims to solve the imbalanced dataset problem by using two oversampling techniques, namely ADASYN and SMOTE. In addition, this study will also detect type 2 diabetes mellitus using the SVM classification and analyze the results of applying two oversampling techniques. Table 1 is a comparison of the contribution of this study with previous studies.

Table 1. Comparison of Contributions

Authors	Imbalanced Method	Classification Method
[6]	Random Oversampling	Random Forest
[9]	SMOTE	Decision Tree
[11]	-	SVM
[12]	ADASYN	Random Forest
[16]	SMOTE	SVM
[17]	-	SVM
This Study	SMOTE, ADASYN	SVM

METHODS

Figure 1 is a schematic diagram in this study.

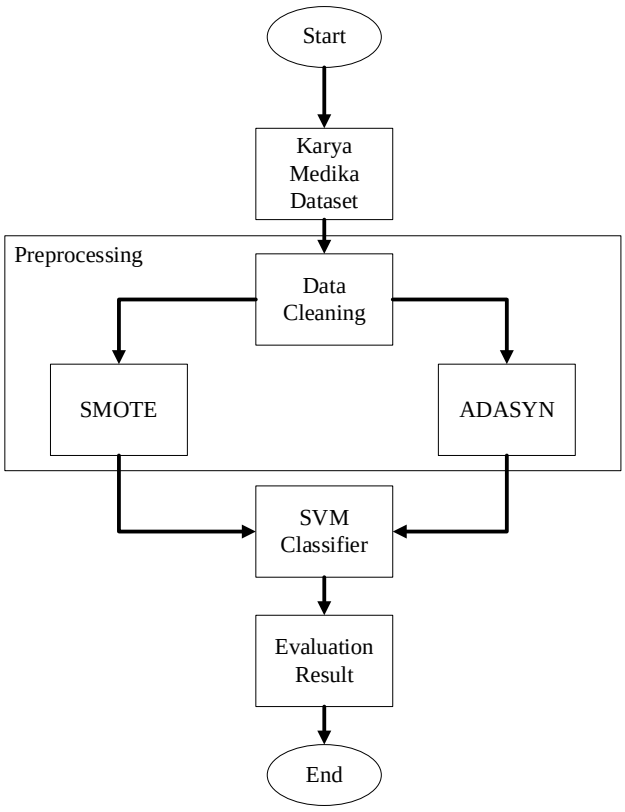

Figure 1. Design System

Dataset

This study uses a dataset originating from the Karya Medika laboratory, Indonesia [6]. The total data consists of 630 rows with 9 features, 290 diabetics, and 340 non-diabetics. Table 2 shows the characteristics of the dataset.

Table 2. Characteristics of Karya Medika Dataset

No	Attribute Description	Unit	Attribute Type
1	Glucose	mg/dl	Numeric
2	Male (1) or Female (0)	-	Nominal
3	Systolic Blood Pressure	mmHg	Numeric
4	Diastolic Blood Pressure	mmHg	Numeric
5	Height	Kg	Numeric
6	Weight	cm	Numeric
7	Age	Year	Numeric
8	Fasting (1) or No-Fasting (0)	-	Nominal
9	Diabetes (1) or No-Diabetes (0)	-	Nominal

Preprocessing

At this stage, data cleaning will be carried out and balance the number of diabetics with non-diabetics.

Data Cleaning

The initial stage of the dataset will be checking for each row and feature whether there are values that are null or unfilled. If there is a null value, can do it in two ways, namely by removing a row that contains a

null value or replacing a null value based on the mean, median, or other statistical values in the same data class. In the Karya Medika dataset, there are several attributes whose rows contain null values, table 3 is the representation.

Table 3. Null Values

Features	% Null Values
Systolic Blood Pressure	10.63%
Diastolic Blood Pressure	10.63%
Height	8%
Weight	7.46%
Age	4.4%

Null values are empty values, {}, NaN in the data rows (Figure 2). It can be seen in the table above that there are 5 features whose rows have a null value with an average percentage of 8.2%. This of course requires special attention so that the classification process produces the best accuracy. This study will apply a statistical model of the median value to replace null values in each row in each feature.

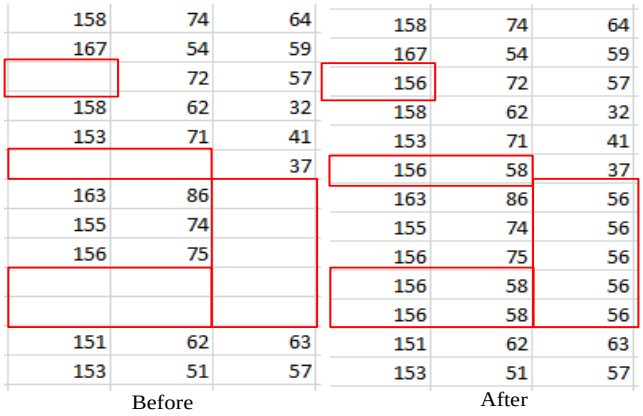

Figure 2. Replacement Null Values

Balancing Data

At this stage, after the dataset used has been cleaned, it will be carried out balancing the data classes using two oversampling models SMOTE and ADASYN.

SMOTE (Synthetic Minority Oversampling Technique)

SMOTE is an oversampling technique used to avoid classifier performance degradation caused by a class imbalance in the dataset [20]. SMOTE works by creating new instances of minority classes "synthetically". The general SMOTE algorithm is as follows [16].

$$Ynew = Y0 + (Y0i - Y0)x\partial \quad (1)$$

Where *Ynew* is a new synthetic sample. *Y0* is the feature vector of each instance in the minority class. *Y0i* is the *i-th* neighbor selected from *Y0*. $x\partial$ represents a random number between 0 and 1.

ADASYN (Adaptive Synthetic Sampling Method)

The ADASYN method is an adaptive data generation method that can generate samples adaptively to reduce class imbalances from the data set [21]. The steps of the ADASYN method are as follows [12].

1. Evaluate the degree of imbalance of all classes, $d = m0/m1, d \in (0,1)$.
2. Calculate the number of samples to be produced $G = (m1 - m0) \times \beta, \beta \in [0,1]$ represents the expected level of unbalance after data generation. If $\beta = 1$, means that the class sample is fully balanced after data generation.
3. For each sample from the minor class, find the deepest K in the n-dimensional space. Calculate $Ii = \Delta i / k \ (i = 1,2,..,m)$. $Ii \in [0,1]$ where is the sample number that belongs to the main class and also the k-nearest neighbor of *xi*.
4. Regularize *Ii* according to $Ii = Ii / \sum_i^m Ii$, then *Ii* is the probability distribution, and $\sum Ii = 1$.

5. Count the number of samples xi in the minor class to produce $gi = Ii \times G$.
6. Choose a sample of the k-nearest neighbors from in the small class. Synthesize a new sample S_z, where $S_z = (xj + xi) \times \lambda$, $\lambda \in [0, 1]$ is a random number.
7. Repeat step 6, *gi* times to get the sample *gi xi*.

Support Vector Machine (SVM)

Support Vector Machine (SVM) is a data mining model used for supervised learning in which data is classified and analyzed linearly [13, 22]. The SVM model represents the examples as points in the mapped space so that the examples from separate categories are divided by the widest possible clear gap [11]. Then the new examples are mapped into the same space and predicted to fall into categories based on which side of the gap they fall on [11].

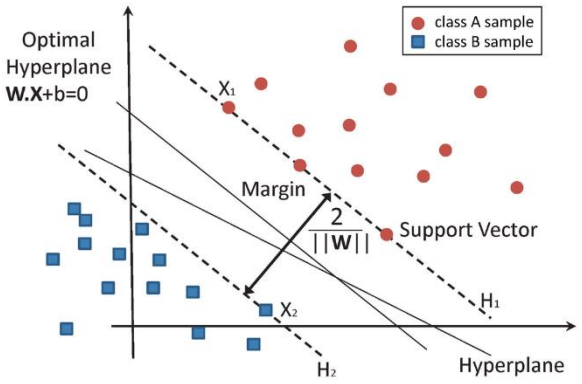

Figure 3. SVM Classification [23]

RESULT AND DISCUSSION

The analysis will be carried out based on the accuracy results obtained from experiments on the SMOTE-SVM and ADASYN-SVM methods. Table 4 is a confusion matrix used to calculate accuracy.

Table 4. Confusion Matrix

	Actual True	Actual False
Predicted True	TP	FP
Predicted False	FN	TN

True Positive (TP) is the amount of data that is predicted to be true and true. False Positive (FP) is the amount of data that is predicted to be true but is false. True Negative (TN) is the amount of data that is predicted to be false and is false. False Negative (FN) is the amount of data that is predicted to be false but is true.

Figure 4 is the confusion matrix of the SMOTE-SVM method.

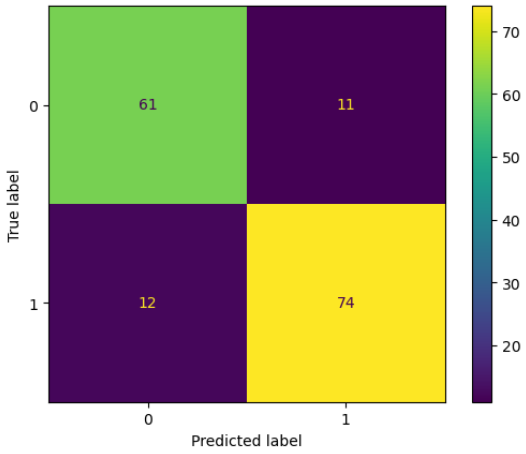

Figure 4. Confusion Matrix SMOTE-SVM

Figure 5 is the confusion matrix of the ADASYN -SVM method.

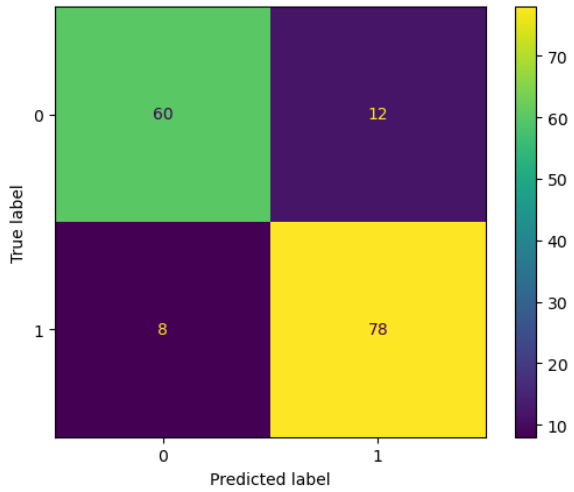

Figure 5. Confusion Matrix ADASYN-SVM

After knowing the value of each confusion matrix in table 4, then the formula for calculating accuracy (1) [17] is then entered.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (2)$$

Table 5 is a comparison of the accuracy produced by SMOTE-SVM and ADASYN-SVM.

Table 5. Classification Result

Methods	Accuracy %
SMOTE-SVM	85.4
ADASYN-SVM	87.3
SVM	83%

As seen from the accuracy results, the ADASYN-SVM method is superior to SMOTE-SVM. Several things make the ADASYN-SVM method superior, such as looking at the confusion matrix results (Figure 2 and Figure 3). The TP and TN values of the ADASYN-SVM method are higher than that of SMOTE-SVM. The value of TP and TN here means that they have the most significant influence on the resulting accuracy. The SMOTE-SVM model in predicting the FN value also has more errors. This study also proves that the problem of unbalanced data will cause lower diabetes detection results (Table 5).

CONCLUSION

Based on the results and analysis that has been carried out in this study, the ADASYN-SVM model is superior to use in diabetes detection problems with unbalanced data classes. Problems with unbalanced data are also important for processing in preprocessing. This study proves that the application of the model to balance the data can increase the accuracy results by 2-4% than using the SVM classification model.

REFERENCES

[1] Khairani. Info Datin (Pusat Data dan Informasi Kementrian Kesehatan Republik Indonesia). Hari diabetes sedunia. [Accessed December 12, 2021]. PDF article, https://pusdatin.kemkes.go.id/download.php?file=download/pusdatin/infodatin/infodatin-Diabetes-2018.pdf, 2018. (In Indonesian)

[2] S. Pangribowo, et.al. Info Datin (Pusat Data dan Informasi Kementrian Kesehatan Republik Indonesia). Tetap produktif, cegah, dan atasi diabetes melitus. [Accessed December 12, 2021]. PDF article, https://pusdatin.kemkes.go.id/download.php?file=download/pusdatin/infodatin/Infodatin-2020-Diabetes-Melitus.pdf. 2020. (In Indonesian)

[3] Ministry of Health RI. “Cegah, cegah, dan cegah: suara dunia perangi diabetes. [Accessed December 12, 2021]. https://www.kemkes.go.id/article/view/18121200001/prevent-prevent-and-prevent-the-voiceof-the-world-fight-diabetes.html. 2018. (In Indonesian)
[4] World Health Organization (WHO). Diabetes country profiles. [Accessed December 12, 2021]. PDF article, https://www.who.int/diabetes/country-profiles/idn_en.pdf. 2021.
[5] https://www.who.int/news-room/fact-sheets/detail/diabetes, Accessed December 12, 2021.
[6] N. G. Ramadhan, Adiwijaya, and A. Romadhony. “Preprocessing handling to enhance detection of type 2 diabetes mellitus based on random forest,” *Int. J. Adv. Comput. Sci. Appl.*, 12(7), pp. 223-228. 2021.
[7] Q. Zou, K. Qu, Y. Luo, D. Yin, Y. Ju, & H. Tang, “Predicting diabetes mellitus with machine learning techniques,” *Front. Genet*, 9, 515, 2018.
[8] Z. Tafa, N. Pervetica, & B. Karahoda, “An intelligent system for diabetes prediction,” In *2015 4th Mediterr. Conf. Embed. Comput. (MECO)* (pp. 378-382), IEEE, June 2015.
[9] S. Mirza, S. Mittal, & M. Zaman, “Decision support predictive model for prognosis of diabetes using SMOTE and decision tree,” *Int. J. Appl. Eng. Res.*, 13(11), pp. 9277-9282, 2018.
[10] A. Azrar, Y. Ali, M. Awais, & K. Zaheer, “ Data mining models comparison for diabetes prediction,” *Int. J. Adv. Comput. Sci. Appl.*, 9(8), pp. 320-323, 2018.
[11] S. Saru, & S. Subashree, “Analysis and prediction of diabetes using machine learning,” *Int. J. Emerg. Technol. Innov. Eng.*, 5(4). 2019.
[12] Q. Wang, W. Cao, J. Guo, J. Ren, Y. Cheng, & D. N. Davis, “DMP_MI: an effective diabetes mellitus classification algorithm on imbalanced data with missing values,” *IEEE Access*, 7, 102232-102238, 2019.
[13] Tyagi, Shivani, and S. Mittal. "Sampling approaches for imbalanced data classification problem in machine learning." *Proc. of ICRIC 2019*. Springer, Cham, pp. 209-221, 2020.
[14] Z. Shi, “Improving k-nearest neighbors algorithm for imbalanced data classification,” In *IOP Conf. Ser.: Mater. Sci. Eng.* (Vol. 719, No. 1, p. 012072), IOP Publishing, 2020.
[15] H. Wu, S. Yang, Z. Huang, J. He, & X. Wang, “Type 2 diabetes mellitus prediction model based on data mining. *Inform. Med. Unlocked*, 10, pp. 100-107, 2018.
[16] M. Shuja, S. Mittal, & M. Zaman, “Effective prediction of type ii diabetes mellitus using data mining classifiers and SMOTE,” In *Adv. Comput. Intell. Syst.*, (pp. 195-211), Springer, Singapore, 2020.
[17] M. D. Purbolaksono, M. I. Tantowi, A. I. Hidayat, & A. Adiwijaya, “ Perbandingan support vector machine dan modified balanced random forest dalam deteksi pasien penyakit diabetes,” *J. RESTI (Rekayasa Sist. Dan Teknol. Inform.*, 5(2), pp. 393-399, 2021.
[18] J. P. Kandhasamy and S. Balamurali. "Performance analysis of classifier models to predict diabetes mellitus," *Procedia Comput. Sci*., 47, pp. 45-51, 2015.
[19] M. Md, et al. "Comparative approaches for classification of diabetes mellitus data: Machine learning paradigm," *Comput. Methods Programs Biomed*, 152, pp. 23-34, 2017.
[20] N. V. Chawla, K. W. Bowyer, L. O. Hall, & W. P. Kegelmeyer, “SMOTE: synthetic minority over-sampling technique,” *J. Artif. Intell. Res.*, 16, pp. 321-357, 2002.
[21] H. He, Y. Bai, E. A. Garcia, & S. Li, “ADASYN: Adaptive synthetic sampling approach for imbalanced learning,” In *2008 IEEE Int. Jt. Conf. Neural. Netw. (IEEE World Congr. Comput. Intell.)* (pp. 1322-1328), IEEE, June 2008.
[22] I. Tomek, “Two modifications of CNN,” IEEE Trans. Syst. Man Cybern. 6, pp. 769–772. 1976.
[23] E. García-Gonzalo, Z. Fernández-Muñiz, P. J. García Nieto, A. Bernardo Sánchez, M. Menéndez & M. Fernández, “Hard-rock stability analysis for span design in entry-type excavations with learning classifiers,” *Materials*, 9(7), pp. 531, 2016.