Recursive Journal of Informatics
Vol. 1, No. 1, March 2023
p-ISSN 2963-5551 https://journal.unnes.ac.id/sju/index.php/rji e-ISSN XXXX-XXXX

Implementation of Synthetic Minority Oversampling Technique and Two-phase Mutation Grey Wolf Optimization on Early Diagnosis of Diabetes using K-Nearest Neighbors

Fathan Arsyadani[1], Aji Purwinarko[2]
[1, 2]Department Computer Science, Faculty of Mathematics and Natural Sciences,
Universitas Negeri Semarang, Indonesia

Abstract. Diabetes is a disease attacking the endocrine system characterized by high blood sugar levels. International Diabetes Federation (IDF) estimates that there were 451 million people with diabetes globally in 2017. Without treatment, this number is expected to rise to 693 million by 2045. One method for preventing increases in the number of diabetics is by early diagnosis. In an era where technology has developed rapidly, early diagnosis can be made with the machine learning method using classification. In this study, we propose a diabetes classification using K-Nearest Neighbors (KNN). Before classifying the data, we select the best feature subset from the dataset using Two-phase Mutation Grey Wolf Optimization (TMGWO) and balance the training data using Synthetic Minority Oversampling Technique (SMOTE). After dividing the dataset into training and testing sets using 10-fold cross validation, we reached an accuracy of 98.85% using the proposed method.
Purpose: This study aims to understand how to apply TMGWO and SMOTE to classify the early stage diabetes risk prediction dataset using KNN and how it affects the results.
Methods/Study design/approach: In this study, we use TMGWO to make a feature selection on the dataset, K-fold cross validation to split the dataset into training and testing sets, SMOTE to balance the training data, and KNN to perform the classification. The desired results in this study are accuracy, precision, recall, and f1-score.
Result/Findings: Performing classification using KNN with only features selected by TMGWO and balancing the training data using SMOTE gives an accuracy rate of 98.85%. From the results of this research, it can be concluded that the proposed algorithm can give higher accuracy compared to previous studies.
Novelty/Originality/Value: Implementing TMGWO to perform feature selection so the model can perform classification with fewer features and implementing SMOTE to balance the training data so the model can better classify the minority class. By doing classification using fewer features, the model can perform classification with a shorter computational time compared to using all features in the dataset.

Keywords: Data Mining, Diabetes, KNN, Machine Learning, SMOTE, TMGWO
Received January 03, 2023 / **Revised** January 19, 2023 / **Accepted** March 06, 2023

INTRODUCTION

Diabetes is a disease attacking the endocrine system and characterized by high level of blood sugar (Cole & Florez, 2020). Increased blood sugar levels are caused by a lack of insulin production by the pancreas or the inability of blood cells to use the insulin produced effectively (Gojka, 2016). According to IDF, it was estimated that there were 451 million diabetic people worldwide in 2017. Without treatment, this number can increase to 693 million in 2045. A reasonably high death rate also accompanies many people with diabetes. In 2017 alone, more than 50 million deaths worldwide were caused by diabetes [1]. The high mortality rate is because many sufferers are late to realize they have diabetes. Most diabetics only realize that they have diabetes after a compilation occurs without any initial treatment [2]. This problem will be resolved if people make an early diagnosis before complications between diabetes and other diseases that can increase the risk of death [3].

The rapid development of technology allows early diagnosis of a disease to be carried out by utilizing data mining methods [4]. In the medical field, especially in disease diagnosis, the data mining method commonly used is supervised learning with classification techniques [5]. Classification is used to diagnose a patient

[1]*Corresponding author.
Email addresses: fathan@students.unnes.ac.id (Arsyadani)
DOI: 10.15294/rji.v1i1.64406

based on the symptoms experienced [6]. One of the classification algorithms that can be used in diagnosing diseases that patients may suffer is K-Nearest Neighbors (KNN).

The problem that is often faced by most classification algorithms is when classifying datasets with an unbalanced amount of data between classes. An imbalanced data between classes will cause machine learning models to be biased toward the majority class and tend to misclassify the minority class [7]. Oversampling strategy can be used to tackle the imbalanced data problem. Oversampling is a strategy in which some artificial data will be generated to be added to the training data so that the data between classes become balanced [8]. Oversampling strategy can be done using the Synthetic Minority Oversampling Technique (SMOTE). In making artificial data, SMOTE works by selecting random samples from the minority class and then looking for the nearest neighbors of the selected samples [9].

Apart from data imbalance, another thing that often becomes a problem in classification is high-dimensional data. High-dimensional data has the potential to make the model difficult to classify, which can lead to low levels of accuracy or even overfitting [10]. To overcome the high dimensionality problem, performing feature selection using Two-phase Mutation Gray Wolf Optimization (TMGWO) can be the chosen solution. Feature selection works by selecting a subset of features that can better represent the entire population with minimal performance degradation [11]. TMGWO has been proven to have a great ability to perform feature selection by selecting fewer features and providing higher accuracy than other feature selection algorithms [12]. Previous studies have used either KNN, SMOTE, or TMGWO to classify the early stage diabetes risk prediction dataset [13]–[15]. This study aims to improve the accuracy produced by previous studies by combining TMGWO, SMOTE, and KNN.

METHODS

This research performs a classification on the early-stage diabetes risk prediction dataset using KNN. Before the classification process begins, TMGWO is used to perform a feature selection, so the classification process will not use all features from the dataset, but it will use only the selected features by TMGWO. After that, 10-fold cross validation is used to divide the dataset into training and testing sets. Before performing the classification, SMOTE is used to balance the training set so the model can better classify the minority class. This research is divided into three main steps, as explained below.

Data Collection

The early-stage diabetes risk prediction dataset collected from UCI Machine Learning Repository is used in this research. This dataset was collected from patients by using a direct questionnaire from the Sylhet Diabetes Hospital of Sylhet, Bangladesh in 2019 [16]. This dataset contains 520 instances. Among those 520 data, 320 belonged to the "Positive" class, and 200 belonged to the "Negative" class.

Table 1. Early stage diabetes risk prediction dataset

ID	Attributes	Values		Type
1	Age	[16, 25, 26, …79, 85, 90]		Continuous
2	Gender	Female: 0	Male: 1	Nominal
3	Polyuria	No: 0	Yes: 1	Nominal
4	Polydipsia	No: 0	Yes: 1	Nominal
5	Sudden weight loss	No: 0	Yes: 1	Nominal
6	Weakness	No: 0	Yes: 1	Nominal
7	Polyphagia	No: 0	Yes: 1	Nominal
8	Genital thrush	No: 0	Yes: 1	Nominal
9	Visual blurring	No: 0	Yes: 1	Nominal
10	Itching	No: 0	Yes: 1	Nominal
11	Irritability	No: 0	Yes: 1	Nominal
12	Delayed healing	No: 0	Yes: 1	Nominal
13	Partial paresis	No: 0	Yes: 1	Nominal
14	Muscle stiffness	No: 0	Yes: 1	Nominal
15	Alopecia	No: 0	Yes: 1	Nominal
16	Obesity	No: 0	Yes: 1	Nominal
17	Class	Negative: 0	Positive: 1	Nominal

This dataset consists of 17 attributes (16 features and one label), with the label being a column named "class" that consists of 2 categorical values: "Positive" (which later will be converted to 1) and "Negative" (which later will be converted to 0). Table 1 shows all attributes, values, and types of the dataset.

Data Preparation

Before the dataset is prepared and can be used for classification, several preprocessing steps are carried out. These steps include:

1. Changing categorical values to numeric (the categorical values and their numeric replacement are shown in Table 1)
2. Normalizing continuous data using a min-max scaler.
3. Performing feature selection using TMGWO.
4. Dividing the dataset into training and testing sets using 10-fold cross validation, and
5. Performing oversampling using SMOTE to balance class distribution in the training set.

Data Mining

After the data is fully prepared, the next step is to classify using KNN. The result of this classification process is a confusion matrix that can be used to measure the model's performance using precision, recall, f1-score, and accuracy. The flowchart of our proposed method is shown in Figure 1.

RESULT AND DISCUSSION

Data Collection

In this study, early-stage diabetes risk prediction dataset collected from UCI Machine Learning Repository is used for diabetes classification. This dataset consists of 520 instances, with 320 (61.5%) instances belonging to the "Positive" class and 200 (38.5%) instances belonging to the "Negative" class, and 17 attributes. All attributes and their type are shown in Table 1.

Data Preparation

Changing categorical features to numeric

Since almost all machine learning algorithms cannot handle string data and can only process numerical data, we need to change the categorical data into numeric. In the early stage diabetes risk prediction dataset, the categorical features are Polydipsia, Polyuria, Sudden weight loss, Weakness, Polyphagia, Genital thrush, Itching, Irritability, Gender, Delayed healing, Partial paresis, Muscle stiffness, Visual blurring, Obesity, Alopecia, and Class. The values of each categorical data and their replacements are shown in Table 1.

Normalize continuous features using min-max scaler

Min-max scaler will scale the data between the range 0 and 1 by dividing each value by its column's maximum and minimum values. The Equation for min-max scaling is shown in Equation 1. Where $X_{i_{scaled}}$ is the normalized value of X_i, X_i is the value that is going to be normalized, X_{min} and X_{max} are the minimum and maximum values of the current column, respectively.

$$X_{i_{scaled}} = \frac{(X_i - X_{min})}{(X_{max} - X_{\min})} \quad (1)$$

In the early stage diabetes risk prediction dataset, the continuous feature is "Age". Age's maximum and minimum values are shown in Table 2.

Table 2. Age's maximum and minimum values

Variable	Value
X_{min}	16
X_{max}	90

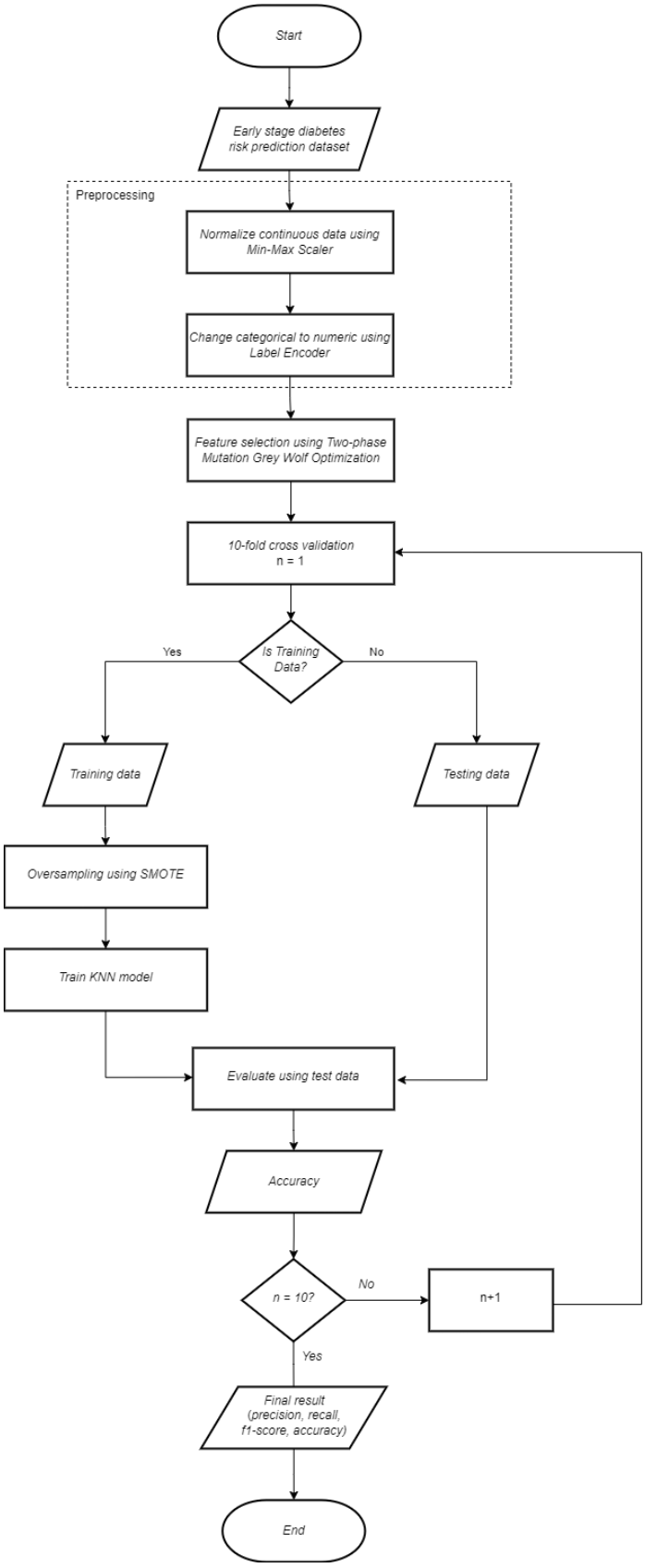

Figure 1. Flowchart of the proposed method

Feature selection using TMGWO
TMGWO is used to select the best subset of feature from all features in the dataset. Table 3 shows the parameters used in this study.

Table 3. Parameter setting of TMGWO

Parameter	Value
Number of wolves	8
Max Iterations	40
Number of dimensions	16
Mutation probability	0.07

The optimization process will stop when it reaches its stopping condition (in this case, the maximum number of iterations). A list of all features, their representation, and their status after 70 iterations are shown in Table 4.

Table 4. Status of each feature after TMGWO

ID	Attributes	Representation	Status
1	Age	1	Selected
2	Gender	1	Selected
3	Polyuria	1	Selected
4	Polydipsia	1	Selected
5	Sudden weight loss	0	X
6	Weakness	0	X
7	Polyphagia	0	X
8	Genital thrush	0	X
9	Visual blurring	1	Selected
10	Itching	1	Selected
11	Irritability	1	Selected
12	Delayed healing	0	X
13	Partial paresis	0	X
14	Muscle stiffness	1	Selected
15	Alopecia	1	Selected
16	Obesity	1	Selected

Dividing dataset into training and testing set using 10-fold cross validation
The dataset will be divided into ten equal folds using 10-fold cross validation, then one fold will be treated as testing data, and the other nine folds will be used as training data. This process will continue until each fold has become testing data, so the entire training process will be carried out ten times/iteration.

Oversampling using SMOTE.
SMOTE is used to balance the data between classes in the training set. The amount of data in each class in each iteration is shown in Table 5.

Table 5. Training data distribution in each iteration

Iteration-	"Positive" class	"Negative" class	Ratio (%)	Total
1	285	183	61:39	468
2	291	177	62:38	468
3	290	178	62:38	468
4	287	181	61.5:38.5	468
5	284	184	60.5:39.5	468
6	289	179	61.5:38.5	468
7	292	176	62.5:37.5	468
8	287	181	61.5:38.5	468
9	291	177	62:38	468
10	284	184	60.5:39.5	468

SMOTE will balance the data by creating synthetic data using a k-nearest neighbor. In this study, we use 3-nearest neighbors for our SMOTE. The amount of synthetic data generated for each iteration is shown in Table 6.

Table 6. Amount of synthetic data generated in each iteration

Iteration-	Amount of synthetic data
1	102
2	114
3	112
4	106
5	100
6	110
7	116
8	106
9	114
10	100

After being oversampled by SMOTE, the training data in each iteration will have 468 data with a 50:50 ratio.

Data Mining

The data mining process is divided into four stages:

1. Classification using KNN without applying TMGWO and SMOTE
2. Classification using KNN using only selected features by TMGWO.
3. Classification using KNN after applying SMOTE on training data, and
4. Classification using KNN using only selected features obtained from TMGWO after applying SMOTE on training data.

In our proposed algorithm, we use 5-nearest neighbors for the classification process. The confusion matrix of KNN, KNN using only selected features by TMGWO, KNN after balancing the training data using SMOTE, and KNN using only selected features by TMGWO and after balancing the training data using SMOTE are shown in Figures 2,3,4 and 5, respectively.

From the confusion matrix shown in Figures 2, 3, 4, and 5, applying either TMGWO to perform feature selection or SMOTE to balance the training data can improve the KNN model's ability. TMGWO can select 10 out of 16 features and still gives good accuracy. SMOTE can balance the training data and improve model's ability to classify data in the minority class (in this case, the "Negative" class). A decrease in the misclassification rate indicates these improvements. On classification with all 16 features, there were 16 misclassifications. This number was reduced to 15 on classification using only selected features by TMGWO. Before implementing SMOTE, there were three misclassifications in the minority class. This number decreases to only one misclassification after implementing SMOTE on training data.

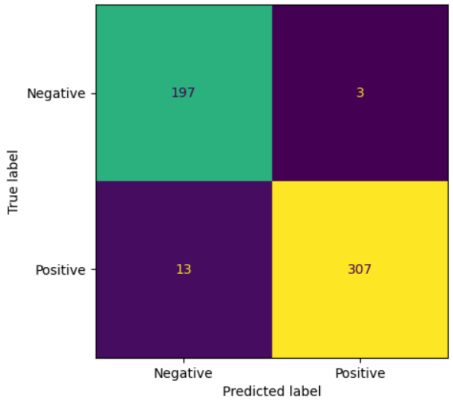

Figure 2. Confusion matrix of KNN

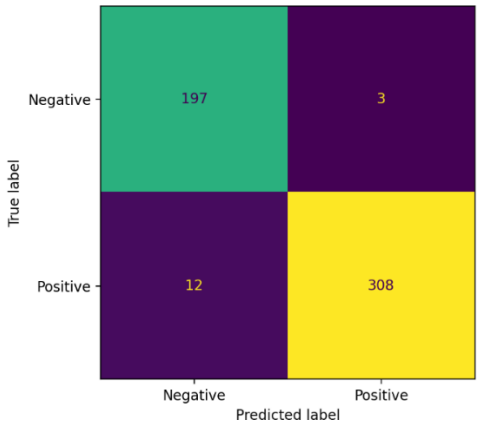

Figure 3. Confusion matrix of KNN+TMGWO

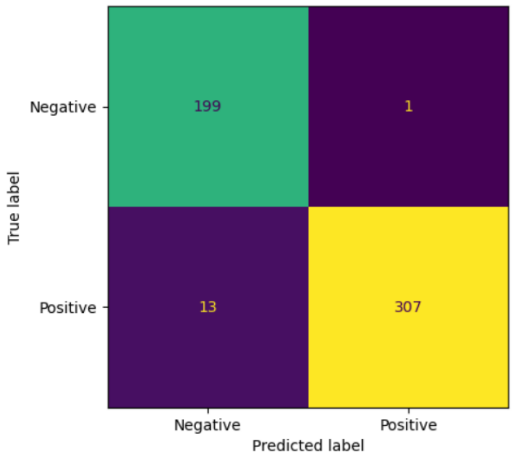

Figure 4. Confusion matrix of SMOTE+KNN

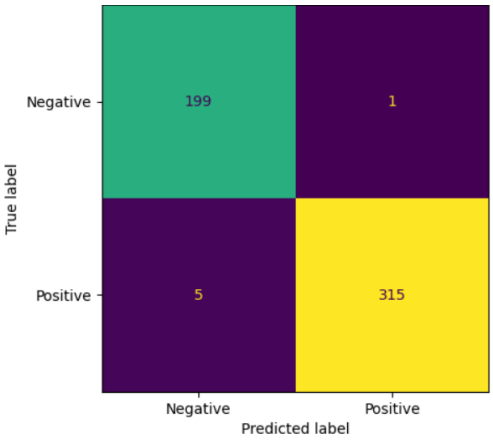

Figure 5. Confusion matrix of TMGWO+SMOTE+KNN

Furthermore, classification using only selected features by TMGWO and balancing the training data using SMOTE gives good results in the classification process. As shown in Figure 5, classification using those two methods give only six misclassifications, 5 in the "Positive" class and 1 in the "Negative" class. From the confusion matrix shown in Figures 2, 3, 4, and 5, we can obtain some metrics that can be used to measure the performance of our model in classifying the early stage diabetes risk prediction dataset. The classification result of each algorithm is shown in Table 7.

Table 7. Classification result of each algorithm (%)

Algorithm(s)	Precision		Recall		F1-score		Accuracy
	Positive	Negative	Positive	Negative	Positive	Negative	
KNN	99.03	93.81	95.94	98.50	97.46	96.10	96.92
KNN+TMGWO	99.03	94.26	96.25	98.50	97.62	96,33	97.12
KNN+SMOTE	99.68	93.78	95.94	99.50	97.77	96.60	97.31
KNN+TMGWO+SMOTE	99.68	97.55	98.44	99.50	99.06	98.51	98.85

Discussion

Table 7 shows that there are some increases in each data mining process. For example, in the accuracy, the classification of the early-stage diabetes risk prediction dataset using KNN using all features in the dataset gets an accuracy of 96.92%. Meanwhile, classification using only selected features by TMGWO gives an accuracy rate of 97.12% with a 0.20% improvement. Furthermore, classification using all features and applying SMOTE on training data before the classification process can increase the accuracy by 0.39%, from 96.92% to 97.31%.

Meanwhile, the proposed method in this study gets the highest accuracy among all other methods. Classification of early-stage diabetes risk prediction with only selected features by TMGWO and by applying SMOTE on training data before the classification process gets an accuracy of 98.85%. Our proposed method gives a 1.73% increase compared to classification using only selected features by TMGWO, a 1.54% increase compared to classification on balanced training data. The most significant improvement occurs when compared to classification using all features and without balancing the training data, with a 1.93% increase. Furthermore, we also compare our accuracy with previous research to prove that our proposed method has advantages compared to their methods. We have chosen three previous studies that use the same early stage diabetes risk prediction dataset. The comparison with previous research can be seen in Table 8.

Table 8. Accuracy comparison with previous research (%)

Algorithms	Accuracy
ANN+IG [13]	98.08
MLP+APGWO [17]	97.12
Stacked Ensemble+MDI [15]	97
TMGWO+SMOTE+KNN	98.85

The results show that our proposed method has the highest accuracy compared to the method proposed by another researcher. In our proposed method, by using TMGWO, SMOTE, and KNN, we can produce an accuracy of 98.85%. Previous research by Chaves & Marques (2021) on classifying the early-stage diabetes risk prediction dataset by using Artificial Neural Network (ANN) as the classification algorithm and Information Gain (IG) for feature selection produces an accuracy of 98.08%. On the other hand, Le et al. (2021) doing the same classification but using different methods, which are Multilayer Perceptron (MLP) as the classification algorithm and Adaptive Particle Grey Wolf Optimization (APGWO) for feature selection, getting an accuracy of 97.12% on early-stage diabetes risk prediction dataset. Saxena et al. (2021), on classifying the early-stage diabetes risk prediction dataset using Stacked Ensemble as the classification algorithm and Mean Discrete in Impurity (MDI) for feature selection, gets an accuracy rate of 97%. The results in Table 8 prove that using SMOTE to balance the training data can increase model accuracy in classifying the early-stage diabetes risk prediction dataset. Besides that, TMGWO can also select the best subset of features, and by only using ten features out of all 16 features in the dataset, it can reach the highest accuracy compared to other research.

CONCLUSION

The application of TMGWO in selecting features in the dataset can produce ten features out of a total of 16 features in the dataset. Furthermore, the dataset with only the ten features will be divided into training and testing sets using 10-fold cross validation. Then, SMOTE will be applied to training data to overcome the problem of data imbalance between classes in the dataset. The application of TMGWO and SMOTE can make the classification carried out by KNN more optimal, characterized by the high levels of accuracy, recall, precision, and f1-score obtained, as shown in Table 7. Furthermore, our proposed method also gives

the highest accuracy compared to previous studies, as shown in Table 8. From the results, it can be concluded that the combination of TMGWO for selecting the best feature subset, SMOTE for balancing the training data, and KNN for classification is the best solution to classify the early stage diabetes risk prediction dataset.

REFERENCES

[1] N. H. Cho *et al.*, "IDF Diabetes Atlas: Global estimates of diabetes prevalence for 2017 and projections for 2045," *Diabetes Res Clin Pract*, vol. 138, pp. 271–281, Apr. 2018.

[2] M. Pikkemaat, K. B. Boström, and E. L. Strandberg, "'I have got diabetes!' - Interviews of patients newly diagnosed with type 2 diabetes," *BMC Endocr Disord*, vol. 19, no. 1, pp. 1–12, May 2019.

[3] J. J. Ofman *et al.*, "Does disease management improve clinical and economic outcomes in patients with chronic diseases? A systematic review," *American Journal of Medicine*, vol. 117, no. 3, pp. 182–192, Aug. 2004.

[4] S. Vijiyarani, S. Sudha, and M. P. Research Scholar, "Disease Prediction in Data Mining Technique-A Survey," *International Journal of Computer Applications & Information Technology*, vol. 2, no. 1, p. 17, 2013.

[5] I. Kavakiotis, O. Tsave, A. Salifoglou, N. Maglaveras, I. Vlahavas, and I. Chouvarda, "Machine Learning and Data Mining Methods in Diabetes Research," *Comput Struct Biotechnol J*, vol. 15, pp. 104–116, 2017.

[6] K. Pingale, S. Surwase, V. Kulkarni, S. Sarage, and A. Karve, "Disease Prediction using Machine Learning," *International Research Journal of Engineering and Technology (IRJET)*, vol. 6, no. 12, pp. 831–833, 2019.

[7] V. López, A. Fernández, S. García, V. Palade, and F. Herrera, "An insight into classification with imbalanced data: Empirical results and current trends on using data intrinsic characteristics," *Inf Sci (N Y)*, vol. 250, pp. 113–141, Nov. 2013.

[8] J. A. Sáez, B. Krawczyk, and M. Woźniak, "Analyzing the oversampling of different classes and types of examples in multi-class imbalanced datasets," *Pattern Recognit*, vol. 57, pp. 164–178, Sep. 2016.

[9] T. Zhu, Y. Lin, and Y. Liu, "Synthetic minority oversampling technique for multiclass imbalance problems," *Pattern Recognit*, vol. 72, pp. 327–340, Dec. 2017.

[10] V. Bolón-Canedo, N. Sánchez-Maroño, and A. Alonso-Betanzos, *Feature Selection for High-Dimensional Data*. Cham: Springer International Publishing, 2015.

[11] A. Destrero, S. Mosci, C. de Mol, A. Verri, and F. Odone, "Feature selection for high-dimensional data," *Computational Management Science*, vol. 6, no. 1, pp. 25–40, 2009.

[12] M. Abdel-Basset, D. El-Shahat, I. El-henawy, V. H. C. de Albuquerque, and S. Mirjalili, "A new fusion of grey wolf optimizer algorithm with a two-phase mutation for feature selection," *Expert Syst Appl*, vol. 139, pp. 1–14, 2020.

[13] L. Chaves and G. Marques, "Data mining techniques for early diagnosis of diabetes: A comparative study," *Applied Sciences (Switzerland)*, vol. 11, no. 5, pp. 1–12, Mar. 2021.

[14] D. H. Lee, J. K. Yang, C. H. Lee, and K. J. Kim, "A data-driven approach to selection of critical process steps in the semiconductor manufacturing process considering missing and imbalanced data," *J Manuf Syst*, vol. 52, pp. 146–156, Jul. 2019, doi: 10.1016/j.jmsy.2019.07.001.

[15] S. Saxena, D. Mohapatra, S. Padhee, and G. K. Sahoo, "Machine learning algorithms for diabetes detection: a comparative evaluation of performance of algorithms," *Evol Intell*, pp. 1–17, 2021.

[16] M. M. F. Islam, R. Ferdousi, S. Rahman, and H. Y. Bushra, "Likelihood Prediction of Diabetes at Early Stage Using Data Mining Techniques," in *International Symposium on Computer Vision and Machine Intelligence in Medical Image Analysis*, 2019, pp. 113–125.

[17] T. M. Le, T. M. Vo, T. N. Pham, and S. V. T. Dao, "A Novel Wrapper-Based Feature Selection for Early Diabetes Prediction Enhanced with a Metaheuristic," *IEEE Access*, vol. 9, pp. 7869–7884, 2020.