Classification Model for Diabetes Mellitus Diagnosis based on K-Means Clustering Algorithm Optimized with Bat Algorithm

Syaiful Anam[1], Zuraidah Fitriah[2*], Noor Hidayat[3], Mochamad Hakim Akbar Assidiq Maulana[4]
Mathematics Department, Brawijaya University, Malang, Indonesia[1, 2, 3]
Undergraduate Student, Mathematics Department, Brawijaya University, Malang, Indonesia[4]

***Abstract*—Diabetes mellitus is a disease characterized by abnormal glucose homeostasis resulting in an increase in blood sugar. According to data from the International Diabetes Federation (IDF), Indonesia ranks 7th out of 10 countries with the highest number of diabetes mellitus patients in the world. The prevalence of patients with diabetes mellitus in Indonesia reaches 11.3 percent or there are 10.7 million sufferers in 2019. Prevention, risk analysis and early diagnosis of diabetes mellitus are necessary to reduce the impact of diabetes mellitus and its complications. The clustering algorithm is one of methods that can be used to diagnose and analyze the risk of diabetes mellitus. The K-mean Clustering Algorithm is the most commonly used clustering algorithm because it is easy to implement and run, computation time is fast and easy to adapt. However, this method often gets to be stuck at the local optima. The problem of the K-means Clustering Algorithm can be solved by combining the K-means Clustering algorithm with the global optimization algorithm. This algorithm has the ability to find the global optimum from many local optimums, does not require derivatives, is robust, easy to implement. The Bat Algorithm (BA) is one of global optimization methods in swarm intelligence class. BA uses automated enlargement techniques into a solution and it's accompanied by a shift from exploration mode to local intensive exploitation. Based on the background that has been explained, this article proposes the development of a classification model for diagnosing diabetes mellitus based on the K-means clustering algorithm optimized with BA. The experimental results show that the K-means clustering optimized by BA has better performance than K-means clustering in all metrics evaluations, but the computational time of the K-means clustering optimized by BA is higher than K-means clustering.**

***Keywords*—Diabetes mellitus; disease diagnosis methods; k-means clustering algorithm; optimization; bat algorithm**

I. Introduction

Diabetes mellitus is a disease characterized by abnormal glucose homeostasis resulting in an increase in blood sugar. According to data from the International Diabetes Federation (IDF), Indonesia ranks 7th out of 10 countries with the highest number of diabetes mellitus in the world. The prevalence of patients with diabetes mellitus in Indonesia reaches 11.3 percent or there are 10.7 million sufferers in 2019 [1]. Diabetes mellitus causes various complications such as cardiovascular disease, atherosclerotic disease, peripheral neuropathy, diabetic retinopathy, severe foot infections, kidney failure, and sexual dysfunction [2, 3].

Early diagnosis of diabetes mellitus is necessary to reduce the impact of diabetes mellitus and its complications. Clustering algorithms have been used to diagnose and analyze the risk of diabetes mellitus [4-6]. In general, clustering is divided into four categories of use, namely data reduction, hypothesis formation, hypothesis testing, and prediction based on groups [7]. Algorithm clustering is automatically able to recognize patterns in the data so that it can analyze the collected data without the label [8].

The K-mean Clustering Algorithm is the most commonly used clustering algorithm because it is easy to implement and run, the computation time is fast, and easy to adapt [9]. This algorithm has been used in various applications including diagnosis of diabetes mellitus [5], segmentation of diseases in plant leaves [10], heart disease prediction and classification [11, 12] and prediction of diabetes mellitus [13]. However, this method has a drawback, namely random centroid initialization causing, the algorithm to be stuck at the local optima [14]. The clustering result of the K-means algorithm becomes worse because the cluster center is stuck at the local optima. Therefore, the robust initialization of centroid is needed to obtain the good clustering result.

Problems of the K-means Clustering Algorithm can be overcome by combining the K-means Clustering with global optimization algorithms, e.g., swarm intelligence algorithm. This algorithm is able to find the global optimum from many local optimums, does not require derivatives, robust, and easy to implement [15]. Anam et al. have used a swarm intelligence-based algorithm (Particle Swarm Optimization) to segment disease in tomato leaves [16]. One of the swarm intelligence methods is Bat Algorithm, with a faster convergence rate than Genetic Algorithm and Particle Swarm Optimization [17]. This is because BA uses automated enlargement techniques into a promising solution. This enlargement is accompanied by a shift from exploration mode to local intensive exploitation. BA also has been used for many applications, for example travelling salesman problem [18, 19], resource scheduling [20, 21], customer churn [22, 23], brain tumor recognition [24, 25], estimating state of health of lithium-ion batteries [26], detection of myocardial infarction [27] and features selection [28, 29].

* Corresponding Author

TABLE I. ATTRIBUTES DESCRIPTION OF DATA SET

No.	Attribute Name	Attribute Description
1	HighBP	Respondent has high blood pressure which is decided by health professional.
2	HighChol	Respondent has ever had high blood cholesterol which is decided by health professional.
3	CholCheck	Cholesterol check in the last five years
4	BMI	Body Mass Index (BMI)
5	Smoker	Respondent has smoked at least 100 cigarettes in his/her lifetime.
6	Stroke	Respondent has a stroke.
7	HeartDiseaseorAttack	Respondent who had reported suffering from coronary heart disease or myocardial infarction.
8	PhysActivity	Respondent who reported engaging in physical activity or sports during the last 30 days apart from their regular job.
9	Fruits	Respondent consumes fruit 1 time or more per day
10	Veggies	Respondent consumes vegetable 1 time or more per day
11	HvyAlcoholConsump	Heavy drinking or not (adult men drink more than 14 drinks per week and adult women drink more than 7 drinks per week)
12	AnyHealthcare	Possession of any health care coverage, including health insurance, prepaid plans, or government plans.
13	NoDocbcCost	Was there a time in the last 12 months when you needed to see a doctor but couldn't because of costs?
14	GenHlth	General health score. [from 1 to 5]
15	MentHlth	Mental health which includes stress, depression and emotional problems, for how many days during the last 30 days your mental health was not good. [from 1 to 30]
16	PhysHlth	Physical health including physical illness and injury, for how many days during the last 30 days your physical health was not good. [from 1 to 30]
17	DiffWalk	Has serious difficulty walking/climbing stairs or not
18	Sex	Indicates the gender of the respondent. [Female: 0, Male: 1]
19	Age	Age category fourteen levels [from 1 to 14]
20	Education	The completed level of education. [from 1 to 6]
21	Income	The household's annual income from all sources: (If the respondent declines at any income level, code "Refuse.")
22	Diabetes	0 is no diabetes, 1 is pre-diabetes or diabetes

Based on the background described, this article proposes the development of a method for diagnosing diabetes mellitus based on the K-means clustering algorithm optimized by BA. The purpose of this research is to develop a rapid method for diagnosing diabetes mellitus using the K-means and BA algorithms. In this article, the K-means algorithm is improved by using the BA algorithm to overcome the problem of the K-means algorithm which is often stuck in the local optima. This research is useful for the prevention and reduction of the impact of diabetes mellitus through a rapid and inexpensive diagnosis of diabetes mellitus by utilizing information technology (machine learning).

II. PROPOSED METHOD

This sub-chapter will explain the dataset which are used, research stages, stages of the proposed method and evaluation tools used.

A. Data Set

The dataset used in this study was taken from the web at https://www.kaggle.com/datasets/alexteboul/diabetes-health-indicators-dataset. The dataset was taken from the Behavioral Risk Factor Surveillance System (BRFSS), which is a health-related telephone survey that is collected annually in America.

The dataset consists of several predictor variables, both medical and non-medical, and one target variable (diabetes mellitus sufferers and not diabetes mellitus sufferers). Description of the attributes of the dataset used can be seen in Table I. Class 1 means people with diabetes mellitus or prediabetes while class 0 means not people with diabetes mellitus. This dataset is used for the training process and for evaluation of the built diabetes mellitus prediction model.

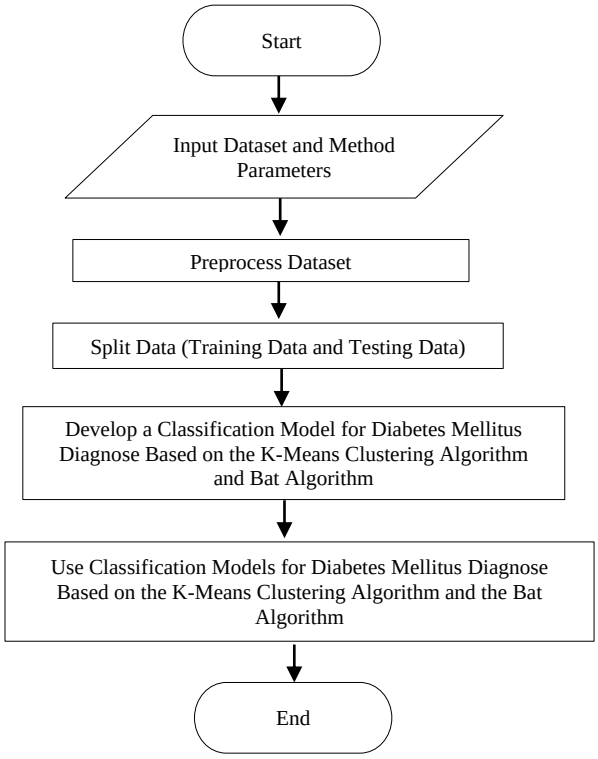

Fig. 1. Flowchart of the proposed method.

B. Classification Model for Diabetes Mellitus Diagnosis Based on K-Means Clustering Algorithm and Bat Algorithm

This section describes the steps of the Classification Model for Diabetes Mellitus Diagnosis Based on K-Means Clustering Algorithm and Bat Algorithm. The steps or stages of the Classification Model for Diabetes Mellitus Diagnosis Based on the K-Means Clustering Algorithm and the Swarm Intelligence Algorithm can be seen in Fig. 1. The method has several steps which are input dataset and method parameters, preprocess dataset, split data (training data and testing data),

develop a Classification Model for Diabetes Mellitus Diagnosis Based on the K-Means Clustering Algorithm and Bat Algorithm, and use Classification Models for Diabetes Mellitus Diagnosis Based on the K-Means Clustering Algorithm and the Bat Algorithm.

C. Parameters Setting

Before the Classification Model for Diabetes Mellitus Diagnosis Based on the K-Means Clustering Algorithm and Bat Algorithm is used, there are several parameters that must be set. Some of these parameters include:

- The number of bats used is n = 20,
- Maximum number of iterations t_{max} = 1000,
- An initial loudness (constant or decreasing) A = 1,
- An initial pulse rate (constant or decreasing) r_0 = 1,
- Alpha (α)= 0.97,
- Gamma (γ) = 0.1,
- A minimum frequency (f_{min}) = 0, and
- A maximum frequency ($f_{\max}$)= 2.

These parameters are taken from [17].

D. Data Preprocessing

Data preprocessing is an initial step in the data mining technique to convert raw data into data that is more efficient and in accordance with the data mining model to be used. Raw data taken from various sources often experience errors, missing values, and are inconsistent, so that the raw data need to be formatted so that data mining results are precise and accurate. In addition, raw data also need to be transformed to change data from its original form into data that is ready to be mined. Data transformation can facilitate the process of extracting data to find new knowledge. One of the data transformation techniques is normalization. Normalization is the process of scaling the attribute values of the data so that they can lie in a certain range. This study uses the Min-Max Normalization Method. The Min-Max Normalization is a normalization method by carrying out a linear transformation of the original data so as to produce a balance of comparison values between the data.

E. Data Splitting

After preprocessing, dataset is divided into two parts for training and testing. The proportion of training and testing data is 80% and 20%. The training data in this study is used to train the model so as to get a clustering model. Data testing is used to test and evaluate the model, as a simulation of using the model in the real world. Data testing should never be used in model training before to make model validation.

F. Develop a the Classification Model for Diabetes Mellitus Diagnosis Based on K-Means Clustering Algorithm and Bat Algorithm

The next step is to build the Classification Model for Diabetes Mellitus Diagnosis Based on K-Means Clustering Algorithm and Bat Algorithm. The diagnostic model is built based on the Clustering Method based on the K-Means Clustering Algorithm and the Bat Algorithm. The first step is to build a K-Means Clustering algorithm that is optimized with the Bat Algorithm. The position of the bat in the Bat Algorithm represents the center of the cluster (centroid). The optimized function in the Bat Algorithm is the objective function of K-means Clustering. After the algorithm is built, the next step is to implement the program with the Python programming language.

The next step is to input the training data, the parameters of the Bat Algorithm and the number of clusters. The training data that will be included in the clustering model training process is the predictor variables of the training data. While the response variable will be used later when evaluating the clustering model after the training phase is complete. Algorithm 1 states the K-means Clustering Algorithm-Based Clustering Method and the Bat Algorithm. Algorithm 2 is used for the association of cluster centers to data classes, while Algorithm 3 is used for the testing process.

Algorithm 1 Clustering Method Based on K-Means Clustering Algorithm and Bat Algorithm

Input:
The training data ($\boldsymbol{X_{train}}$) with size of $n \times m$
The number of cluster (K)
The parameters of Bat Algorithm

Output:
Best ($\boldsymbol{x}_*$) is the best solution produced

a. Initialize the bat positions and velocities $\boldsymbol{x}_i^0$ *and* $\boldsymbol{v}_i^0$ *,*$(i = 1, 2, \dots, N)$*. Each* $\boldsymbol{x}_i$ *represents a candidate from the centroid or cluster center. For example, matrix* C_i *is the i-th centroid candidate represented in equation (1).*

$$C_i = \begin{bmatrix} c_{1,1} & \cdots & c_{1,m} \\ \cdots & \cdots & \cdots \\ c_{K,1} & \cdots & c_{K,m} \end{bmatrix} \quad (1)$$

Therefore, C_i is reshaped to get a matrix of size $1 \times (m.K)$ and saved in $\boldsymbol{x}_i^0 = (x_{i,1}, \dots x_{i,m.K}) = (c_{i,1}, \dots, c_{1,m}, \dots, c_{K,1}, \dots, c_{K,m})$. It aims to facilitate the calculation process on the Bat Algorithm. $\boldsymbol{v}_i$ also sized $1 \times (m.K)$.

b. Initialize a frequency f_i*, a pulse rate* r_i*, and a loudness* A_i

c. t=0

d. ***while*** *(*$t < Maximum\ Iteration$ *)* ***do***

1. **for** $i = 1{:}N$ **do**
   i. Generate the new solutions by adjusting the frequency, updating the velocity and the position of bats using equations (2), (3), and (4).

$$f_i = f_{min} + (f_{max} - f_{min})\beta, \quad (2)$$

$$\boldsymbol{v}_i^{t+1} = \boldsymbol{v}_i^t + (\boldsymbol{x}_i^t - \boldsymbol{x}_*)f_i, \quad (3)$$

$$\boldsymbol{x}_i^{t+1} = \boldsymbol{x}_i^t + \boldsymbol{v}_i^{t+1}. \quad (4)$$

   ii. **if** ($rand < r_i$ **then**

Generate the local solutions randomly by

using equation (5),

$$\boldsymbol{x}_{new} = \boldsymbol{x}_{old} + \sigma\epsilon_t A^{(t)} \quad (5)$$

where ϵ_t is random numbers obtained from the normal Gaussian distribution $N(0,1)$, $A^{(t)}$ is the average loudness of all bats over time t, and σ is the scale factor, for simplification, can be used $\sigma = 0.01$.

end if

iii. Evaluate the fitness using the objective function of K-means clustering which is stated in equation (6),

$$J = \sum_{j=1}^{K}\sum_{i=1}^{a_j}\left\|\boldsymbol{X}_{train}^{j} - \boldsymbol{c}_j\right\|^2 \quad (6)$$

where $\boldsymbol{c}_j$ represents centroid j of K centroid. $\boldsymbol{c}_j$ reshaping results were obtained $\mathbf{x}_i$ of size $1 \times K.m$ to matrix $\boldsymbol{C}_i$ of size $K \times m$.

iv. **if** $(rand > A_i$ **and** $f(\boldsymbol{x}_i) < \mathcal{F}(\boldsymbol{x}_*))$ **then**
Update the current solution using one of the solutions from step (i) or (ii)
end if

v. Increase r_i and reduce A_i by using Equations (7) and (8),

$$A_i^{t+1} = \alpha A_i^t, \quad (7)$$

$$r_i^{t+1} = r_i^0[1 - \exp(-\gamma t)], \quad (8)$$

where $0 < \alpha < 1$, and $\gamma > 0$. According to Yang (2014), to facilitate the search process A_i and r_i can equate the value of α and γ, with value $\alpha = \gamma = 0.9$.

vi. Sort the bats and determine the best solution $(\boldsymbol{x}_*)$

end for
end while

e. Do a reshape on Best $(\boldsymbol{x}_*)$ to get centroid $\boldsymbol{C}$.

Algorithm 2 Centroid analysis on K-Means Clustering Algorithm-Based Clustering Method and Bat Algorithm

Input:
The training data ($\boldsymbol{X_{train}}$) with size of n × m
The centroids
y_{train} (class label of each training data)
Output:
Accuracy, Recall, Precision, *F1 Score*.

1. Determine which centroid represents the target class based on the majority value of the labels in each cluster.

2. Calculate the label prediction y_{pred} based on centroid cluster.

3. Calculate Accuracy, Recall, Precision and F_1 Score.

Algorithm 3 Testing Algorithm of Classification model of Diabetes Mellitus Diagnosis Based Clustering Method K-Means Clustering and Bat Algorithm

Input:
The testing data
The centroids
$y_{testing}$ (class labels of each testing data)
Output:
Accuracy, Recall, Precision, *F1 Score*.

1. Calculate the label prediction y_{pred} based on centroid cluster.

2. Calculate Accuracy, Recall, Precision and F1 Score.

G. Evaluation Metrics

The performance of the proposed method is evaluated by using accuracy, recall, precision and *F1 Score*. The performance of the proposed method is compared to the previous method, namely the K-means Clustering method. If the proposed method is better than the standard method, it can be said that the performance of this method can be improved. The evaluation metrics used are:

1) Classification Rate / Accuracy which is calculated using the formulation in equation (9),

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (9)$$

where TP states that diabetics and is detected as a diabetic. TN stated that healthy person and is detected as a healthy person. FN is the healthy person but detected as diabetics. FP stated that diabetics but detected as a healthy person. Accuracy is used to measure the ratio of correct predictions to the total number of instances evaluated.

2) Recall is calculated by the formulation in equation (10). Recall is used to measure the fraction of a correctly classified positive pattern.

$$Recall = \frac{TP}{TP+FN} \quad (10)$$

3) Precision is calculated by the formula in equation (11). Precision is used to measure the correctly predicted positive pattern from the total predicted pattern in the positive class.

$$Precision = \frac{TP}{TP+FP} \quad (11)$$

4) F1 Score is calculated by the formula in equation (12). F1 Score is harmonic mean of precision and recall.

$$F1\ Score = 2 \cdot \frac{Precision \cdot Recall}{Precision+Recall} \quad (12)$$

After the evaluation metrics are calculated, the experimental results are analyzed to obtain conclusions.

III. Results and Discussions

The dataset has different scale on each feature/variable; therefore the algorithm cannot work well. So that, the dataset is needed to be pre-processed to solve this problem by using normalization technique. Furthermore, the data are normalized by using the minmax normalization method. This process will make data that has the same range, namely between values 0 and 1. The normalized data is then divided into two parts, namely, 80% of the data is used for training and the remaining 20% is used for testing. After the data is appropriate with the model to be used, the data can be input for the algorithm to be executed.

The evaluation tools used to measure the quality of each algorithm are the objective function (f_{min}), accuracy, precision, recall, *F1 score*, and the computational time. The accuracy, precision, recall and *F1 score* are calculated for both training data and testing data. The evaluation tool will reach the optimum value when the objective function is minimum, the accuracy, precision, recall and *F1 score* are maximum, and the computation time is not too long.

The first algorithm to run is the standard K-means Algorithm. The parameters used are the number of clusters of 2 which correspond to the expected number of targets. The iterations are carried out until one of stopping conditions is reached. The parameters used in the Bat Algorithm are initialized with the parameters described in sub-section II.C. These algorithm uses two stopping conditions which are the maximum iteration and convergence condition. The maximum iteration is 1000 times. The algorithm is assumed convergence if the global best doesn't have improvement in 100 iteration. The experiments in this study were repeated 25 times, because the Bat Algorithm and the K-means Algorithm used in the Classification Model for Diabetes Mellitus Diagnosis Based on the K-Means Clustering Algorithm and the Bat Algorithm involve random numbers in obtaining the optimum value of the objective function. Then the average and standard deviation of the evaluation tool used are calculated. The standard deviation is used to measure the spread of recall, accuracy, precision and *F1 scores*, as well as objective function values. While the average value is used to concentrate the results of recall, accuracy, precision and *F1* scores, as well as objective function values.

Table II shows the comparison of objective function value of Classification Model for Diabetes Mellitus Diagnosis based on the K-means Clustering Algorithm and the Bat Algorithm (training data). It can be shown that the Classification Model for Diabetes Mellitus Diagnosis based on the K-means Clustering Algorithm and the Bat Algorithm has better performance than K-means Clustering Method. It results the smaller the objective values.

TABLE II. COMPARISON OF OBJECTIVE FUNCTION VALUE OF CLASSIFICATION MODEL FOR DIABETES MELLITUS DIAGNOSIS BASED ON THE K-MEANS CLUSTERING ALGORITHM AND THE BAT ALGORITHM (TRAINING DATA)

Method	Average	Deviation Standard
K-means	6389.589	82.41
K-means + Bat Algorithm, *n*=10	6341.573	3.9882
K-means + Bat Algorithm, *n*=20	**6340.507**	**0.0291**

TABLE III. MEAN OF EVALUATION METRICS FOR CLASSIFICATION MODEL FOR DIABETES MELLITUS DIAGNOSIS BASED ON THE K-MEANS CLUSTERING ALGORITHM AND THE BAT ALGORITHM (TRAINING DATA)

Method	Accuracy	Precision	Recall	F1 Score	Time (s)
K-means	0.7009	0.70797	0.71500	0.70498	**0.02856**
K-means + Bat Algorithm, *n*=10	0.72427	**0.73632**	0.69402	0.71443	1008.645
K-means + Bat Algorithm, *n*=20	**0.72431**	0.73459	**0.69744**	**0.71553**	2794.129

TABLE IV. DEVIATION STANDARD OF EVALUATION METRICS FOR CLASSIFICATION MODEL FOR DIABETES MELLITUS DIAGNOSIS BASED ON THE K-MEANS CLUSTERING ALGORITHM AND THE BAT ALGORITHM (TRAINING DATA)

Method	Accuracy	Precision	Recall	F1 Score	Time (s)
K-means	0.06834	0.07391	0.09370	0.04588	0.00927
K-means + Bat Algorithm, *n*=10	0.002607	0.003948	0.014368	0.006265	544.6731
K-means + Bat Algorithm, *n*=20	0.000591	0.00041	0.002546	0.001166	4542.019

TABLE V. MEANS OF EVALUATION METRICS FOR CLASSIFICATION MODEL FOR DIABETES MELLITUS DIAGNOSIS BASED ON THE K-MEANS CLUSTERING ALGORITHM AND THE BAT ALGORITHM (TESTING DATA)

Method	Accuracy	Precision	Recall	F1 Score
K-means	0.6956	0.7010	0.6550	0.6764
K-means + Bat Algorithm, *n*=10	0.7155	**0.7311**	0.6938	**0.7118**
K-means + Bat Algorithm, *n*=20	**0.7156**	0.7128	**0.6971**	0.70471

TABLE VI. DEVIATION STANDARD OF EVALUATION METRICS FOR CLASSIFICATION MODEL FOR DIABETES MELLITUS DIAGNOSIS BASED ON THE K-MEANS CLUSTERING ALGORITHM AND THE BAT ALGORITHM (TESTING DATA)

Method	Accuracy	Recall	Precision	F1 Score
K-means	0.06068	0.06819	0.04169	0.04988
K-means + Bat Algorithm, *n*=10	0.00146	0.02574	0.01117	0.01363
K-means + Bat Algorithm, *n*=20	0.00083	0.02778	0.00276	0.0136

Table III shows the mean of evaluation metrics for Classification Model for Diabetes Mellitus Diagnosis based on the K-means Clustering Algorithm and the Bat Algorithm (training data), while Table IV shows the deviation standard of evaluation metrics for Classification Model for Diabetes Mellitus Diagnosis based on the K-means Clustering Algorithm and the Bat Algorithm (training data). The experimental results show that the that the Model for Diabetes Mellitus Diagnosis based on the K-means Clustering Algorithm and the Bat Algorithm are superior to the standard K-means method. that the Model for Diabetes Mellitus Diagnosis based on the K-means Clustering Algorithm and the Bat Algorithm are superior in all evaluation tools (accuracy, recall, precision and *F1 Score*). The standard deviation of the accuracy, recall, precision and *F1 score* of the Diabetes Mellitus Diagnostic Model Based on the K-Means Clustering Algorithm and the Swarm Intelligence Algorithm is very small. This means that this method results in low variation. The computational time required for training the Model for Diabetes Mellitus Diagnosis based on the K-means Clustering Algorithm and the Bat Algorithm is much longer than the standard K-means method.

Table V shows the results of the classification model evaluation for data testing. The experimental results show that the Classification Model for Diabetes Mellitus Diagnosis based on the K-means Clustering Algorithm and the Bat Algorithm yields much better results compared to the standard K-means method for accuracy, recall, precision and *F1 Score values*. Tables V and VI also show the accuracy, recall, precision and *F1 scores* produced by the Model for Diabetes

Mellitus Diagnosis based on the K-means Clustering Algorithm and the Bat Algorithm are not much different between training and testing data. This shows that the proposed method has good performance and neither overfitting nor underfitting occurs.

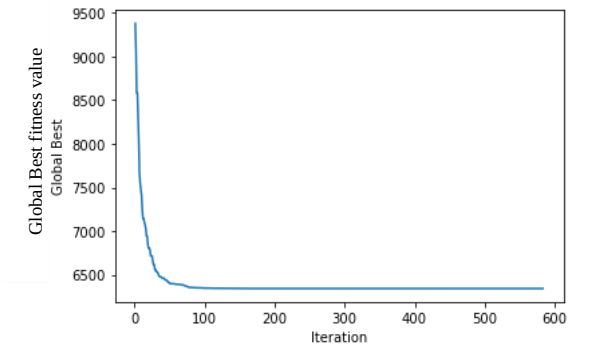

Fig. 2. Graph of iteration and global best relationships of the diabetes mellitus diagnostic method based on the K-Means clustering algorithm and the swarm intelligence algorithm with 20 particles (global best fitness value 6340,499).

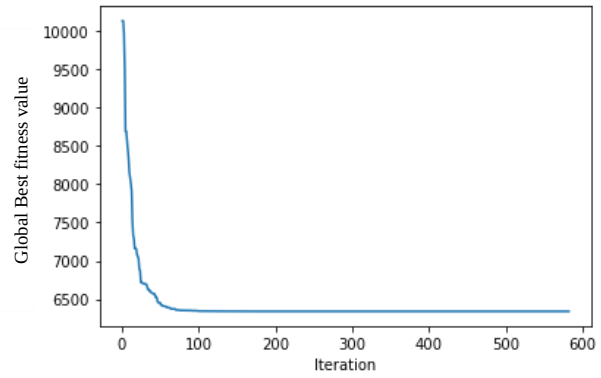

Fig. 3. Graph of iteration and global best relationships of the diabetes mellitus diagnostic method based on the K-Means clustering algorithm and the swarm intelligence algorithm with 10 particles (global best fitness value = 6340.497).

Fig. 2 and 3 show graphs of the iteration and global best relationships of the Classification Model for Diabetes Mellitus Diagnosis based on the K-means Clustering Algorithm and the Bat Algorithm with 20, and 10 Particles. The figures show the method converges less than 600 iterations for the number of particles 20 and 10. The global best convergent fitness value is around 6340.5.

IV. Conclusions

Based on the experimental results and analysis of the experimental results, several conclusions were obtained. that the Model for Diabetes Mellitus Diagnosis based on the K-means Clustering Algorithm and the Bat Algorithm developed from the K-Means Clustering Algorithm by adding the Bat Algorithm optimizer to determine the centroid of the cluster. The experimental results revealed that the number of bats has an effect on the method's convergence speed and processing time. The experimental results reveal that the Model for Diabetes Mellitus Diagnosis based on the K-means Clustering Algorithm and the Bat Algorithm are able to diagnose diabetes mellitus quite well. The accuracy obtained is around 72.4% and the *F1 score* is 71.4% for training data, and the accuracy obtained is around 71.55% and the *F1 score* is 71.18% for data testing. The evaluation results show that the performance of the Model for Diabetes Mellitus Diagnosis based on the K-means Clustering Algorithm and the Bat Algorithm is better than the standard K-means for evaluating accuracy, precision, recovery, f1 score, but the recovery time is quite large.

Acknowledgment

We would like to thank Brawijaya University for funding this research through the Hibah Penelitian Pemula (HPP) research grant.

References

[1] InfoDatin, “Pusat data dan informasi Kementrian Kesehatan RI”, 2020.

[2] K. Papatheodorou, M. Banach, M. Edmonds, N. Papanas, and D. Papazoglou1, “Complications of diabetes”, Journal of Diabetes Research, vol. 2015, 2015, http://dx.doi.org/10.1155/2015/189525

[3] D. Tomic, J. E. Shaw, and D. J. Magliano, “The burden and risks of emerging complications of diabetes mellitus”, Nature Reviews Endocrinology, vol. 18, 2022, pp. 525–539.

[4] H. Wu, S. Yang, Z. Huang, J. He, and X. Wang, “Type 2 diabetes mellitus prediction model based on data mining”, Informatics in Medicine Unlocked, vo.10, 2018, pp.100-107

[5] T. Santhanam and M. S. Padmavathi, “Application of k-means and genetic algorithms for dimension reduction by integrating SVM for diabetes diagnosis”, Procedia Computer Science, vol. 47, 2015, pp. 76 – 83

[6] A. M. Simarmata, M. A. P. Sianipar, S. Singh, I. I. M. Gulo, and J. B. R. Purba, “Grouping diabetes diagnosis based on age range with k-means algorithm”, Jurnal Mantik, vol. 5, no. 2, 2021, pp. 1408-1412

[7] S. Theodoridis, and K. Koutroumbas, “Clustering: basic concepts”, Pattern Recognition, Edisi 4, Academic Press, 2009, pp. 595–625. doi:10.1016/b978-1-59749-272-0.50013-x

[8] Y. Lei, Z. He, Y. Zi, and X. Chen, “New clustering algorithm-based fault diagnosis using compensation distance evaluation technique”, Mechanical Systems and Signal Processing, vol. 22, 2008, pp. 419–435.

[9] M. Mancas, and B. Gosselin, “Fuzzy tumor segmentation based on iterative watersheds”, Proc. STW Conf. of ProRISC, Veldhoven, Netherlands, 2003

[10] S. Zhang, H. Wang, W. Huang, and Z. You, “Plant diseased leaf segmentation and recognition by fusion of superpixel, k-means and PHOG”, Optik, vol. 157, 2018, pp. 866–872

[11] M. Thangamani, R. Vijayalakshmi, M. Ganthimathi, M. Ranjitha, P. Malarkodi and S. Nallusamy, “Efficient classification of heart disease using k-means clustering algorithm”, International Journal of Engineering Trends and Technology, vol. 68, no. 12, 2020, pp. 48-53.

[12] R. Shinde, S. Arjun, P. Patil and J. Waghmare, “An intelligent heart disease prediction system using K-means Clustering and Naïve Bayes Algorithm”, International Journal of Computer Science and Information Technologies, vol. 6, no. 1, 2015, pp. 637-639.

[13] C. Fiarni, E. M. Sipayung and S. Maemunah, “Analysis and prediction of diabetes complication disease using data mining algorithm”, Procedia Computer Science, vol. 161, 2019, pp. 449–457

[14] K., Singh, D. Malik, and N. Sharma, “Evolving limitations in k-means algorithm in data mining and their removal,” International Journal of Computational Engineering & Management, vol. 12, 2011, pp. 105-109.

[15] A. K Kordon, “Swarm intelligence: the benefits of swarms”, Applying Computational Intelligence, 2009, pp. 145-174.

[16] S. Anam, and Z. Fitriah, “Early blight disease segmentation on tomato plant using k-means algorithm with swarm intelligence-based algorithm”, International Journal of Mathematics and Computer Science, vol. 16, no. 4, 2021

[17] X. Yang, "A new metaheuristic bat-inspired algorithm", Nature Inspired Cooperative Strategies for Optimization, Springer, 2010, pp 65-74.

[18] E. Osaba, X. Yang, F. Diaz, P. Lopez-Garcia, and R. Carballedo, "An improved discrete bat algorithm for symmetric and asymmetric Traveling Salesman Problems", Engineering Applications of Artificial Intelligence, vol. 48, 2016, pp. 59-71.

[19] Y. Saji and M. Barkatou, "A discrete bat algorithm based on Lévy flights for Euclidean traveling salesman problem", Expert Systems with Applications, vol. 172, 2021, 114639.

[20] L. Jacob, "Bat Algorithm for resource scheduling in cloud computing", International Journal for Research in Applied Science and Engineering Technology, vol. 2 (4), 2014, pp. 53-57.

[21] J. Zheng and Y.Wang, "A hybrid multi-objective bat algorithm for solving cloud computing resource scheduling problems", Sustainability, vol. 13, pp. 1-25.

[22] S. Induja, and V. P. Eswaramurthy, "Customer churn prediction and attribute selection in telecom industry using kernelized extreme learning machine and bat algorithms", International Journal of Science and Research, vol. 5 (12), 2016, pp. 258-565.

[23] M. Li, | C. Yan, W. Liu anf X. Liu, "An early warning model for customer churn prediction in telecommunication sector based on improved bat algorithm to optimize ELM", International Journal of Intelligent Systems, vol. 36, no. 7, 2021, pp. 1-28.

[24] R. Chawla, S. M. Beram, C. R. Murthy, T. Thiruvenkadam, N. P G. Bhavani, R. Saravanakumar, and P. J. Sathishkumar, "Brain tumor recognition using an integrated bat algorithm with a convolutional neural network approach", Measurement: Sensors, vol. 24, 2022.

[25] G. R. Sreekanth, A. F. Alrasheedi, K. Venkatachalam, M. Abouhawwash, and S. S. Askar, "Extreme Learning Bat Algorithm in brain tumor classification", Intelligent Automation & Soft Computing, vol. 34, no.1, 2022, pp. 249-265.

[26] D. Ge, Z. Zhang, X. Kong and Z. Wan, "Extreme Learning Machine using Bat Optimization Algorithm for estimating state of health of lithium-ion batteries", Appl. Sci. 2022, 12, 1398

[27] P. Kora and S. R. Kalva, "Improved Bat algorithm for the detection of myocardial infarction", Springer Plus, vol. 4, 2015, pp. 1-18.

[28] S. Jeyasingh and M. Veluchamy, "Modified Bat Algorithm for feature selection with the Wisconsin Diagnosis Breast Cancer (WDBC) dataset", Asian Pacific Journal of Cancer Prevention, vol. 18, no. 5, pp. 1257-1264.

[29] B. Yang, Y. Lu, K. Zhu, Gu. Yang, J. Liu and H. Yin, "Feature Selection Based on Modified Bat Algorithm", IEICE Trans. Inf. & Syst., vol. E100–D, no. 8, 2017, pp. 1860-1869.