Prediction Analysis of Diabetes Mellitus Based on Machine Learning Algorithm

[1st] Rumini
Department of Information Systems
Universitas Amikom Yogyakarta
Yogyakarta, Indonesia
rumini@amikom.ac.id

[2nd] Sri Ngudi Wahyuni
Department of Informatics Management
Universitas Amikom Yogyakarta
Yogyakarta, Indonesia
yuni@amikom.ac.id

[3rd] Bambang Sudaryatno
Department of Information Systems
Universitas Amikom Yogyakarta
Yogyakarta, Indonesia
bambang_s@amikom.ac.id

[3rd] Arvi Pramudyantoro
Department of Informatics
Amikom University Yogyakarta
Yogyakarta, Indonesia
arvi.pramudyantoro@students.amikom.ac.id

2022 5th International Seminar on Research of Information Technology and Intelligent Systems (ISRITI) | 978-1-6654-5512-1/22/$31.00 ©2022 IEEE | DOI: 10.1109/ISRITI56927.2022.10052794

Abstract — **Diabetes Mellitus is a metabolic disease of several etiologies characterized by chronic hyperglycemia with impaired carbohydrate, fat, and protein metabolism due to defects in insulin secretion, insulin action, or both. Diabetes will make approximately 6.7 million deaths in 2021, so one person dies every 5 seconds. Diabetes accounts for at least USD 966 billion dollars in health spending and has seen a 316% increase over the last 15 years. In Indonesia, Diabetes Mellitus will reach 21 million people in 2030 and is increasing. In the long term, it will place a significant financial burden on the health care system as well as on individuals and society as a whole. Based on the description above, the purpose of this study is to predict people with Diabetes Mellitus using the K-Nearest Neighbor (K-NN) algorithm. The data is taken from Kaggle and data processing uses Python. The testing model in this study used a confusion matrix and the testing accuracy used Recall and precision. The results show that K-NN has an accuracy value of 84.41%, which means for every 77 test samples there are 17 people diagnosed as positive. This is indicated by the recall value of 92.85% and the precision value of 86%. This result shows the K-NN algorithm has good accuracy in the prediction problem.**

Keywords— Diabetes Mellitus, Prediction, K-NN, Machine learning

I. INTRODUCTION

Diabetes is a metabolic disease of multiple etiologies characterized by chronic hyperglycemia with impaired carbohydrate, fat, and protein metabolism due to defects in insulin secretion, insulin action, or both [1]. Diabetes is a long-life disease because of high levels of sugar in the blood [2], approximately 90-95% of people have type 2 diabetes in the world [3]. In 2035 the number of people with diabetes is predicted to be around 592 million, which is almost double the current number [4]. Based on 2021 International Diabetes Federation data, more than 537 million adults (20-79 years) live with diabetes -1 in 10 [5]. In 2030, the predicted of number diabetes mellitus patients will increase to 643 million and to 783 million in 2045. In 2021, DiabeteMellitusus will generate approximately 6.7 million deaths, which means there are one person dies every 5 seconds. Diabetes accounts for at least USD 966 billion dollars in health spending and has seen a 316% increase over the last 15 years.

During the COVID-19 pandemic, diabetes mellitus patients are the largest contributor to death worldwide. This disease needs to be watched at early age so as to be able to make early preparations for treatment or healing treatment. In Indonesia, the number of people with Diabetes Mellitus is increasing. Thus in the long term, will place a significant financial burden on the health care system as well as on individuals and society as a whole [1]. Prediction is one way to determine the first step in understanding and knowing about the early symptoms of Diabetes Mellitus. So as to prevent and reduce the prevalence rate so as not to increase and provide opportunities for healing to sufferers.

Several Diabetes Mellitus predictive based on machine learning research conducted by many researchers. Olisa et al. used Random Forest, Support Vector Machine (SVM) to predict type 2 Diabetes Mellitus patients [2]. Several studies use the Naïve Bayes algorithm, Decision tree, Neural Network, Adaboost [3], Logistics Regression [4], [5]. Mujumdar compared the prediction algorithms of Supervised Learning, Unsupervised Learning, and Semi-supervised Learning for the case of Diabetes Mellitus type 2 prediction, and the results showed that the semi-supervised learning algorithm had higher accuracy than all algorithms in machine learning [6]. The KNN algorithm is also widely used in the fields of economics and business [7],[8], education [9], prediction of plant diseases [10]–[13], or humans diseases, including rheumatism [14], Parkinson [15], Alzheimer [16],

Based on the description above, no one study about Diabetes Mellitus prediction using parameter number of times of delivery, glucose levels, blood pressure, skin thickness, insulin levels, body weight, family history of diabetes, and age. The purpose of this study is to predict people with Diabetes Mellitus using machine learning-based K-Nearest Neighbor. The research method will be discussed in the second session, the results and discussion in the third session and conclusions in the fourth session, the fifth session is a bibliography.

II. METHODOLOGY

A. Dataset processing

The dataset in this study was taken in Kaggle. The total data is 768. The data is divided into 90% as training data and 10% as testing data. The predictive parameters used were the number of times of delivery, glucose levels, blood pressure, skin thickness, insulin levels, body weight, family history of diabetes, and age. These parameters will be normalized into a category for further processing. Then determine the class in the dataset, namely 0 and 1. Where class 1 is a positive class and 0 is a negative class. The data is divided into two parts, namely 691 data for training data and 77 data for testing data.

B. Propose method

The step of methodology in this study shows in figure 1.

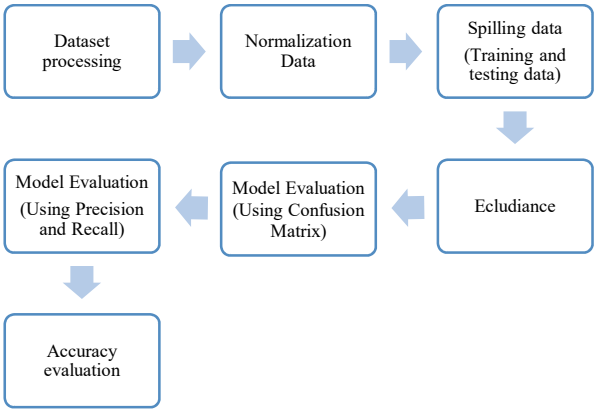

Fig 1. Proposed method

The steps of research in Fig 1. the research begins with dataset collection via Kaggle. Then the data were normalized using the MinMax Schaler model. To get the value 0 or 1 in the dataset. Furthermore, the data will be split into 90% training data and 10% testing data. So that the composition of the data is

C. K-Nearest Neighbor

K-Nearest Neighbor is an algorithm used to classify data based on attributes and samples. KNN was first introduced by T. Cover and P. Hart in 1967 where this algorithm classifies sample classes based on their closest neighbor classes [17], [18]. The K-NN algorithm is used for predictions, especially the neighboring classification as the prediction value of the new instance. The near and far distance can be calculated by the Euclidean distance (d) and the formula to find the distance between 2 points in two - dimensional space is as follows $X_1 = (x^{11}, x^{21}, x^{31} \ldots x^{m1})$and $X_2 = (x^{12}, x^{22}, x^{32} \ldots x^{m2})$. Where, constant, x_1 is the i-th data entry from the test data, and $x_2$is the i-th data entry from the training data.

D. Exception distance

Euclidean Distance is a technique to avoid a probability value of 0, due to the absence of data for a particular attribute in a class . In this technique, one data value is added to each calculation, so that it can avoid the case of a probability value of 0. The distance formulation is presented in the following formula,

a. Euclidean distance:

$$d(x_1, x_2) = \sqrt{\sum i = 1m(x_{i1} - x_{i2})^2} \quad (2)$$

b. Manhattan distance:

$$d(x_1, x_2) = \sum i = 1m|x_{i1} - x_{i2}| \quad (3)$$

c. Minkowski distance:

$$d(x_1, x_2) = (\sum i = 1m|x_{i1} - x_{i2}|^p)^{1/p} \quad (4)$$

Where: m is the number of training data, p is a constant, x_1 is the i-th data entry from the test data, and $x_2$is the i-th data entry from the training data [2]. If p = 1, then the Minskowski distance function is the same as the Manhattan distance function. Meanwhile, if p = 2 then the Minskowski distance function is the same as the Euclidean distance function. Laplace Correction has the following formula:

$$P_i = \frac{m_i+1}{n+k} \quad (5)$$

where k is the number of bins or class of attributesm_i.

E. Accuracy testing

The evaluation method used in this study is the Confusion Matrix. This method is a table consisting of the number of true or false data rows from a tested/predicted classification model. The aim is to determine the performance of a classification model. The steps of the confusion matrix are presented in Table 1.

TABLE 1. CONFUSION MATRIX

		Prediction Class	
		+	-
Actual Class	+	True Positive (TP)	False Negative (FN)
	-	False Positive (FP)	True Negative (TN)

Where:

1. *True Positive* (TP) = the result of a positive system prediction and in accordance with the actual positive
2. *True Negative* (TN) = the result of the system prediction which is negative and corresponds to the actual negative.
3. *False Positive* (FP) = the result of the positive system prediction, while the actual negative. *False Negative* (FN) = the result of the negative system prediction, while the actual positive.

Precision is a true positive classification and the overall predicted data is a positive class. The precision equation is presented in formula 6.

$$Precision = \frac{TP}{TP+FP} \quad (6)$$

Recall is the number of documents that have a *true positive classification* of all data that are truly positive (including *false negatives*). The recall equation is presented in formula 7.

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

Accuracy is the closeness value between the number of documents that have a correct classification (*true positive and true negative*) and the sum of all data. The accuracy equation is presented in formula 8.

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (8)$$

III. ANALYSIS AND DISCUSSION

A. Data normalization

The preprocessing stage in this research is normalization of data using the MinMax Schaler method to obtain the same range of data, which is between 0 to 1. The data normalization method used is the min - max schaler method. The results of normalization are shown in Figure 2.

	0	1	2	3	4	5	6	7
0	0.470588	0.929032	0.551020	0.147135	0.079086	0.607362	0.025192	0.366667
1	0.411765	0.696774	0.653061	0.402174	0.079086	0.650307	0.110589	0.250000
2	0.117647	0.354839	0.285714	0.086957	0.096154	0.130879	0.238685	0.000000
3	0.058824	0.419355	0.326531	0.152174	0.145433	0.143149	0.322374	0.033333
4	0.117647	0.283871	0.510204	0.130435	0.046875	0.220859	0.064475	0.016667

Fig 2. Results of data normalization using MinMax Schaller.

After the normalization process, the data will be divided, namely 90% for training data and 10% for testing data. The next stage is to place the basic assumptions as the basis for making predictions, namely:

1. Women with multiple pregnancies have a higher chance of being positive for diabetes.
2. A person with a high plasma glucose concentration will have a higher chance of being positive for diabetes.
3. A person with high blood pressure will have a higher chance of being positive for diabetes.
4. A person with a higher tricep average (body fat) will have a higher chance of being positive for diabetes.
5. A person with a high amount of insulin will have a higher chance of being positive for diabetes.
6. A person with a high BMI will have a higher chance of being positive for diabetes.
7. Someone with old age will have a higher chance of being positive for diabetes.

Based on the description above, the proportion between positive and negative data is depicted in Figure 2.

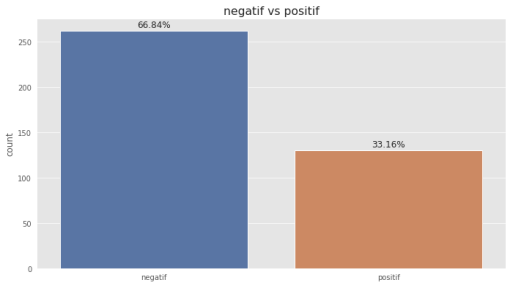

Figure 2. Proportion of positive and negative in the dataset

Figure 2 describes the proportion of negative patients with diabetes mellitus of 66.94%, the remaining 33.16% of patients tested positive for diabetes. Based on the basic assumptions, the following analyzes are used, including:

1. The first assumption is, if it is known that the higher the *Glucose Level* , the higher the chance that a person will be positively diagnosed with diabetes. The prediction results are presented in Figure 3.

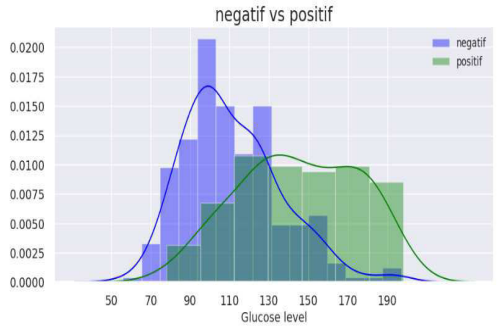

Figure 3. The results of the diagnosis of Diabetes Mellitus based on the *Glucose Level parameter*

2. The second assumption is that the higher the blood pressure, the higher the potential for suffering from Diabetes Mellitus. In this case as many as 3.32% of positive patients, presented in Figure 4.

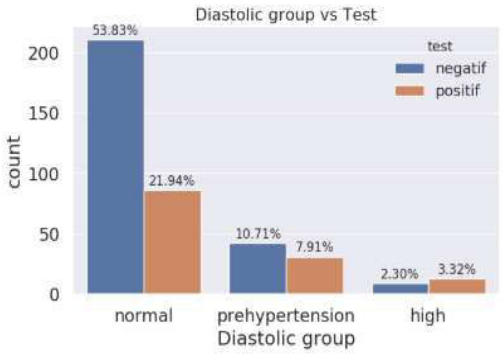

Figure 4. The results of the Diabetes Mellitus analysis based on the height of blood pressure

The third assumption is that the higher the triceps or body fat, the higher the chance for a person to be positively diagnosed with Diabetes Mellitus. Prediction results are presented in Figure 5.

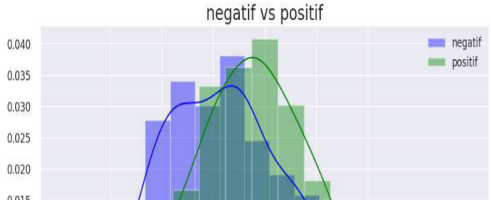

Figure 5. Results of Diabetes Mellitus analysis based on body fat height

2. The fourth assumption is that the higher the BMI or the patient 's weight, the higher the chance that a person will be positively diagnosed with Diabetes Mellitus . Prediction results are presented in Figure 6.

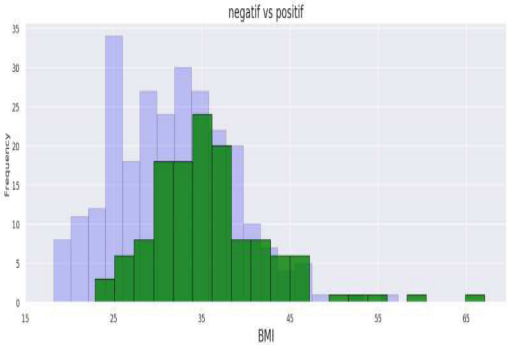

Figure 6. Diagnosis of Diabetes Mellitus based on parameters weight

3. The fifth assumption is that the higher the amount of insulin, the higher the chance that a person will be positively diagnosed with diabetes. Prediction results are presented in Figure 7.

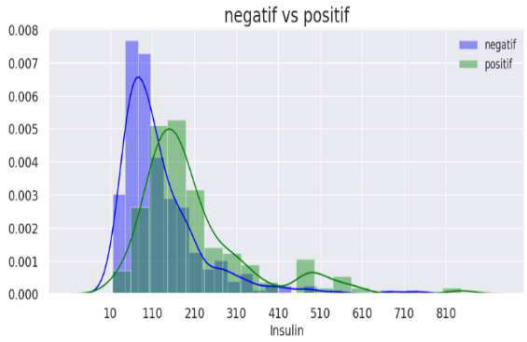

Figure 7. Diagnosis of Diabetes Mellitus based on parameters of insulin

4. The sixth assumption is that the older a person is, namely over 30 years of age, the higher the chance that a person will be positively diagnosed with Diabetes Mellitus. Prediction results are presented in Figure 8.

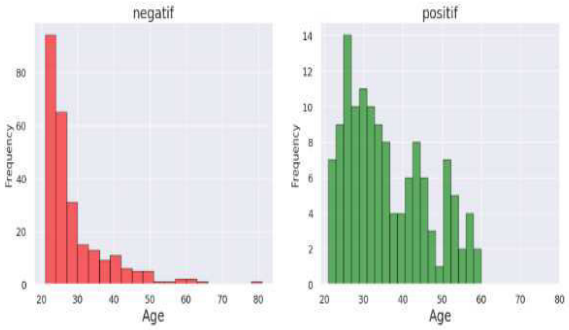

Figure 8. Analysis of Diabetes Mellitus by age

B. Determine the ecludience value

Using formula 2, the resulting euclidience distance value is in Table 3. In this process, 8 data are used from testing data.

TABLE II. EUCLIDENCE RESULT

Test_1	Test_2	…	…	Test_8
1.600553	1.363975	…	…	1.327675
0.887649	1.333159	…	…	0.912648
1.279874	1.655012	…	…	1.347433
1.29512	1.666787	…	…	1.567657
…	…	…	…	…
…	…	…	…	…
…	…	…	…	…
…	…	…	…	…
…	…	…	…	…
…	…	…	…	…
1.424362	1.794991	…	…	1.57382
1.302593	1.10551	…	…	1.323893
0.90149	1.497238	…	…	0.919813

After getting the value of the exclusion distance in Table II, the value of k will be searched. for the first value of k we take randomly, i.e. k=5. Furthermore, based on the value of k = 5, the value of the ecludience disatance will be calculated 5 times. So the results are shown in Table III.

TABLE III. RESULT OF CALCULATION VALUE K=5

Test_1	Test_2			Test_7	Test_8
5	5			5	5
…	…	…	…	…	…
Positive		…	…	…	Positive
…	…	…	…	…	…
…	…	…	…	…	…
…	…	…	…	Negative	Negative
…	Positive	…	…	…	…
…		…	…	…	…
…	Negative	…	…	…	…
Negative	Negative	…	…	…	…
…	…	…	…	Negative	…
…	…	…	…	…	…
…	Positive	…	…	…	…
Negative	…	…	…	Negative	Negative
…	…	…	…	…	…
…	…	…	…	…	…
…	…	…	…	…	…
…	…	…	…	…	Positive
Positive	Positive	…	…	…	Positive
…	…	…	…	Negative	…
Negative	…	…	…	Negative	…
2	3	2	1	0	3
3	2	3	4	5	2

Based on Table III, predictions will be made between positive cases and negative cases. If the number of positive cases of Diabetes Mellitus is greater than negative cases, it will produce positive data and vice versa. It is rule based:

Positive > Negative = Positive
or
Negative > Positive = Negative

C. Testing accuracy

Based on the above formulation, the Confusion Matrix value is generated in Table IV.

TABLE IV. TEST RESULTS USING CONFUSSION MATRIX

Actual Diabetes	Diabetes Prediction	Confusion Matrix
Negative	Negative	TN
Positive	Positive	TP
Positive	Negative	FN
…	…	…
…	…	…
Positive	Negative	FN
Positive	Positive	TP

Based on Table IV, the resulting number of negative and positive data is presented in Table V.

TABLE V. FIRST TEST RESULTS USING CONFUSION MATRIX

		Diabetes Prediction		Amount
		Positive	Negative	
Diabetes current	Positive	52	8	60
	Negative	4	13	12
Amount		56	21	77

Table V explains that there are 77 people who are categorized as positive and 13 people who are categorized as negative. Furthermore, there are 4 people who are predicted to be positive but fall into the negative class and 8 negative people with Diabetes Mellitus who fall into the positive category. Based on this, it is necessary to test accuracy, recall and precision to get more accurate results. The results of the accuracy test using the Confusion Matrix are presented in Figure 8.

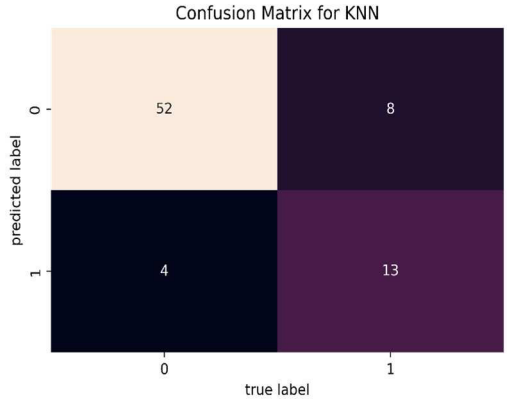

Figure 9. The results of the accuracy test using the Confusion Matrix.

Based on Table 6. Then the model accuracy values are:

Accuracy = (TP+TN) / (TP+TN+FP+FN)
= (52+13) / (52+13+4+8)
= 65 / 77
= 0.8441 * 100%
= 84.41 %

This shows that the model accuracy value of 84.41% is indicated by the results of the above accuracy calculation. The next test is the accuracy test using Recall based on formula 7.

Recall = TP / (TP+FN)
= 52 / (52+4)
= 52 /56
= 0.9285 * 100
= 92.85%

From the calculation using Formlua 7, the recall accuracy result is 92.85%. The following is an accuracy test using precision in formula 8.

Precision = TP / (TP+FP)
= 52 / (52+8)
= 52 /60
= 0.86 *100
= 86%

Based on the above precision calculation, the precision value of the K-NN model is 86%. This means that the K-NN value still needs to be improved.

V. CONCLUSION

The conclusion of this study is that K-NN is an algorithm that has a fairly good accuracy value in machine learning prediction algorithms. This is indicated by an accuracy value of 84.41% with a recall value of 92.85% and a precision value of 86%. This shows that the accuracy of K-NN is quite good. In future research, a comparison test will be carried out on the accuracy of machine learning-based prediction models, especially for classification cases, to find the best model accuracy.

ACKNOWLEDGMENT

This work was supported by Universitas Amikom Yogyakarta, Indonesia

REFERENCES

[1] "Diabetes." https://www.who.int/news-room/fact-sheets/detail/diabetes (accessed Feb. 18, 2022).

[2] C. C. Olisah, L. Smith, and M. Smith, "Diabetes mellitus prediction and diagnosis from a data preprocessing and machine learning perspective," *Computer Methods and Programs in*

Biomedicine, vol. 220, Jun. 2022, doi: 10.1016/J.CMPB.2022.106773.

[3] V. Rawat, S. Joshi, S. Gupta, D. P. Singh, and N. Singh, "Machine learning algorithms for early diagnosis of diabetes mellitus: A comparative study," *Materials Today: Proceedings*, vol. 56, pp. 502–506, Jan. 2022, doi: 10.1016/J.MATPR.2022.02.172.

[4] F. Alaa Khaleel and A. M. Al-Bakry, "Diagnosis of diabetes using machine learning algorithms," *Materials Today: Proceedings*, Jul. 2021, doi: 10.1016/J.MATPR.2021.07.196.

[5] N. P. Tigga and S. Garg, "Prediction of Type 2 Diabetes using Machine Learning Classification Methods," *Procedia Computer Science*, vol. 167, pp. 706–716, 2020, doi: 10.1016/J.PROCS.2020.03.336.

[6] A. Mujumdar and V. Vaidehi, "Diabetes Prediction using Machine Learning Algorithms," *Procedia Computer Science*, vol. 165, pp. 292–299, 2019, doi: 10.1016/J.PROCS.2020.01.047.

[7] S. B. Imandoust and M. Bolandraftar, "Application of k-nearest neighbor (knn) approach for predicting economic events: Theoretical background," *Int J Eng Res Appl*, vol. 3, no. 5, pp. 605–610, 2013.

[8] O. Dogan and B. Oztaysi, "Genders prediction from indoor customer paths by Levenshtein-based fuzzy kNN," *Expert Systems with Applications*, vol. 136, pp. 42–49, 2019.

[9] A. Maghari, "Prediction of student's performance using modified KNN classifiers," in *Alfere, SS, & Maghari, AY (2018). Prediction of Student's Performance Using Modified KNN Classifiers. In The First International Conference on Engineering and Future Technology (ICEFT 2018)*, 2018, pp. 143–150.

[10] E. Hossain, M. F. Hossain, and M. A. Rahaman, "A color and texture based approach for the detection and classification of plant leaf disease using KNN classifier," in *2019 International Conference on Electrical, Computer and Communication Engineering (ECCE)*, 2019, pp. 1–6.

[11] N. Krithika and A. G. Selvarani, "An individual grape leaf disease identification using leaf skeletons and KNN classification," in *2017 International Conference on Innovations in Information, Embedded and Communication Systems (ICIIECS)*, 2017, pp. 1–5.

[12] M. P. Vaishnnave, K. S. Devi, P. Srinivasan, and G. A. P. Jothi, "Detection and classification of groundnut leaf diseases using KNN classifier," in *2019 IEEE International Conference on System, Computation, Automation and Networking (ICSCAN)*, 2019, pp. 1–5.

[13] U. Shruthi, V. Nagaveni, and B. K. Raghavendra, "A review on machine learning classification techniques for plant disease detection," in *2019 5th International conference on advanced computing & communication systems (ICACCS)*, 2019, pp. 281–284.

[14] S. Rajathi and G. Radhamani, "Prediction and analysis of Rheumatic heart disease using kNN classification with ACO," in *2016 International Conference on Data Mining and Advanced Computing (SAPIENCE)*, 2016, pp. 68–73.

[15] R. A. Shirvan and E. Tahami, "Voice analysis for detecting Parkinson's disease using genetic algorithm and KNN classification method," in *2011 18th Iranian conference of biomedical engineering (ICBME)*, 2011, pp. 278–283.

[16] A. bin Tufail, A. Abidi, A. M. Siddiqui, and M. S. Younis, "Automatic classification of initial categories of Alzheimer's disease from structural MRI phase images: a comparison of PSVM, KNN and ANN methods," *International Journal of Biomedical and Biological Engineering*, vol. 6, no. 12, pp. 713–717, 2012.

[17] J. M. Keller, M. R. Gray, and J. A. Givens, "A fuzzy k-nearest neighbor algorithm," *IEEE Trans Syst Man Cybern*, no. 4, pp. 580–585, 1985.

[18] T. Cover and P. Hart, "Nearest neighbor pattern classification," *IEEE Trans Inf Theory*, vol. 13, no. 1, pp. 21–27, 1967.