Scientific Journal of Informatics
Vol. 10, No. 3, Aug 2023

p-ISSN 2407-7658 http://journal.unnes.ac.id/nju/index.php/sji e-ISSN 2460-0040

A Comparative Analysis on the Evaluation of KNN and SVM Algorithms in the Classification of Diabetes

Agus Fahmi Limas[1*], Rika Rosnelly[2], Hartono[3], Aly Nursie[4]
[1,2,3,4] Department of Computer Science, Faculty of Engineering and Computer Science,
Potensi Utama University, Indonesia

Abstract.
Purpose: Diabetes has received a great deal of attention in medical research because of its profound effect on human health. Many factors cause this disease in the human body. Can be from food or drink that is often consumed by the human body. Diabetes cannot be cured and can only be controlled.
Methods: In this study, using 2 data mining techniques namely Support Vector Machine and K-Nearest Neighbor were applied to predict diabetes. In this study, 768 diabetes data were used as trial data, consisting of training data that had been pre-processed data and 400 data cleaning data, 278 data testing data, and 50 diabetes data samples used as samples in the calculation.
Result: The performance of each algorithm is analyzed differently, the results of each best algorithm will be analyzed to determine which algorithm can provide better results for predicting diabetes. The results obtained in this study get a value of 0 where the predicted value of the target class for new data is the negative class (Suffer).
Novelty: This study compares the SVM and K-NN methods for diabetes classification. So, successfully implemented for data on the classification target

Keywords: Data mining, Diabetes, K-Nearest Neighbor, Support Vector Machine
Received May 2023 / **Revised** May 20223/ **Accepted** June 2023

INTRODUCTION

Diabetes is a chronic disease caused by a lack of insulin production (a hormone produced by the pancreas to regulate glucose levels) in the human body [1], [2]. Glucose is the main source of energy for body cells. Lack of insulin secretion causes blood glucose levels to increase and exceed the normal limit that should be in the blood, resulting in a buildup of glucose. Glucose that accumulates in the blood because it is not properly absorbed by body cells can cause various organ disorders [3]–[5]. Diabetes is a disease that cannot be taken lightly and needs to be treated quickly. If diabetes is not controlled properly, it can cause various complications that endanger the lives of sufferers.

Diabetes Mellitus is a multifactorial disease, characterized by chronic hyperglycemia syndrome and disturbances of carbohydrate, fat and protein metabolism caused by insufficiency of insulin secretion or endogenous insulin activity or both. Uncontrolled hyperglycemia can also cause many complications such as neuropathy, stroke and peripheral vascular disease. The classic symptoms of Diabetes Mellitus are polyuria (urinating a lot), [3], [6] polydipsia (drinking a lot) and polyphagia (eating a lot). These symptoms may occur, especially in children (months or weeks), but they may be indistinguishable or completely absent, in addition to growing much more slowly, there may also be weight loss/load (other than a normal or heavier diet) and fatigue that can no longer be minimized. These symptoms can also turn into diabetes in patients who are not well-controlled.

In general, the incidence of Diabetes Mellitus is influenced by a lack of exercise or activity. Physical activity has a strong influence on energy balance and can be said to be the main modifiable factor through which many external forces that trigger weight gain work. Physical exercise in people with diabetes mellitus has a very important role in controlling blood sugar levels, when doing physical exercise there is an increase in

[*] Corresponding author.
Email addresses: agusfahmilimasptr@gmail.com (Limas)
DOI: 10.15294/sji.v10i3.44269

the use of glucose by active muscles so it directly causes a decrease in blood glucose. Most of the causes of diabetes are the increasing number of people who are overweight or obese [7].

K-Nearest Neighbour (KNN) algorithm is a simple data classification algorithm that uses the shortest distance computation to categorize a new instance based on a similarity metric. The KNN algorithm is a supervised algorithm, which means that it is formed by a learning process on previously classified records, and the results of this learning are used to categorize new records with unknown output [8]. The K-Nearest Neighbour algorithm is a previously classified classification method that classifies the results of a new query instance based on the majority of the proximity of the existing categories [9]–[11]. The SVM (Support Vector Machine) technique is a supervised learning approach for determining the optimum hyperplane by maximizing the distance between classes. Since its inception in 1995, the Support Vector Machine (SVM) method has been widely employed as a classification tool. Data is frequently separated into training and test sets during the categorization process. Each instance has one goal value and various attributes in the training data set. SVM's purpose is to create a model based on training data that can predict the target value of the provided test data based only on its properties [7], [12]–[16].

Research conducted by [14] regarding the classification of underdeveloped districts using a Support Vector Machine yielded that districts in KTI could be classified correctly at 87.23%. Based on the ROC curve, an AUC value of 0.862 is obtained, which means the model performance is good. Research conducted by [17] also uses the K-Nearst Neighbor method and SupportVector Machine for data classification of prospective blood donors, which gives results in the form of the accuracy of 90% greater than the K-Nearest Neighbor Algorithm, which only has an accuracy of 76%. With a data ratio of 60:40 with a total dataset of 200 data [18]–[20]. Research conducted by [21] for classifying households receiving electricity subsidies in Gorontalo province also uses the K-Nearest Neighbor and Support Vector Machine, resulting in the KNN method having better accuracy in carrying out classification, namely 98.07% [22]–[26]. Overall, there is a significant difference between KNN and SVM classification, where KNN performance is much better than SVM in classifying [27]–[30]. Based on several journal sources explaining the advantages and disadvantages of the KNN and SVM algorithms for their accuracy in predicting the results of a study, the researchers tried to carry out a comparative analysis of the performance of KNN and SVM.

To overcome this problem, this study tested two algorithm methods: the KNN algorithm (K-Nearest Neighbors) and SVM (Support Vector Machine). The two methods will be examined to determine which is most accurately used to predict diabetes, which often occurs in every human being. With the prediction of determining diabetes, it is hoped that it can simplify and add value to knowledge in determining diabetes, making it easier for a doctor or medical staff to classify diabetes based on data generated from existing processes.

METHODS

Data Collection

The data collection process is a very important component in research. To avoid errors in the analysis process, it requires a careful and correct data collection process, besides that the results and conclusions obtained are also ambiguous if the data collection is carried out incorrectly. In this sub-chapter, we will discuss the concept of collecting data sets that are carried out. Data to be used in the analysis process as well as sampling techniques in data collection, with details as follows:

1. Data Category
   Based on sources from the website of the Ministry of Health, in this study, the required data was divided into 4 diabetes categories (classes), namely type 1 diabetes, type 2 diabetes, gestational diabetes and other types of diabetes.
2. Data Collection
   Diabetes data collection was taken from the UCI Machine Learning Repository website as many as 768 data.
3. Sampling Technique
   The sampling technique in this study was purposive sampling, which means a sampling technique in which the sample is taken purposively according to the required sample data criteria. The amount of data on diabetics depends on the needs of system users. In this study, the data is divided into two parts, namely training data and testing data. Training data is data that contains a collection of diabetic data which will be used as training data for the system. Here the data collected already has the categories of type 1 diabetes, type 2 diabetes, gestational diabetes and other types of diabetes. While data testing

is data that contains a collection of data on diabetics in the category of type 1 diabetes, type 2 diabetes, gestational diabetes and other types of diabetes. Later this data will be analyzed and tested by the system based on existing training data. The comparison to be used in this sampling technique is 80:20 and 60:40 [31]. Determination of training data and test data can be seen from the equation below:

$$Test\ Data = \frac{Distribution \times Amount\ of\ data}{100} \tag{1}$$

$$Train\ Data = Amount\ of\ data - Test\ Data \tag{2}$$

Preprocessing Dataset

After identifying the problem and collecting data, the next step is the Preprocessing Dataset. Dataset preprocessing is the process of converting raw data into a form that is easier to understand. This process aims to facilitate the processing of diabetes patient data during the analysis process using the SVM and KNN methods [32], [33].

Data Training and Data Testing

Training data and data testing are carried out after the dataset is obtained, or later training will be used to train the algorithm in finding an appropriate model, while data testing will be used to test and find out the performance of the model obtained at the testing stage [34].

1. Preprocessing Analysis

   Data preprocessing is an important step in carrying out classification analysis which aims to clean data from elements that are not needed to speed up the classification process.

   The stages in the data cleaning process take sample test data from diabetes data collection from the UCI Machine Learning Repository site, as much as 768 data. Type 2, gestational diabetes, and other types of diabetes. Moreover, In the next stage, the data is tested using a sampling technique where sampling is deliberately taken according to the required sample data criteria [3], [15], [35]–[38].
2. Analysis with KNN and SVM

   At this stage, an analysis of the classification method that has been obtained with training data is carried out.

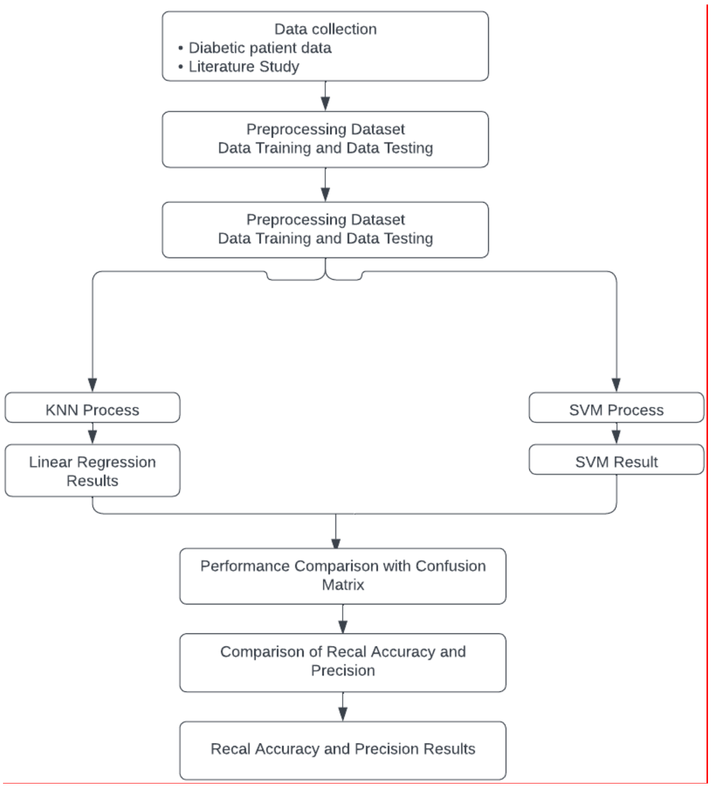
Data collection
• Diabetic patient data
• Literature Study
Preprocessing Dataset
Data Training and Data Testing
Preprocessing Dataset
Data Training and Data Testing
KNN Process
Linear Regression Results
SVM Process
SVM Result
Performance Comparison with Confusion Matrix
Comparison of Recal Accuracy and Precision
Recal Accuracy and Precision Results

Figure 1. Stages of research methodology

In Figure 1 above are the methodological stages of this study. First, collect data on diabetes patients and collect literature studies, then preprocess the dataset, and conduct trials on training data and data testing. The next stage is to perform calculations with the KNN algorithm and the SVM algorithm so that it can be seen from the results of the linear regression at the KNN stage and the results of the SVM. Then a performance comparison is made by looking at the confusion matrix. At performance of the recall accuracy and precision, the results of the recall accuracy and precision are obtained [17], [18], [28], [32], [39], [40].

RESULT AND DISCUSSION

Data analysis is the most critical process carried out during research. Data analysis is used to solve research problems. The data source used in this study comes from kaggle.com data. In this study, 768 diabetes data were used, consisting of 400 training data, 278 testing data, and 50 diabetes data samples used as manual calculation samples. This data is used for analysis and testing.

Table 1. Transformation data

No	Pregnancy	Glucose	Blood pressure	Insulin	Decision Status
1.	2	8,5	77	48	Very good
2.	9	6,2	65	40	Good
3.	4	9,6	78	46	Enough
4.	1	8,5	77	46	Enough
5.	5	14	98	50	Enough
6.	2	11,4	84	50	Enough
7.	3	12	91	49	Enough
8.	3	11,8	92	49	Enough
9.	9	6,9	65	40	Enough
10.	5	13,3	101	48	Kurang
11.	4	15,1	96	49	Enough
12.	1	6,1	67	46	Not Enough
13.	4	12,1	90	49	Enough
14.	2	14,3	87	51	Enough
15.	8	5,4	61	40	Not Enough
16.	3	12,8	87	50	Enough
17.	3	13,1	92	50	Enough
18.	4	13,3	99	49	Enough
19.	2	15,3	80	50	Enough
20.	5	11	106	49	Enough
21	4	17,2	102	44	Enough
22	1	6,5	71	46	Enough
23	3	11,5	90	49	Enough
24	2	8,1	81	50	Enough
25	3	11,9	90	50	Enough
26	2	10,9	83	51	Enough
27	1	5,4	66	41	Enough
28	1	9,1	73	46	Enough
29	3	15,5	105	51	Enough
30	4	16	98	51	Enough
31	3	9,5	87	48	Enough
32	4	8,1	72	45	Enough
33	5	14	100	48	Enough
34	2	10,1	80	57	Enough
35	4	14,5	99	50	Enough
36	2	10,5	86	48	Enough
37	2	10	83	48	Enough
38	8	6,1	65	40	Enough
39	5	15	107	51	Enough
40	1	11,4	82	49	Enough
41	11	16	105	50	Enough
42	3	68	67	43	Enough
43	3	19,6	92	47	Enough
44	1	12	92	51	Enough
45	1	9,7	73	47	Enough
46	5	10,3	78	48	Not Enough
47	2	42	57	40	Not Enough
48	1	10,4	82	48	Not Enough
49	3	10,1	80	46	Not Enough
50	3	11,3	89	49	Not Enough

Table 1 describes the stages of the data transformation process, which includes changing the decision status on each attribute in the data to be processed. Data transformation is a crucial stage in data analysis because it can help correct missing or invalid values, change the scale of the data, and improve the shape of the data distribution. Data transformations can also help adapt data to certain statistical model assumptions. In some cases, data transformation can improve the quality of the model and make the analysis more accurate and meaningful. Therefore, selecting and applying appropriate data transformation techniques can significantly affect the final results of data analysis.

The next stage is the calculation with the KNN algorithm:

1. First, we determine the parameter K. First, we must determine the value of K first. There is no exact formula for determining the value of K. However, one tip that can be considered is that if the number of classes is even, then the K value should be odd, conversely, if the class number is odd, then the K value should be even. For example, let's make the number of nearest neighbors K = 3.
2. Second, we calculate the distance between the new data and all the training data. We use Euclidean distance. We calculate based on the formula:

$$dis = \sqrt{\sum_{i=0}^{n} (X1i - X2i)^2 + \ (Y1i - Y2i)^2 + ..}$$

$$Line\ 1 = \sqrt{(2-10)^2 + (8{,}5 - 11{,}2)^2 + (77-92)^2 + (48-48)^2} = 17{,}21307642$$

$$Line\ 2 = \sqrt{(9-10)^2 + (6{,}2 - 11{,}2)^2 + (65-92)^2 + (40-48)^2} = 28{,}61817604$$

$$Line\ 50 = \sqrt{(3-10)^2 + (11{,}3 - 11{,}2)^2 + (89-92)^2 + (49-48)^2} = 7{,}681796665$$

Table 2. Euclidean distance calculation results

No.	Euclidean Distance Calculation Results	No.	Euclidean Distance Calculation Results
1.	17,21307642	26.	12,41329932
2.	28,61817604	27.	28,97654224
3.	15,44538766	28.	21,22286503
4.	17,81263596	29.	15,6681205
5.	8,534635317	30.	10,2
6.	11,49086594	31.	8,768694316
7.	7,186097689	32.	21,32158531
8.	7,096478	33.	9,84073168
9.	28,50421022	34.	17,03555106
10.	10,50761629	35.	9,994498487
11.	8,258934556	36.	10,02447006
12.	27,12950423	37.	12,10123961
13.	6,466065264	38.	28,68815086
14.	10,37352399	39.	16,53602129
15.	32,59815946	40.	13,49221998
16.	8,975522269	41.	14,03709372
17.	7,523961722	42.	62,65173581
18.	9,508417324	43.	10,97998179
19.	15,12646687	44.	9,520504188
20.	14,90100668	45.	21,10094785
21.	13,7113092	46.	14,89328708
22.	23,41132205	47.	47,97541037
23.	7,354590403	48.	13,47738847
24.	14,09290602	49.	14,07870733
25.	7,582216035	50.	7,681796665

3. We sort the distance from the new data with the training data and determine the nearest neighbor based on the minimum distance K=3. As for sorting from the largest to the smallest number using the sort function. So that it becomes like Table 3.

Table 3. Distance order

No	Euclidean Distance Calculation Results	Sorting Distance Largest to Smallest	Sorting Results
1.	17,21307642	62,65173581	1
2.	28,61817604	47,97541037	2
3.	15,44538766	32,59815946	3
4.	17,81263596	28,97654224	4
5.	8,534635317	28,68815086	5
6.	11,49086594	28,61817604	6
7.	7,186097689	28,50421022	7
8.	7,096478	27,12950423	8
9.	28,50421022	23,41132205	9
10.	10,50761629	21,32158531	10
11.	8,258934556	21,22286503	11
12.	27,12950423	21,10094785	12
13.	6,466065264	17,81263596	13
14.	10,37352399	17,21307642	14
15.	32,59815946	17,03555106	15
16.	8,975522269	16,53602129	16
17.	7,523961722	15,6681205	17
18.	9,508417324	15,44538766	18
19.	15,12646687	15,12646687	19
20.	14,90100668	14,90100668	20
21.	13,7113092	14,89328708	21
22.	23,41132205	14,09290602	22
23.	7,354590403	14,07870733	23
24.	14,09290602	14,03709372	24
25.	7,582216035	13,7113092	25
26.	12,41329932	13,49221998	26
27.	28,97654224	13,47738847	27
28.	21,22286503	12,41329932	28
29.	15,6681205	12,10123961	29
30.	10,2	11,49086594	30
31.	8,768694316	10,97998179	31
32.	21,32158531	10,50761629	32
33.	9,84073168	10,37352399	33
34.	17,03555106	10,2	34
35.	9,994498487	10,02447006	35
36.	10,02447006	9,994498487	36
37.	12,10123961	9,84073168	37
38.	28,68815086	9,520504188	38
39.	16,53602129	9,508417324	39
40.	13,49221998	8,975522269	40
41.	14,03709372	8,768694316	41
42.	62,65173581	8,534635317	42
43.	10,97998179	8,258934556	43
44.	9,520504188	7,681796665	44
45.	21,10094785	7,582216035	45
46.	14,89328708	7,523961722	46
47.	47,97541037	7,354590403	47
48.	13,47738847	7,186097689	48
49.	14,07870733	7,096478	49
50.	7,681796665	6,466065264	50

4. After sorting, choose as many as K (=3), while the provisions for decisions that can be made and will be used as a dataset to look for support vector machine calculations can be seen in Table 4.

Table 4. Dataset decision

No	Pregnancy	Glucose	Blood pressure	Insulin	Classification Results	Distance
13	3	6.7	33	9	Enough	6.466065264
8	2	6.4	35	9	Enough	7.096478
7	2	6.6	34	9	Enough	7.186097689

The next stage is calculating using the Support Vector Machine (SVM) to get the decision results. The dataset consisting of training data as a sample is given as input-output pairs. Each input-output pair consists of 2 inputs, namely X1 and X2, and 1 output, Y. Then, the contents of the dataset are changed using the principle of random numbers which can be seen in Table 5.

Table 5. Conversion datasets

No.	X1	X2	Y
1	90	95	1
2	85	92	1
3	75	72	0

The steps to calculate SVM are as follows:

1. The first step is to calculate the mean and standard deviation of each feature (X1 and X2) in the training data set. This is done to normalize the data.
   Average value X1
   = (90 + 85 + 75 + 60 + 70 + 80 + 66 + 72 + 85 + 68) / 10
   = 75.6
   Standard deviation of values X1
   $= \sqrt{[(1/10) * ((90-75.6)^2 + (85-75.6)^2 + (75-75.6)^2 + (60-75.6)^2 + (70-75.6)^2 + (80-75.6)^2 + (66-75.6)^2 + (72-75.6)^2 + (85-75.6)^2 + (68-75.6)^2)]}$
   = 9.55
   Average value X2
   = (95 + 92 + 72 + 65 + 82 + 85 + 70 + 95 + 92 + 75) / 10
   = 81.3
   Standard deviation of values X2
   $= \sqrt{[(1/10) * ((95-81.3)^2 + (92-81.3)^2 + (72-81.3)^2 + (65-81.3)^2 + (82-81.3)^2 + (85-81.3)^2 + (70-81.3)^2 + (95-81.3)^2 + (92-81.3)^2 + (75-81.3)^2)]}$
   = 11.27
2. Data normalization. Each feature value (X1 value and X2 value) in the training data set will be reduced by the average and divided by the standard deviation.
   Value normalized data X1
   Data 1: (90 - 75.6) / 9.55 = 1.51
   Data 2: (85 - 75.6) / 9.55 = 1.00
   Data 3: (75 - 75.6) / 9.55 = -0.06
   Value normalized data X2
   Data 1: (95 - 81.3) / 11.27 = 1.22
   Data 2: (92 - 81.3) / 11.27 = 0.95
   Data 3: (72 - 81.3) / 11.27 = -0.82
   The next stage is calculating the hyperplane equation using the formula w1 * x1 + w2 * x2 + b = 0.
   = w1 * x1 + w2 * x2 + b = 0 -3.87 * x1 - 5.10 * x2 - 0.43
   = 0

From the results of the discussion of the calculations above, we have found the values of αi, w1, w2, and b needed to create the SVM model when placing apprenticeship locations using the linear kernel method. Define the target class for new data by using a hyperplane equation. Suppose the result of the hyperplane equation is more significant than zero. In that case, the new data is classified as a positive class (Not Suffering), whereas if the result is less than zero, the new data is classified as a negative class (Suffering).

CONCLUSION

Conclusion of this study from the results of the calculations above, it shows that the predicted value of the target class for the new data is the negative class (Suffer). In this research, Support Vector Machine and K-Nearest Neighbor are applied to predict diabetes. The performance of each algorithm is analyzed differently, the results of each best algorithm will be analyzed to determine which algorithm can provide better results for predicting diabetes. This research can be developed with the addition of analysis using other methods.

REFERENCES

[1] P. P. Siswazah, "Universiti Malaysia Perlis Pusat Pengajian Siswazah Centre for Graduate Studies," pp. 1–3.

[2] G. A. Naikoo and M. U. D. Sheikh, "Development of Highly Efficient NiO Based Composite Materials for Ultra-sensitive Glucose Sensors Non Enzymatic Glucose Sensors," in *Proc. Int. Conf. Sens. Technol., ICST*, 2019, pp. 1–4, doi: 10.1109/ICST46873.2019.9047722.

[3] R. Patil and S. Tamane, "A Comparative Analysis on the Evaluation of Classification Algorithms in the Prediction of Diabetes," *Int. J. Electr. Comput. Eng.*, vol. 8, no. 5, pp. 3966–3975, 2018, doi: 10.11591/ijece.v8i5.pp3966-3975.

[4] L. Qiao, Y. Zhu, and H. Zhou, "Diabetic Retinopathy Detection Using Prognosis of Microaneurysm and Early Diagnosis System for Non-Proliferative Diabetic Retinopathy Based on Deep Learning Algorithms," *IEEE Access*, vol. 8, pp. 104292–104302, 2020, doi: 10.1109/ACCESS.2020.2993937.

[5] K. Puttananjegowda, A. Takshi, and S. Thomas, "Silicon Carbide Nanoparticles Electrospun Nanofibrous Enzymatic Glucose Sensor," *Biosens. Bioelectron.*, vol. 186, 2021, doi: 10.1016/j.bios.2021.113285.

[6] P. M. Arabi, G. Joshi, S. Lohith, M. Mubashir, P. Y. Pallavi, and M. Surabhi, "Early Diagnosis of Diabetic Vasculopathy by Thermo-Regulatory Vascular Impairment," in *2020 5th Int. Conf. Commun. Electron. Syst. (ICCES)*, 2020, pp. 1368–1373, doi: 10.1109/icces48766.2020.9137907.

[7] S. A. D. Alalwan, "Diabetic Analytics: Proposed Conceptual Data Mining Approaches in Type 2 Diabetes Dataset," *Indones. J. Electr. Eng. Comput. Sci.*, vol. 14, no. 1, pp. 85–95, 2019, doi: 10.11591/ijeecs.v14.i1.pp88-95.

[8] S. Wong, "Stock Price Prediction Model Based on the Short-term Trending of KNN Method," in *2020 7th Int. Conf. Inf. Sci. Control Eng., ICISCE 2020*, 2020, pp. 1355–1360, doi: 10.1109/ICISCE50968.2020.00273.

[9] L. R. Somula and M. Meena, "K-Nearest Neighbour (KNN) Algorithm based Cooperative Spectrum Sensing in Cognitive Radio Networks," in *Proc. 4th Int. Conf. Cybern. Cogn. Mach. Learn. Appl., ICCCMLA 2022*, 2022, pp. 1–6, doi: 10.1109/ICCCMLA56841.2022.9988996.

[10] R. Zhang, T. Du, S. Qu, and H. Sun, "Adaptive Density-based Clustering Algorithm with Shared KNN Conflict Game," *Inf. Sci. (Ny).*, vol. 565, pp. 344–369, 2021, doi: 10.1016/j.ins.2021.02.017.

[11] S. DiLmaç, Z. Dokur, and T. Ölmez, "Comparative Analysis of MABC with KNN, SOM, and ACO Algorithms for ECG Heartbeat Classification," *Turkish J. Electr. Eng. Comput. Sci.*, vol. 26, no. 6, pp. 2819–2830, 2018, doi: 10.3906/elk-1712-328.

[12] A. Purwanto and L. Parningotan Manik, "Software Effort Estimation Using Logarithmic Fuzzy Preference Programming and Least Squares Support Vector Machines," *Sci. J. Inform.*, vol. 10, no. 1, pp. 1–12, 2023, doi: 10.15294/sji.v10i1.39865.

[13] U. I. Larasati, M. A. Muslim, R. Arifudin, and A. Alamsyah, "Improve the Accuracy of Support Vector Machine Using Chi Square Statistic and Term Frequency Inverse Document Frequency on Movie Review Sentiment Analysis," *Sci. J. Inform.*, vol. 6, no. 1, pp. 138–149, 2019, doi: 10.15294/sji.v6i1.14244.

[14] N. A. Noori and A. A. Yassin, "A Comparative Analysis for Diabetic Prediction Based on Machine Learning Techniques," *J. Basrah Res.*, vol. 47, no. 1, pp. 180–190, 2021, [Online]. Available: https://www.iasj.net/iasj/journal/260/issues.

[15] L. Mohan, J. Pant, P. Suyal, and A. Kumar, "Support Vector Machine Accuracy Improvement with Classification," in *2020 12th Int. Conf. Comput. Intell. Commun. Netw., CICN 2020*, 2020, pp. 477–481, doi: 10.1109/CICN49253.2020.9242572.

[16] S. E. Pandarakone, S. Gunasekaran, Y. Mizuno, and H. Nakamura, "Application of Naive Bayes Classifier Theorem in Detecting Induction Motor Bearing Failure," in *2018 23rd Int. Conf. Electr. Mach., ICEM 2018*, 2018, pp. 1761–1767, doi: 10.1109/ICELMACH.2018.8506836.

[17] N. G. Ramadhan, "Comparative Analysis of ADASYN-SVM and SMOTE-SVM Methods on the Detection of Type 2 Diabetes Mellitus," *Sci. J. Inform.*, vol. 8, no. 2, pp. 276–282, 2021, doi: 10.15294/sji.v8i2.32484.

[18] C. Chethana, "Prediction of Heart Disease Using Different KNN Classifier," in *5th Int. Conf. Intell. Comput. Control Syst., ICICCS 2021*, 2021, pp. 1186–1194, doi: 10.1109/ICICCS51141.2021.9432178.

[19] J. Huang, Y. Wei, J. Yi, and M. Liu, "An Improved KNN based on Class Contribution and Feature Weighting," in *10th Int. Conf. Meas. Technol. Mechatron. Autom., ICMTMA 2018*, 2018, pp. 313–

316, doi: 10.1109/ICMTMA.2018.00083.

[20] Q. Yunneng, “A New Stock Price Prediction Model based on Improved KNN,” in *2020 7th Int. Conf. Inf. Sci. Control Eng., ICISCE 2020*, 2020, pp. 77–80, doi: 10.1109/ICISCE50968.2020.00026.

[21] S. A. W. Fitra A. Bachtiar, Indra K. Syahputra, “Perbandingan Algoritme Machine Learning Untuk Memprediksi,” *J. Teknol. Inf. dan Ilmu Komput.*, vol. 6, no. 5, pp. 543–548, 2019, doi: 10.25126/jtiik.2019611755.

[22] Y. Yustikasari, H. Mubarok, and R. Rianto, “Comparative Analysis Performance of K-Nearest Neighbor Algorithm and Adaptive Boosting on the Prediction of Non-Cash Food Aid Recipients,” *Sci. J. Inform.*, vol. 9, no. 2, pp. 205–217, 2022, doi: 10.15294/sji.v9i2.32369.

[23] A. Satria, O. S. Sitompul, and H. Mawengkang, “5-Fold Cross Validation on Supporting K-Nearest Neighbour Accuration of Making Consimilar Symptoms Disease Classification,” in *2nd Int. Conf. Comput. Sci. Eng.: Eff. Digit. World Pandemic (EDWAP), IC2SE 2021*, 2021, pp. 1–5, doi: 10.1109/IC2SE52832.2021.9792094.

[24] C. J. Ma and Z. sheng Ding, “Improvement of K-Nearest Neighbor Algorithm based on Double Filtering,” in *2020 5th Int. Conf. Mech. Control Comput. Eng., ICMCCE 2020*, 2020, pp. 1567–1570, doi: 10.1109/ICMCCE51767.2020.00343.

[25] M. A. Alanezi and A. Alqaddoumi, “Applying Parallel Processing to Improve the Computation Speed of K-Nearest Neighbor Algorithm,” in *2020 Int. Conf. Data Anal. Bus. Ind.: Way Towards Sustain. Econ., ICDABI 2020*, 2020, pp. 1–6, doi: 10.1109/ICDABI51230.2020.9325700.

[26] Novita, Suyanto, and Y. Prasti Eko, “Bonferroni Mean Fuzzy K-Nearest Neighbors Based Handwritten Chinese Character Recognition,” in *2021 Int. Conf. Data Sci. Appl., ICoDSA 2021*, 2021, pp. 118–123, doi: 10.1109/ICoDSA53588.2021.9617488.

[27] R. Zhao, “The Water Potability Prediction Based on Active Support Vector Machine and Artificial Neural Network,” in *2021 Int. Conf. Big Data Artif. Intell. Risk Manag., ICBAR 2021*, 2021, pp. 110–114, doi: 10.1109/ICBAR55169.2021.00032.

[28] D. A. Anggoro and D. Novitaningrum, “Comparison of Accuracy Level of Support Vector Machine (SVM) and Artificial Neural Network (ANN) Algorithms in Predicting Diabetes Mellitus Disease,” *ICIC Express Lett.*, vol. 15, no. 1, pp. 9–18, 2021, doi: 10.24507/icicel.15.01.9.

[29] P. C. Upadhyay, L. Karanam, J. A. Lory, and G. N. Desouza, “Classifying Cover Crop Residue from RGB Images: A Simple SVM versus a SVM Ensemble,” in *2021 IEEE Symposium Series on Computational Intelligence, SSCI 2021*, 2021, pp. 1–7, doi: 10.1109/SSCI50451.2021.9660147.

[30] X. Huang and S. Wang, “Prediction of Bottom-hole Flow Pressure in Coalbed Gas Wells based on GA Optimization SVM,” in *Proc. 2018 IEEE 3rd Adv. Inf. Technol. Electron. Autom. Control Conf., IAEAC 2018*, 2018, pp. 138–141, doi: 10.1109/IAEAC.2018.8577488.

[31] V. Ariyani, P. Putri, A. B. Prasetijo, and D. Eridani, “Perbandingan Kinerja Algoritme Naïve Bayes Dan K-Nearest Neighbor (Knn) Untuk Prediksi Harga Rumah,” *Transm. J. Ilm. Tek. Elektro*, vol. 4, no. 24, pp. 162–171, 2022, doi: https://doi.org/10.14710/transmisi.24.4.162-171.

[32] N. H. Othman, K. Y. Lee, A. R. M. Radzol, W. Mansor, and U. R. M. Rashid, “Classification of Salivary Adulterated NS1 SERS Spectra Using PCA-Cosine-KNN,” in *ICIIBMS 2019 - 4th Int. Conf. Intell. Inform. Biomed. Sci.*, 2019, pp. 159–163, doi: 10.1109/ICIIBMS46890.2019.8991490.

[33] R. Ramadhan and A. Triyudi, “Perbandingan Klasifikasi Data Calon Pendonor Darah Dengan Algoritma K-Nearest Neighbor Dan Support Vector Machine,” *J. Tek. Inform. dan Sist. Inf.*, vol. 9, no. 2, pp. 1250–1260, 2022, doi: 10.35957/jatisi.v9i2.2018.

[34] M. I. Perangin-angin, A. H. Lubis, and A. Ikhwan, “Association Rules Analysis on FP-Growth Method in Predicting Sales,” *Int. J. Recent Trends Eng. Res.*, vol. 3, no. 10, pp. 58–65, 2017, doi: 10.23883/ijrter.2017.3453.dhcoa.

[35] S. Wang *et al.*, “K-Means Clustering With Incomplete Data,” *IEEE Access*, vol. 7, pp. 69162–69171, 2019, doi: 10.1109/ACCESS.2019.2910287.

[36] D. Hartama, A. Perdana Windarto, and A. Wanto, “The Application of Data Mining in Determining Patterns of Interest of High School Graduates,” *J. Phys. Conf. Ser.*, vol. 1339, no. 1, p. 012042, Dec. 2019, doi: 10.1088/1742-6596/1339/1/012042.

[37] T. Yue, Y. Li, and Z. Hu, “Dwsa: An Intelligent Document Structural Analysis Model for Information Extraction and Data Mining,” *Electron.*, vol. 10, no. 19, pp. 1–16, 2021, doi: 10.3390/electronics10192443.

[38] A. Ibrahim, “Forecasting the Early Market Movement in Bitcoin Using Twitter’s Sentiment Analysis: An Ensemble-based Prediction Model,” in *2021 IEEE Int. IOT Electron. Mechatron. Conf. IEMTRONICS 2021 - Proc.*, 2021, pp. 1–5, doi:

10.1109/IEMTRONICS52119.2021.9422647.

[39] A. Kumar, A. Verma, G. Shinde, Y. Sukhdeve, and N. Lal, “Crime Prediction Using K-Nearest Neighboring Algorithm,” in *Int. Conf. Emerg. Trends Inf. Technol. Eng., ic-ETITE 2020*, 2020, pp. 1–4, doi: 10.1109/ic-ETITE47903.2020.155.

[40] Y. Jusman *et al.*, “Comparison between Support Vector Machine and K-Nearest Neighbor Algorithms for Leukemia Images Classification Using Shape Features,” in *ICSEC 2021 - 25th Int. Comput. Sci. Eng. Conf.*, 2021, pp. 70–74, doi: 10.1109/ICSEC53205.2021.9684585.