Recursive Journal of Informatics
Vol. 1, No. 2, September 2023
p-ISSN 2963-5551 https://journal.unnes.ac.id/sju/index.php/rji e-ISSN 2986-6588

Comparison of Probabilistic Neural Network (PNN) and k-Nearest Neighbor (k-NN) Algorithms for Diabetes Classification

Diah Siti Fatimah Azzahrah[1], Alamsyah[2]
[1,2]Department Computer Science, Faculty of Mathematics and Natural Sciences,
Universitas Negeri Semarang, Indonesia

Abstract.
Purpose: This study aims to compare the PNN an K-NN algorithms to determine the accuracy and the speed used for diabetes classification.
Methods: There are two algorithms used in this study, namely Probabilistic Neural Network (PNN) and k-Nearest Neighbor (k-NN). The data used is the Pima Indians Diabetes Database. The dataset contains 768 data with 8 attributes and 1 target class, namely 0 for no diabetes and 1 for diabetes. The dataset has been divided into 80% training data and 20% testing data.
Result: Accuracy is obtained after implementing k-fold cross validation (k = 4). The accuracy results show that the k-NN algorithm is superior and has better quickness compared to the PNN. The k-NN algorithm obtains an accuracy of 74.6% for all features and 78.1% for four features.
Novelty: The novelty of this paper is optimizing and improving accuracy which is implemented with by focusing on data preprocessing, feature selection, and k-fold cross validation in the classification algorithm.

Keywords: Data Mining, Feature Selection, K-Fold Cross Validation, k-Nearest Neighbor, Probabilistic Neural Network
Received February 06, 2023 / **Revised** February 09, 2023 / **Accepted** September 14, 2023

INTRODUCTION

Diabetes mellitus is a metabolic disorder characterized by a prolonged increase in blood sugar levels beyond normal limits [1]. Initially, diabetes mellitus does not show any obvious symptoms. However, when it is detected late in its treatment there is a risk of complications [2]. Complications that could happen contain cardiovascular disease, stroke, kidney failure, eye, and other disease [3]. There are three types of diabetes mellitus including type 1 diabetes, type 2 diabetes, and type 3 gestational diabetes [4]. The World Health Organization (WHO) says that the number of diabetes patients has increased from 108 million to 422 million in 2014. Estimates show that in 2045 this number will reach 629 million. In 2016, there were reports of around 1.6 million people dying due to diabetes [5].

Diabetes is a major health problem in the world. Because the number of cases of diabetes continues to increase from year to year. In this case, of course, it is necessary to find a solution to be able to predict diabetes. The existence of a diabetes predictor tool can make it preventive for patients to find out whether the disease is present early. It can also prevent diabetes from getting worse. Diabetes can be detected using historical patient data containing information about a patient's symptoms or conditions [6]. Nowadays, technology has growing rapidly. With current technological advances, a disease will be detected more quickly through these symptoms [7]. In the health sector, there is a lot of data or big data can be processed and produce new information. The processing of data that can be extracted as information or knowledge pattern from sets of data, then it is called data mining approach [8]. Solving these problems can be implemented with current technological developments, namely data mining. Data mining is a technique to process a data to find the patterns in a specific domain. Moreover, it is also finding relationships between the attributes of data samples using statistical and computer science technique [9].

[1]*Corresponding author.
Email addresses: diah.azzahrah@students.unnes.ac.id (Diah)
DOI: 10.15294/rji.v1i2.66078

This data mining method can be used in the world of health in making models to predict and classify the health problems that are being faced. Classification is one of the most frequently used types of data mining. The decision in data mining can be predicted by using classification technique [8]. Classification is used to classify data into predefined categorical class labels. In classifying data, the classification algorithm will create a classification model. This technique is a two-step process consisting of training and testing. The classification technique has been used to determine the medical diagnosis and prognosis of person with the symptoms and health conditions dataset [10]. Prediction is an important matter for knowing future events by recognizing patterns of past events. By knowing the events that will occur, everyone will be more careful and of course prepare for all possibilities both in life and related to property [11].

As for research with diabetes objects conducted by [12], using six algorithms namely Support Vector Machine (SVM) , Random Forest (RF), Logistic Regression (LR), K-Nearest Neighbor (KNN), Decision Tree (DT), and Naïve Bayes (NB). SVM and KNN obtained the highest accuracy of 77%. Then there are limitations to research on diabetes, namely the size of the dataset and the missing attribute values. Further research from [13], using Decision Tree (DT), Discriminant Analysis (DA), Logistic Regression (LR), Support Vector Machine (SVM), k-Nearest Neighbors (k-NN), and ensemble learners. The highest accuracy results by LR with an average accuracy of 77.9%. The research suggests using the feature selection method to improve classification accuracy. Then, research [14], using Decision Tree (DT), Support Vector Machine (SVM), and Naïve Bayes (NB) algorithms. The highest accuracy by the NB algorithm is 76,30%. This study didn't use cross validation. In addition, accuracy can be improved by using other machine learning algorithms. Then, research from [15] uses four algorithms namely Naïve Bayes (NB), Support Vector Machine (SVM), Random Forest (RF), and Simple CART. This study didn't implement cross validation techniques. Furthermore, research from [16] uses the J48 Decision Tree, Random Forest (RF), and Naïve Bayes (NB) algorithms. In this research, it is necessary to increase accuracy by using appropriate preprocessing techniques for data management and analysis.

Previous studies have explained that the classification system can handle predicting a problem, including predicting a disease. In this study, two algorithms will be used, namely Probabilistic Neural Network (PNN) and k-Nearest Neighbor (k-NN). Probabilistic Neural Network and k-Nearest Neighbor algorithms are included in the algorithms used in classification problems. Probabilistic Neural Network and k-Nearest Neighbor algorithms are included in the algorithms used in classification problems. Probabilistic Neural Network algorithm has faster training compared to Multilayer Perceptron Network [17]. Then, k-Nearest Neighbor (k-NN) algorithm is able to classify datasets using a training model similar to the test query by calculating the closest k training data points (neighbors) to the query being tested [18]. The purpose of this study is to compare the PNN and k-NNalgorithms in terms of the accuracy and speed of the two algorithms used for diabetes classification. Focus in this study on data preprocessing with filling the data with a value of 0 for the mean or median value. In addition, the final results of this study are the application of feature selection of the Pearson correlation type, and the k-fold cross validation test. So that it can be seen which algorithm produces better accuracy and speed in classifying diabetes

METHODS

This study focuses on comparing the accuracy of the PNN and k-NN classification algorithms for classification of diabetes. This accuracy comparison is accompanied by focusing on data preprocessing, and the use of feature selection and k-fold cross validation. The process starts from input dataset, data preprocessing, feature selection, data split, modeling classification algorithms, and model testing. The process stages can be seen in Figure 1.

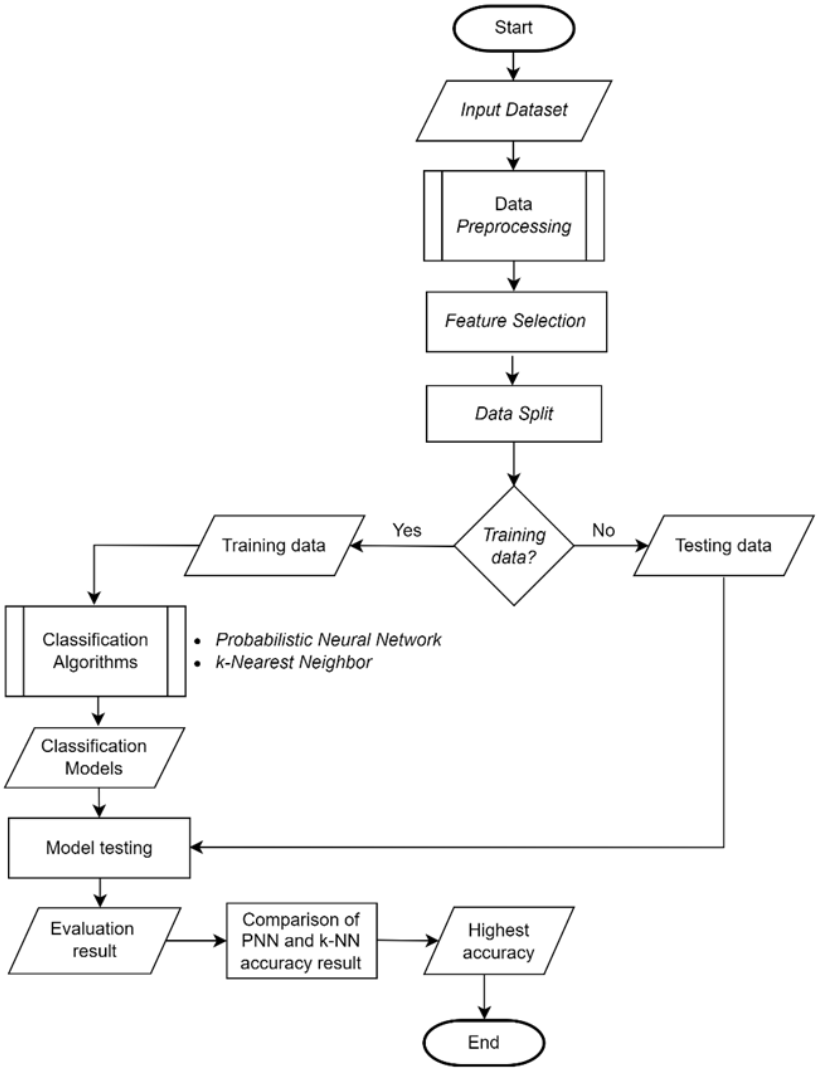

Figure 1. Flowchart Research Design

Dataset

The dataset that used is the dataset found on the Kaggle website. This study took a dataset from the Pima Indians Diabetes Database which can be downloaded via https://www.kaggle.com/datasets/uciml/pima-indians-diabetes-database. This dataset has a total of 768 data with 8 attributes and 1 target class [19]. The dataset can be seen in Table 1.

Table 1. Dataset Pima Indians

Variables	Description
Pregnancies	Number of times pregnant
Glucose	Plasma glucose concentration 2 hours in an oral glucose tolerance test
BloodPressure	Diastolic blood pressure
SkinThickness	Triceps skinfold thickness
Insulin	2-hour serum insulin
BMI	Body Mass Index
DiabetesPedigreeFunction	Diabetes pedigree function
Age	Age (years)
Outcomes	Class variables (0 or 1)

Data Preprocessing

This stage is an initial technique in data mining and data analysis. This process will transform or process raw data into useful and efficient formats and information. It aims to be understood and analysed by computers

in machine learning. The format of the raw data that taken from various dataset sources, often have experiences errors, missing values, and inconsistent data. Data preprocessing consists of the following stages.

1. Data cleaning includes dealing with missing values, removing noise and irrelevant data [20]. In this case, duplicate data will be removed, overcoming the outliers in the variables, and filling in values that are 0 with the median or mean in each variable.
2. Data integration is a process of combining data from several existing sources. Data integration is only done if the data comes from different places [21].
3. Data selection selects the data attributes to be used so that the data can be processed according to the needs of the data mining stages [21].
4. Data transformation will change the shape and format of the data in a form that is suitable for the data mining process [20]. The technique used at these stages in this study is the StandardScaler.

Feature Selection

Feature selection is a preprocessing technique that exists in data mining to select relevant features and reduce data by eliminating unnecessary attributes. The goals of feature selection include getting better predictive performance, speeding up the prediction process, reducing costs, and understanding the process of obtaining data better [22]. In this study, the Pearson correlation will be used in the feature selection process. Pearson correlation is a statistical matrix that calculates the strength and linear relationship between two random variables. This process has been applied to various indexes in statistics including data analysis, classification, clustering, decision making, financial analysis, biological research and others [23]. This Pearson correlation method works to calculate the correlation or relationship between variables in the dataset used.

Data Split

This split data stage is carried out after the data preprocessing has been completed. Split data will be divided into datasets such as training data and testing data. Training data is used as a data pattern in the formation of data mining models. While data testing is a stage that is used after the training process is complete. Data testing is used to carry out tests in applying the classification model. This study uses the composition of data division into training data and testing data, namely 80% for training data and 20% for testing data according to research conducted by [24].

Modeling Classification Algorithms

The classification stage consists of two classification processes using the PNN and k-NN algorithms. The classification technique has two stages of the process consisting of the learning stage and the classification stage. The learning stage is a classification model built or created. While the classification stage, namely the model is used to predict class labels in the given data [25]. This techniques is often used in the world of health, one of which is to predict disease [22].

In this research, the Probabilistic Neural Network algorithm provides a solution to the problem of pattern classification by following the approach developed in statistics, which is called a Bayesian classifier. The network paradigm also uses the Parzen estimations which adds up to builds density probabilities function required by Bayes theorem [26]. In addition, PNN has a structure consisting of four layers, namely the input layer and three information processing layers starting from the pattern layer to the summation layer, and continuing to the output layer [27].

As for the k-Nearest Neighbor (k-NN) algorithm is used in solving problems related to regression and classification [3]. Classification in the k-Nearest Neighbor algorithm is carried out based on the distance between training data and testing data. This distance can be calculated using the Euclidean distance. Based on the similarity between training data and testing data, the k nearest neighbors are selected. At the input the value of k is a positive integer. The value of k is a variable regarding the number of nearest neighbors. The label associated with the neighbor is taken as a reference.

Testing data is associated with the class that has the majority among the k nearest neighbors [28]. The stages of the classification of the PNN and k-NN in this study can be seen in Figure 2 and Figure 3.

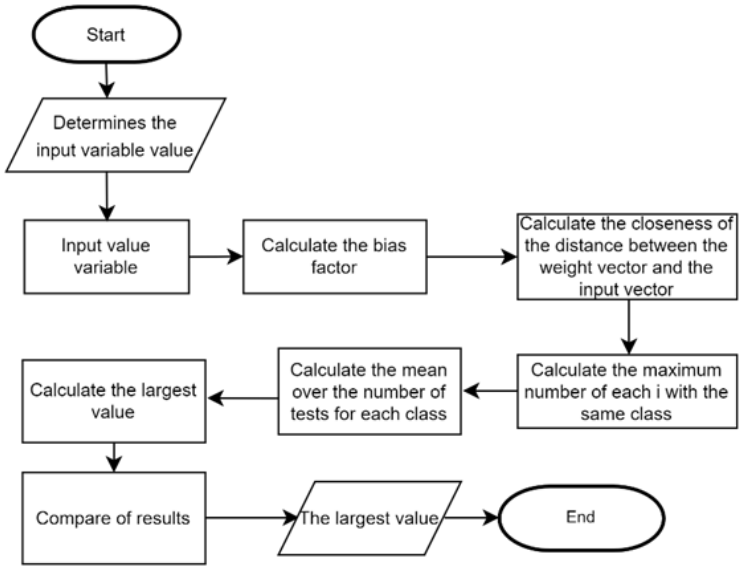

Figure 2. Flowchart Probabilistic Neural Network

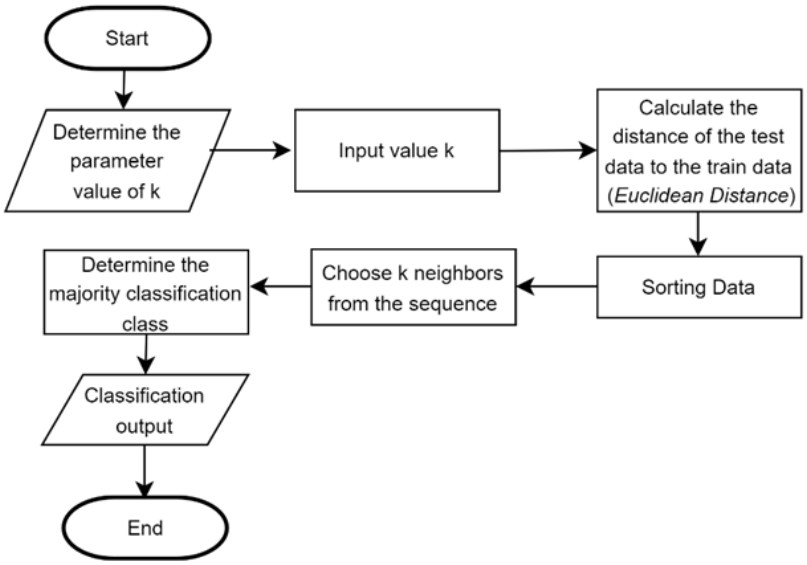

Figure 3. Flowchart k-Nearest Neighbor

Model Testing

In this study, model testing was carried out using the accuracy of the confusion matrix and k-fold cross validation with a the number of k = 4. In machine learning, the performance of an algorithm can be show by using the confusion matrix. Confusion matrix is a special table layout that allows visualize the performance of each class category [29]. Calculation of accuracy using Equation 1.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (1)$$

Accuracy is the ratio between data that is correctly classified and the total sample. The accuracy calculation in Equation 1 has several variables where TP is a True Positive value, TN is a True Negative value, FP is a False Positive value and FN is a False Negative value. So, the sum of TP and TN divided by the total number of TP, TN, FP, and FN.

RESULT AND DISCUSSION

This study applies feature selection with Pearson correlation and k-fold cross validation for diabetes classification using the PNN and k-NN classification algorithms. After the data is retrieved, the next process is data preprocessing, feature selection, and data split. The Pima Indians dataset is called which can be seen in Figure 4.

	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age	Outcome
0	6	148	72	35	0	33.6	0.627	50	1
1	1	85	66	29	0	26.6	0.351	31	0
2	8	183	64	0	0	23.3	0.672	32	1
3	1	89	66	23	94	28.1	0.167	21	0
4	0	137	40	35	168	43.1	2.288	33	1
...	...	...	...	...	...	...	...	...	...
763	10	101	76	48	180	32.9	0.171	63	0
764	2	122	70	27	0	36.8	0.340	27	0
765	5	121	72	23	112	26.2	0.245	30	0
766	1	126	60	0	0	30.1	0.349	47	1
767	1	93	70	31	0	30.4	0.315	23	0

768 rows × 9 columns

Figure 4. Calling dataset Pima Indians

The next process is data preprocessing, such as data cleaning, data integration, data selection and data transformation. Result from data preprocessing can be seen in Figure 5.

	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age	Outcome
0	0.647150	0.865276	-0.019315	0.936123	-0.515442	0.181733	0.588927	1.445691	1
1	-0.848970	-1.205989	-0.531737	0.277236	-0.515442	-0.868800	-0.378101	-0.189304	0
2	1.245598	2.015979	-0.702545	-0.652184	-0.515442	-1.364051	0.746595	-0.103252	1
3	-0.848970	-1.074480	-0.531737	-0.381651	-0.211679	-0.643686	-1.022787	-1.049828	0
4	-1.148194	0.503626	-2.752234	0.936123	1.371253	1.607456	2.596563	-0.017199	1
...	...	...	...	...	...	...	...	...	...
763	1.844045	-0.679954	0.322300	2.363711	1.627945	0.076680	-1.008772	2.564372	0
764	-0.549746	0.010468	-0.190122	0.057607	-0.515442	0.661976	-0.416642	-0.533513	0
765	0.347926	-0.022409	-0.019315	-0.381651	0.173359	-0.928831	-0.749497	-0.275356	0
766	-0.848970	0.141977	-1.044160	-0.652184	-0.515442	-0.343534	-0.385109	1.187534	1
767	-0.848970	-0.942972	-0.190122	0.496865	-0.515442	-0.298511	-0.504236	-0.877723	0

768 rows × 9 columns

Figure 5. Result from data preprocessing

Then, result of feature selection with the Pearson correlation which can be seen in Figure 6.

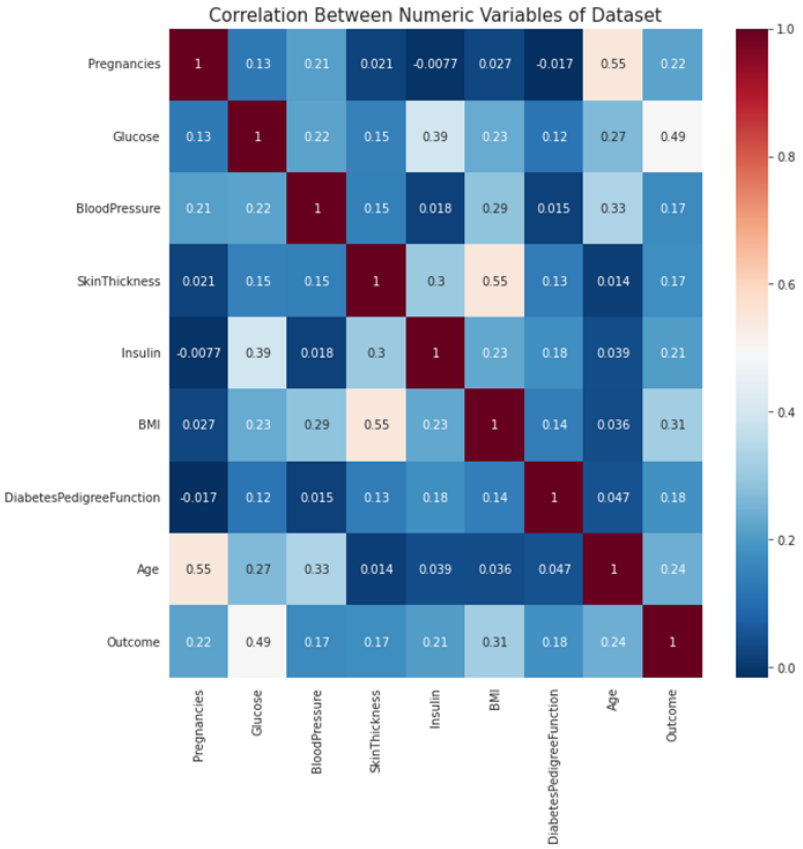

Figure 6. Result of feature selection

Figure 5 show that the four features that have a high correlation include pregnancies, glucose, BMI, and age. Therefore, this study will use these four features and all features to compare the accuracy results. After that, the dataset division of 80% training data and 20% testing data is 614 data for training data and 154 data for testing data. The results of the accuracy of the Probabilistic Neural Network and k-Nearest Neighbor on the implementation of all features with k-fold cross validation, namely the value of k is 4 can be seen in Table 2.

Table 2. Accuracy of all features

Test	All Features Classification Algorithms	
	Probabilistic Neural Network (PNN)	k-Nearest Neighbor (k-NN)
1	75.5 %	78.1 %
2	73.9 %	70.3 %
3	77.0 %	77.6 %
4	70.8 %	72.3 %
Means	74.3 %	74.6 %

Then, the results of the accuracy of the Probabilistic Neural Network and k-Nearest Neighbor algorithms with the application of four features and k-fold cross validation, namely the value of k is 4 can be seen in Table 3.

Table 3. Accuracy of four features

Test	Four Features Classification Algorithms	
	Probabilistic Neural Network (PNN)	k-Nearest Neighbor (k-NN)
1	76.5 %	78.6 %
2	72.9 %	79.6 %
3	77.0 %	76.5 %
4	76.0 %	77.6 %
Means	75.7 %	78.1 %

Looking at Table 2 and Table 3, the final results of the comparison of the average accuracy of the k-fold cross validation on the two classification algorithms are illustrated in tabular form which can be seen in Table 4.

Table 4. Accuracy Final Results

Accuracy	Classification Algorithms	
	Probabilistic Neural Networks (PNN)	k-Nearest Neighbor (k-NN)
All features	74.3 %	74.6 %
4 features	75.7 %	78.1 %

Based on the accuracy value obtained from each algorithm. Both of these algorithms produce accuracy values above 70% by implementing k-fold cross validation on all features and four features. The accuracy value of the k-Nearest Neighbor (k-NN) algorithm is superior to that of the Probabilistic Neural Network (PNN). So, the k-Nearest Neighbor algorithm is able to work well in the classification of diabetes. Then the comparison of this study with previous research can be seen in Table 5.

Table 5. The related research

Writer	**Datasets**	**Algorithms**	**Accuracy Results**
Sarwar et al (2018)	Pima Indians	Support Vector Machine, K-Nearest Neighbor, Logistic Regression, Decision Tree, Random Forest, Naïve Bayes	SVM and KNN = 77 %
Al-Zebari & Sengur (2019)	Pima Indians	Decision Tree, Discriminant Analysis, Logistic Regression, Support Vector Machines, k-Nearest Neighbor, Ensemble Learners → 24 classifiers	Logistic Regression = 77.9 %
Sisodia & Sisodia (2018)	Pima Indians	Decision Tree, Support Vector Machine, Naïve Bayes	Naïve Bayes = 76.30 %
Proposed Method	Pima Indians	Probabilistic Neural Network, k-Nearest Neighbor	k-fold accuracy with 4 features k-NN = 78.1 %, all features k-NN = 74.6 %

Based on related research, the strengths of this study are that using the PNN and k-NN algorithms can produce optimal accuracy and can classify diabetes properly. The results of increasing and optimizing accuracy are supported by focusing on data preprocessing, implementing feature selection and k-fold cross validation. In addition, the classification quickness of the Probabilistic Neural Network (PNN) has a faster training process, including the k-Nearest Neighbor algorithms. However, this research has a drawback, namely the dataset used only has data in the number and attributes that are not large enough. Furthermore, the k-Nearest Neighbor algorithm has a weakness regarding sensitivity to noise data, missing values, and outliers so that handling is needed to overcome this in research to optimize the accuracy obtained. As for the Probabilistic Neural Network algorithm, it is necessary to do try and error to determine the parameters.

CONCLUSION

This study describes the classification algorithm of the PNN and k-NN by applying feature selection and k-fold cross validation in the classification of diabetes. The aim is to determine the accuracy and quickness resulting from the two algorithms. Increasing and optimizing accuracy is supported by focusing on data preprocessing, implementing feature selection, and k-fold cross validation. The value of k from k-fold cross validation is 4. The results obtained show that the quickness of the Probabilistic Neural Network algorithm only takes a short time in the training process. The quickness of k-nearest neighbor has fast, and effective training used on large training data. However, in this study the quickness of the k-Nearest Neighbor algorithm is superior. The accuracy obtained 74.6 % for all features and 78.1 % for the four features of the k-Nearest Neighbor algorithm. Then, the accuracy obtained by the Probabilistic Neural Network is 74.3 % for all features and 75.7 % for four features. However, these results cannot be tested directly in the health world because the error value is still above 0.5%. Future research can use the same algorithm or different algorithm but add the combination method and the same dataset and implement different types of feature selection. In addition, subsequent research, can use datasets with larger data.

REFERENCES

[1] M. Maniruzzaman, M. J. Rahman, B. Ahammed, and M. M. Abedin, "Classification and Prediction of Diabetes Disease using Machine Learning Paradigm," *Heal. Inf. Sci. Syst.*, vol. 8, no. 7, pp. 1–14, 2020, doi: 10.1007/s13755-019-0095-z.

[2] A. Viloria, Y. Herazo-Beltran, D. Cabrera, and O. B. Pineda, "Diabetes Diagnostic Prediction Using Vector Support Machines," in *Procedia Computer Science*, 2020, vol. 170, pp. 376–381. doi: 10.1016/j.procs.2020.03.065.

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

[4] P. Rajendra and S. Latifi, "Prediction of Diabetes using Logistic Regression and Ensemble Techniques," *Comput. Methods Programs Biomed. Updat.*, vol. 1, no. 100032, pp. 1–8, 2021, doi: 10.1016/j.cmpbup.2021.100032.

[5] F. Anwar, Qurat-UI-Ain, M. Y. Ejaz, and A. Mosavi, "A comparative analysis on diagnosis of diabetes mellitus using different approaches – A survey," *Informatics Med. Unlocked*, vol. 21, no. 100482, pp. 1–10, 2020, doi: 10.1016/j.imu.2020.100482.

[6] D. S. F. Azzahrah and Alamsyah, "Klasifikasi Penyakit Diabetes Menggunakan Algoritma K-Nearest Neighbor," in *Seminar Nasional Ilmu Komputer (SNIK 2022)*, 2022, pp. 70–75. [Online]. Available: https://conf.unnes.ac.id/index.php/snik/snik2022/paper/view/610/499

[7] J. L. Handarko and Alamsyah, "Implementasi Fuzzy Decision Tree Untuk Mendiagnosa Penyakit Hepatitis," *Unnes J. Math.*, vol. 4, no. 2, pp. 157–164, 2015.

[8] H. A. Prihanditya and Alamsyah, "The Implementation of Z-Score Normalization and Boosting Techniques to Increase Accuracy of C4.5 Algorithm in Diagnosing Chronic Kidney Disease," *J. Soft Comput. Explor.*, vol. 1, no. 1, pp. 63–69, 2020, doi: 10.52465/joscex.v1i1.8.

[9] S. M. Birjandi and S. H. Khasteh, "A Survey on Data Mining Techniques used in Medicine," *J. Diabetes Metab. Disord.*, vol. 20, pp. 2055–2071, 2021, doi: 10.1007/s40200-021-00884-2.

[10] I. Yoo *et al.*, "Data Mining in Healthcare and Biomedicine: A Survey of the Literature," *J. Med. Syst.*, vol. 36, pp. 2431–2448, 2012, doi: 10.1007/s10916-011-9710-5.

[11] Z. S. Hikmawati, R. Arifudin, and A. Alamsyah, "Prediction The Number of Dengue Hemorrhagic Fever Patients Using Fuzzy Tsukamoto Method at Public Health Service of Purbalingga," *Sci. J. Informatics*, vol. 4, no. 2, pp. 115–124, 2017, doi: 10.15294/sji.v4i2.10342.

[12] M. A. Sarwar, N. Kamal, W. Hamid, and M. A. Shah, "Prediction of Diabetes Using Machine Learning Algorithms in Healthcare," in *2018 24th International Conference on Automation and Computing (ICAC)*, 2018, pp. 1–6. doi: 10.23919/IConAC.2018.8748992.

[13] A. Al-Zebari and A. Sengur, "Performance Comparison of Machine Learning Techniques on Diabetes Disease Detection," *2019 1st Int. Informatics Softw. Eng. Conf.*, pp. 1–4, 2019, doi: 10.1109/UBMYK48245.2019.8965542.

[14] D. Sisodia and D. S. Sisodia, "Prediction of Diabetes using Classification Algorithms," *Procedia Comput. Sci.*, vol. 132, no. Iccids, pp. 1578–1585, 2018, doi: 10.1016/j.procs.2018.05.122.

[15] A. Mir and S. N. Dhage, "Diabetes Disease Prediction Using Machine Learning on Big Data of Healthcare," in *2018 4th International Conference on Computing, Communication Control and Automation (ICCUBEA)*, 2018, pp. 1–6. doi: 10.1109/ICCUBEA.2018.8697439.

[16] V. Chang, J. Bailey, Q. A. Xu, and Z. Sun, "Pima Indians Diabetes Mellitus Classification Based on Machine Learning (ML) Algorithms," *Neural Comput. Appl.*, 2022, doi: 10.1007/s00521-022-07049-z.

[17] L. M. Raposo, M. B. Arruda, R. M. de Brindeiro, and F. F. Nobre, "Lopinavir Resistance Classification with Imbalanced Data Using Probabilistic Neural Networks," *J. Med. Syst.*, vol. 40, no. 69, pp. 1–7, 2016, doi: 10.1007/s10916-015-0428-7.

[18] S. Uddin, I. Haque, H. Lu, M. A. Moni, and E. Gide, "Comparative Performance Analysis of K-Nearest Neighbour (KNN) Algorithm and its Different Variants for Disease Prediction," *Sci. Rep.*, vol. 12, no. 6256, pp. 1–11, 2022, doi: 10.1038/s41598-022-10358-x.

[19] U. M. Learning, "Pima Indians Diabetes Database," *Kaggle*. https://www.kaggle.com/datasets/uciml/pima-indians-diabetes-database

[20] H. Sulastri and A. I. Gufroni, "Penerapan Data Mining Dalam Pengelompokan Penderita Thalassaemia," *J. Nas. Teknol. dan Sist. Inf.*, vol. 3, no. 2, pp. 299–305, 2017, doi: 10.25077/teknosi.v3i2.2017.299-305.

[21] R. R. Rerung, "Penerapan Data Mining dengan Memanfaatkan Metode Association Rule untuk Promosi Produk," *J. Teknol. Rekayasa*, vol. 3, no. 1, p. 89, 2018, doi: 10.31544/jtera.v3.i1.2018.89-98.

[22] F. F. Firdaus, H. A. Nugroho, and I. Soesanti, "A Review of Feature Selection and Classification Approaches for Heart Disease Prediction," *IJITEE (International J. Inf. Technol. Electr. Eng.*, vol. 4, no. 3, pp. 75–82, 2020, doi: 10.22146/ijitee.59193.

[23] I. A. Nikmatun and I. Waspada, "Implementasi Data Mining untuk Klasifikasi Masa Studi Mahasiswa Menggunakan Algoritma K-Nearest Neighbor," *J. SIMETRIS*, vol. 10, no. 2, pp. 421–432, 2019, [Online]. Available: https://jurnal.umk.ac.id/index.php/simet/article/view/2882/1855

[24] A. Gholamy, V. Kreinovich, and O. Kosheleva, "Why 70/30 or 80/20 Relation Between Training and Testing Sets : A Pedagogical Explanation," *Dep. Tech. Reports*, pp. 1–6, 2018.

[25] J. Han, M. Kamber, and J. Pei, *Data Mining: Concepts and Techniques*, 3rd ed. San Fransisco: Morgan Kaufmann, 2012. [Online]. Available: https://doi.org/10.1016/C2009-0-61819-5

[26] M. S. Bascil and H. Oztekin, "A Study on Hepatitis Disease Diagnosis Using Probabilistic Neural Network," *J. Med. Syst.*, vol. 36, pp. 1603–1606, 2012, doi: 10.1007/s10916-010-9621-x.

[27] S. J. Siregar, A. I. Lubis, and E. F. Ginting, "Penerapan Neural Network Dalam Klasifikasi Citra Permainan Batu Kertas Gunting dengan Probabilistic Neural Network," *Build. Informatics, Technol. Sci.*, vol. 3, no. 3, pp. 420–425, 2021, doi: 10.47065/bits.v3i3.1143.

[28] A. Giri, M. V. V. Bhagavath, B. Pruthvi, and N. Dubey, "A Placement Prediction System using K-Nearest Neighbors Classifier," in *2016 Second International Conference on Cognitive Computing and Information Processing (CCIP)*, 2016, pp. 1–4. doi: 10.1109/CCIP.2016.7802883.

[29] N. Yilmaz, O. Inan, and M. S. Uzer, "A New Data Preparation Method Based on Clustering Algorithms for Diagnosis Systems of Heart and Diabetes Diseases," *J. Med. Syst.*, vol. 38, no. 48, pp. 1–12, 2014, doi: 10.1007/s10916-014-0048-7.