Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 6, Number 1, October 2021
DOI : https://doi.org/10.33395/sinkron.v6i1.11234
e-ISSN : 2541-2019
p-ISSN : 2541-044X

Analysis of dimensional reduction effect on K-Nearest Neighbor classification method

Taufiqurrahman[1)], Erna Budhiarti Nababan[2)*], Syahril Efendi[3)]
[1)2)3)] Graduate Program of Computer Science, Universitas Sumatera Utara, Medan, Indonesia
[1)] tfqrrhmn@hotmail.com, [2)] ernabrn@usu.ac.id,[3)] syahril1@usu.ac.id

Submitted : Nov 14, 2021| **Accepted** : Dec 20, 2021 | **Published** : Dec 23, 2021

Abstract: Classification algorithms mostly become problematic on data with high dimensions, resulting in a decrease in classification accuracy. One method that allows classification algorithms to work faster and more effectively and improve the accuracy and performance of a classification algorithm is by dimensional reduction. In the process of classifying data with the K-Nearest Neighbor algorithm, it is possible to have features that do not have a matching value in classifying, so dimension reduction is required. In this study, the dimension reduction method used is Linear Discriminant Analysis and Principal Component Analysis and classification process using KNN, then analyzed its performance using Matrix Confusion. The datasets used in this study are Arrhythmia, ISOLET, and CNAE-9 obtained from UCI Machine Learning Repository. Based on the results, the performance of classifiers with LDA is better than with PCA on datasets with more than 100 attributes. Arrhythmia datasets can improve performance on K-NN K=3 and K=5. The best performance is obtained by LDA+K-NN K=3 which produces an accuracy value of 98.53%, the lowest performance found in K-NN without reduction with K=3. ISOLET datasets, the best performance results are also obtained by data that has been reduced with LDA, but the best performance is obtained when the classification of K-NN with K=5 and the lowest performance is found in PCA+ K-NN with a value of K=3. As for the best performance, dataset CNAE-9 is also achieved by LDA+K-NN, while the lowest performance is PCA+K-NN with the value of K=3.

Keywords: Dimension Reduction, LDA, PCA, Confusion Matrix, K-NN

INTRODUCTION

Classification algorithms mostly become problematic in data with high dimensions, resulting in a decrease in classification accuracy. One method that allows classification algorithms to work faster and effectively and improve the accuracy and performance of a classification algorithm is by dimensional reduction. Dimensional reduction can eliminate irrelevant features, reduce noise, and reduce the curse of dimensionality(Hasdyna, Nababan, & Effendi, 2019). Dimensional reduction can also reduce the amount of time and memory required by classification algorithms. One algorithm that can be used in classification is the K-NN algorithm. The K-NN method is one of the most widely used methods in data mining and machine learning research, such as text categorization, pattern recognition, and classification (Syaliman, Nababan, & Sitompul, 2018). Classifying data with the K-NN algorithm allows for features that do not have a matching value in the classification, so there is a need for dimension reduction. The number of features that affect computing time, the more features used, the more it increases the computing time required (Rosadi, Sanjaya, & Hakim, 2018).

Previous research (Cahyani, Wiryasaputra, & Gustriansyah, 2018) analyzing Linear Discriminant Analysis in identifying handwritten capital letters, LDA, and Euclidean Distance can gain accuracy of 75.39% and a computational time of 0.41999 seconds. Other related studies (Hariadi, Rambu, & Enda, 2019) face detection using the Linear Discriminant Analysis (LDA) and Support Vector Machine (SVM) methods prove that Linear Discriminant Analysis can improve the accuracy performance of classification algorithms. Another study conducted by (Budiman, Santoso, & Afirianto, 2017) in detecting this type of autism in early childhood using the Linear Discriminant Analysis (LDA) method obtained good accuracy results, using 75 training data, can produce

*name of corresponding author

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 6, Number 1, October 2021
DOI : https://doi.org/10.33395/sinkron.v6i1.11234
e-ISSN : 2541-2019
p-ISSN : 2541-044X

an accuracy value of 88%. For PCA, previous research (Wibawa & Novianti, 2017) showed that classification results showed the highest accuracy achieved from PCA use with K-NN, which was 0.9736. Other PCA studies(Lubis, Sihombing, & Nababan, 2020) modifications to the K-NN and PCA methods yielded an average accuracy of 88%, where the value of K was at K = 3 to K = 9. And another study (Suyanto, Siregar, Nababan, & Fikri, 2020) showed that the data used in the study of 2,098 complete blood test results taken from one of the hospitals in Medan and classified using K-NN resulted in a classification accuracy of 92%.

Based on previous research, the authors used K-NN to analyze the dimension reduction methods to be used in the study. The focus of the study was to analyze the performance of K-NN algorithms using reduced datasets compared to non-reduced datasets by measuring accuracy, precision, recall, F1-SCORE, and Matthew Correlation Coefficient (MCC).

LITERATURE REVIEW

Dataset

The dataset used in the study is data taken from UCI Machine Learning that can be accessed in https://archive.ics.uci.edu/ml/index.php. Datasets have more than 100 attributes to see the effect of dimensional reduction on datasets with many attributes. Each dataset will be reduced to the provision of the number of classes minus one (n-1). The data specifications used can be seen in table 1.

Table 1. Dataset Specifications

Dataset	Data	Attribute	Class
Arrhythmia	452	280	12
ISOLET	7797	618	26
CNAE-9	1080	857	9

StandardScaler

The Standard Scaler assumes data is usually distributed within each feature and scales them such that the distribution is centered around 0, with a standard deviation of 1.

$$Z = \frac{(X-\mu)}{\sigma} \quad (1)$$

with μ is the sample mean and σ is the standard deviation of the sample.

Principal Component Analysis (PCA)

The first steps of PCA work are to collect n from m −dimension data $\underset{x_1}{\rightarrow} \ldots \underset{x_n}{\rightarrow}$ vector x_n in $\mathbb{R}$ m. Then the reduction of the mean (average) of each dimension of the data with the following formula :

$$\underset{\mu}{\rightarrow} = \frac{1}{N_i}\left(\underset{x_1}{\rightarrow} + \cdots + \underset{x_n}{\rightarrow}\right) \quad (2)$$

Where N is the number of samples or the number of observational data, after the reduction of the mean, the form of the data matrix (B) and the $covariance$ (S) with B is a matrix measuring $m \times n$, with the i column $\underset{x_1}{\rightarrow} - \underset{\mu}{\rightarrow}$. To form a data matrix, use the formula :

$$B = \left|\underset{x_1}{\rightarrow} - \underset{\mu}{\rightarrow}\right| \ldots \left|\underset{x_n}{\rightarrow} - \underset{\mu}{\rightarrow}\right| \quad (3)$$

As for forming the covariance matrix (which is $m \times m$ in size), use the formula:

$$S = \frac{1}{n-1} BB^T \quad (4)$$

By n is the number of samples or the number of observational data. The next step is, calculate eigenvectors $\underset{\mu_1}{\rightarrow} \ldots \underset{\mu_m}{\longrightarrow}$ and eigenvalues $\underset{\lambda_1}{\rightarrow} \ldots \underset{\lambda_m}{\longrightarrow}$ of the Covariance S matrix. Then sort eigenvectors by eigenvalues value from largest to lowest and select k eigenvectors with the largest eigenvalues to form matrix W with dimensions $m \times k$ (where each column presents eigenvectors). After that, form a new dataset. This new dataset was obtained with a formula:

$$Y = v_{row} * B \quad (5)$$

Where v_{row} obtained from Feature Vector^T and Y is the final data set.

To determine how many eigenvectors were selected, the study used a cumulative proportion of variance (eigenvalues) to total variance (eigenvalues). The proportion of variance indicates the percentage of information

*name of corresponding author

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 6, Number 1, October 2021
DOI : https://doi.org/10.33395/sinkron.v6i1.11234
e-ISSN : 2541-2019
p-ISSN : 2541-044X

of the original variables contained in each feature vector (eigenvector) based on the eigenvalues. It provides an interpretation of how much data can be represented in reduced dimensions (Ma & Wisesty, 2018). The proportion of variances for each major component (eigenvector) can be calculated using the following formulas:

$$\left(\frac{\lambda_i}{\sum\lambda_i} \times 100\%\right) + \left(\frac{\lambda_{i-1}}{\sum\lambda_i} \times 100\%\right) \quad (6)$$

Where λ_i is the eigenvalue and λ_{i-1} is the previous eigenvalue. Eigenvectors are selected based on thresholds (criteria for the selection of eigenvectors or main components) against predefined PPV results. The eigenvector selection formula is as follows:

$$\frac{\lambda_i}{\sum\lambda_i} \times 100\% > threshold \quad (7)$$

Linear Discriminant Analysis (LDA)

LDA is one of the methods used for pattern recognition in statistical calculations by finding linear projections of data that will maximize the distance between classes and minimize the distance of data that has similarities (Hana, 2020).

Although the most popular method for characteristic extraction is Principal Component Analysis (PCA), PCA has the disadvantage of separating between classes that are less than optimal, so the LDA method is made to overcome the shortage of PCA. The LDA method can separate data between classes by maximizing the value of between-class scatter and minimizing Within-class scatter. PCA and LDA have very clear differences because classifying traits can be done by PCA, while LDA focuses on classifying data. In the extraction of features using LDA, the location data set is fixed, but the classes formed to become more separate so that this condition causes the distance between classes to be greater, while the distance between training data in one class becomes smaller. The number of features produced by the LDA is calculated from the number of classes minus one. In other words, the number of features the LDA produces depends on the number of classes and, some number of poses that the LDA has trained, and the LDA does not affect the number of features produced so it will take less time during the feature extraction process as well as the image recognition process. The steps of the trait extraction process using LDA (Cahyani et al., 2018) the first is to convert a two-dimensional matrix into a single dimension or a vector row or column vector. Then group the training data into a matrix of several classes (xi) and calculate the mean of each class (μi).

Calculation of the mean value of each class can be calculated by formula (8).

$$\mu i = \frac{1}{N_i}\sum_{x\in\omega i} x \quad (8)$$

The dimension mean calculation uses a column model if each training data is transformed into a row vector. If the training data is transformed into column form, the dimension mean is calculated based on rows. The number of mean dimensions generated is equal to the number of dimensions one training data instead of the number of datasets.

Calculate the total mean value of all classes (μ). The formula can calculate the calculation of the total mean value of the entire class:

$$\mu = \frac{1}{N_i + \cdots + N_c}\sum_{x\in\omega i} x \quad (9)$$

Then calculate the Matrix Between Class Scatter (S_B) and The Matrix Within Class Scatter (S_w). Matrix calculations S_B can be calculated by formula (10) while S_w matrix can be calculated using formulas (11) :

$$S_B = \sum_{i=1}^{c} Ni(\mu_i - \mu)^T(\mu_i - \mu) \quad (10)$$

$$S_w = \sum_{i=1}^{c}\sum_{j=1}^{N_i}((x_j - \mu_i)^T(x_j - \mu_i)) \quad (11)$$

Where c represents the total number of classes, Ni is the sample number of each class, and i represents the number of classes of the entire class. The next step is to calculate the matrix covariance value (S). Calculation of covariance matrix values can be calculated using formulas (12)

$$ArgMax\ S = S_B * (S_w)^{-1} \quad (12)$$

*name of corresponding author

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 6, Number 1, October 2021
DOI : https://doi.org/10.33395/sinkron.v6i1.11234
e-ISSN : 2541-2019
p-ISSN : 2541-044X

Where ArgMax S is looking for the highest value of the covariance matrix which is a reduction matrix of the linear discriminant analysis (LDA) extraction process that has smaller dimensions compared to the original image matrix dimensions.
And count eigenvalue (λ) and eigenvector (d). Once the ArgMaxS value is determined, the next step is to determine the eigenvalue matrix (λ) and the eigenvector matrix (v) can be calculated by formula (13):

$$|(S_B\ {S_w}^{-1})^T - \lambda \mathrm{I}| = 0 \quad (13)$$

After that the factorization process is carried out, it will get a value that will later be used to find the value of eigenvector (v), with formula (14):

$$[v, d] = eig\ (S)|S - \lambda \mathrm{I}|v = 0 \quad (14)$$

Where v is a matrix of columns with elements (x1,x2,..., xi) in them, this matrix is called eigenvector. The two matrices are multiplied until a formula is obtained (15):

$$(S_{11} - \lambda_1)x_1 + S_{12}x_2 + \cdots S_{cn}x_n = 0 \quad (15)$$

Eigenvector projection results that correlate with eigenvalue are easier to separate than using eigenvalue that correlates with smaller eigenvalue. And calculate the projection matrix and the weight matrix. Calculation of the projection matrix can be calculated by equation (16). After that, calculating the weight matrix of the LDA extraction process can be calculated by the formula (17).

$$W_{Train} = (x - \mu_i)^T * v \quad (16)$$

$$Eigen_{Train} = (x - \mu_i) * v \quad (17)$$

K Nearest Neighbor (K-NN)

The working principle of the K-Nearest Neighbor algorithm is to find the closest distance between the data to be evaluated and the nearest k neighbor in training data. Here is the working process of the K-Nearest Neighbor algorithm:

Determine parameter k the number of closest neighbors and then calculate the Euclidean Distance of each object against the existing data sample,

$$d_i = \sum_{i=1}^{p} (x_{2i} - x_{li})^2 \quad (18)$$

Then sort the objects into groups that have a small Euclidean distance. After sorting according to the smallest Euclidean distance, then adjust category Y (Nearest Neighbor Classification). With μ is the sample mean, and σ is the standard deviation of the sample.

Confusion Matrix

Confusion Matrix is a method usually used to perform accurate calculations on the concept of data mining. Measurement of the performance of a classification system is essential (Mutawalli, Zaen, & Bagye, 2019). The performance of the classification system describes how well the system classifies the data. The confusion matrix contains information that compares the system's classification results with the actual classification results (Hana, 2020). Analysis of the performance used in the study included accuracy (19), precision (20), recall (21), F1-SCORE (22), and MCC (23).

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (19)$$

$$Precision = \frac{TP}{TP + FP} \quad (20)$$

$$Recall = \frac{TP}{TP + FN} \quad (21)$$

$$F_1 = 2\frac{PPV.TPR}{PPV + TPR} = \frac{2TP}{2TP + FP + FN} \quad (22)$$

$$MCC = \frac{\mathrm{TPxTN} - \mathrm{FPxFN}}{\sqrt{(\mathrm{TP} + \mathrm{FP}\ \mathrm{TP} + \mathrm{FN}\ \mathrm{TN} + \mathrm{FP}\ \mathrm{TN} + \mathrm{FN})}} \quad (23)$$

*name of corresponding author

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 6, Number 1, October 2021
DOI : https://doi.org/10.33395/sinkron.v6i1.11234
e-ISSN : 2541-2019
p-ISSN : 2541-044X

METHOD

In this study, the classification method used to classify datasets was K Nearest Neighbor (K-NN) with distances K = 3 and K = 5. Before classifying the dataset will be reduced using PCA and LDA methods, then it will be compared to K-NN performance without reduction, K-NN that has been reduced with PCA, and K-NN that has been reduced with LDA. Classification performance measurements will be measured based on accuracy, precision, recall, F1-SCORE and MCC values. More easily the following is the work procedure carried out in this study can be seen in Figure 1.

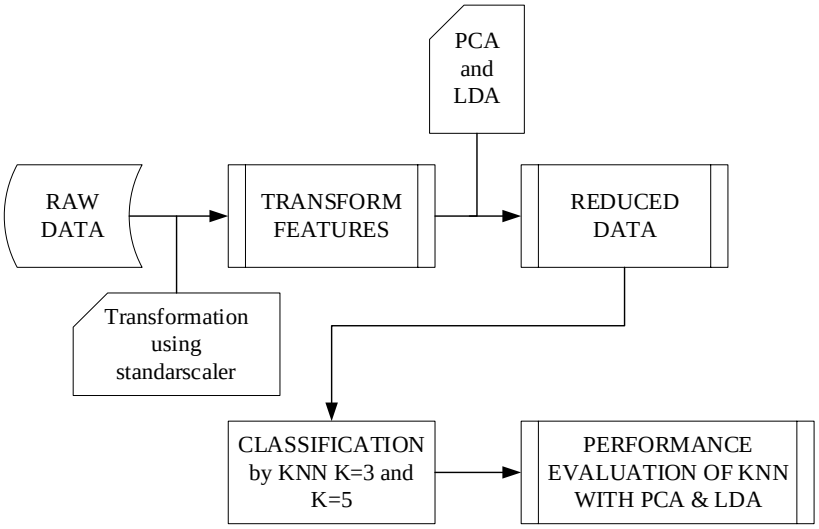


Fig. 1 Work procedure

RESULT

This study used three datasets as stated in table 1. The dataset used in this study has more than 100 attributes to see the effect of dimensional reduction on datasets with many attributes. The attributes contained in the data will go through the dimension reduction stage using Principal Component Analysis (PCA) and Linear Discriminant Analysis (LDA) to reduce the number of attributes in the data. Each dataset will be reduced by the provision of the number of classes minus one (n-1). After the reduction of dimensions, then implement the K Nearest Neighbor (K-NN) classification's model.

K-NN Performance Without Reduction

Table 2. K-NN K=3 Performance Results Without Reduction

Dataset	Accuracy	Precision	*Recall*	*F1*	*MCC*
Arrhythmia	63.24%	0.517	0.632	0.551	0.225
ISOLET	83.76%	0.851	0.837	0.837	0.831
CNAE-9	84.26%	0.850	0.840	0.840	0.823

Table 2, K-NN testing with K = 3 without reduction conducted on all three datasets, shows that the accuracy in the ISOLET and CNAE-9 datasets are more than 80%. In contrast, in the Arrhythmia dataset, only get the accuracy of 63.24%. The ISOLET dataset achieves the highest precision and MCC, while CNAE-9 achieves recall and F1-SCORE. K-NN K=5 performance results can be seen in table 3.

Table 3. K-NN K=5 Performance Results Without Reduction

Dataset	Accuracy	Precision	*Recall*	*F1*	*MCC*
Arrhythmia	63.24%	0.458	0.632	0.521	0.174
ISOLET	86.37%	0.875	0.864	0.864	0.859
CNAE-9	85.19%	0.859	0.852	0.850	0.834

Table 3 shows that K-NN testing with K=5 without reduction in the Arrhythmia dataset showed nothing changed in K-NN K=3 accuracy without the previous reduction of 63.24%, as did recalls with the same result of 0.632,

*name of corresponding author

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 6, Number 1, October 2021
DOI : https://doi.org/10.33395/sinkron.v6i1.11234
e-ISSN : 2541-2019
p-ISSN : 2541-044X

while in precision, F1-SCORE and MCC decreased. In the ISOLET dataset there is a change in position where the ISOLET dataset gets the highest accuracy of 86.37% and CNAE-9 by 85.19%.

PCA and K-NN performance

In this step, the dimensions of the dataset will be reduced using Linear Discriminant Analysis. The number of dimensions/non-class attributes will be reduced as the number of classes minus one (n-1). The dimension-reduced data uses Linear Discriminant Analysis and will be classified using K-NN K=3 and K=5.

Table 4. PCA+K-NN K=3 Performance Results

Dataset	Accuracy	Precision	*Recall*	*F1*	*MCC*
Arrhythmia	63.24%	0.474	0.632	0.536	0.226
ISOLET	81.15%	0.829	0.812	0.813	0.805
CNAE-9	72.84%	0.736	0.728	0.728	0.695

Table 4, the Arrhythmia dataset after being reduced using Linear Discriminant Analysis of K-NN performance results with K =3 did not change so much while the ISOLET and CNAE-9 datasets showed decreased performance compared to the results in table 2.

Table 5. PCA+K-NN K=5 Performance Results

Dataset	Accuracy	Precision	*Recall*	*F1*	*MCC*
Arrhythmia	63.97%	0.495	0.640	0.539	0.221
ISOLET	83.25%	0.842	0.832	0.831	0.826
CNAE-9	73.15%	0.739	0.731	0.731	0.699

For table 5, the performance results of K-NN with K = 5, whose dataset has been reduced using Linear Discriminant Analysis, show an increase compared to the results shown in table 4. In contrast, when compared to the results obtained in table 3, there is a decrease after being reduced with Linear Discriminant Analysis.

LDA and K-NN performance

The dimensions of the dataset will be reduced using Linear Discriminant Analysis (LDA). The number of dimensions/non-class attributes will be reduced to the number of classes minus one (n-1). Data reduced in dimensions using Linear Discriminant Analysis will be classified using K-NN K=3 and K=5.

Table 6. LDA+K-NN K=3 Performance Results

Dataset	Accuracy	Precision	*Recall*	*F1*	*MCC*
Arrhythmia	98.53%	0.986	0.985	0.985	0.975
ISOLET	96.15%	0.962	0.962	0.962	0.960
CNAE-9	99.07%	0.991	0.991	0.991	0.990

Table 6, after being reduced using Linear Discriminant Analysis, the results of K-NN performance with K = 3 increased to produce performance with a value of more than 90%. The most performance was increased rapidly in the Arrhythmia dataset when compared to KKN performance without reduction. For accuracy, precision, recall, F1-SCORE, and MCC, the highest was obtained from the CNAE-9 dataset with a value of more than 99%.

Table 7. LDA+K-NN K=5 Performance Results

Dataset	Accuracy	Precision	*Recall*	*F1*	*MCC*
Arrhythmia	94.12%	0.908	0.941	0.923	0.898
ISOLET	96.84%	0.969	0.968	0.968	0.967
CNAE-9	99.07%	0.991	0.991	0.991	0.990

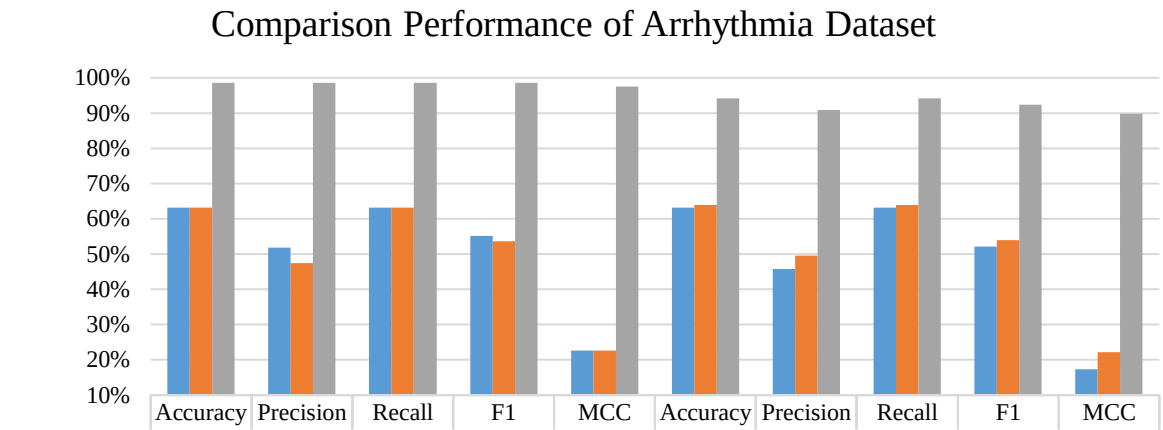
While in table 7, the results of K-NN performance with K = 5 whose datasets have been reduced using Linear Discriminant Analysis showed that there was a decrease in performance in the Arrhythmia dataset when compared to the results shown in table 6, but in the ISOLET dataset there was an increase in the results of accuracy, precision, recall, F1-SCORE, and MCC while the CNAE-9 dataset did not occur such significant changes.

*name of corresponding author

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 6, Number 1, October 2021
DOI : https://doi.org/10.33395/sinkron.v6i1.11234
e-ISSN : 2541-2019
p-ISSN : 2541-044X

DISCUSSIONS

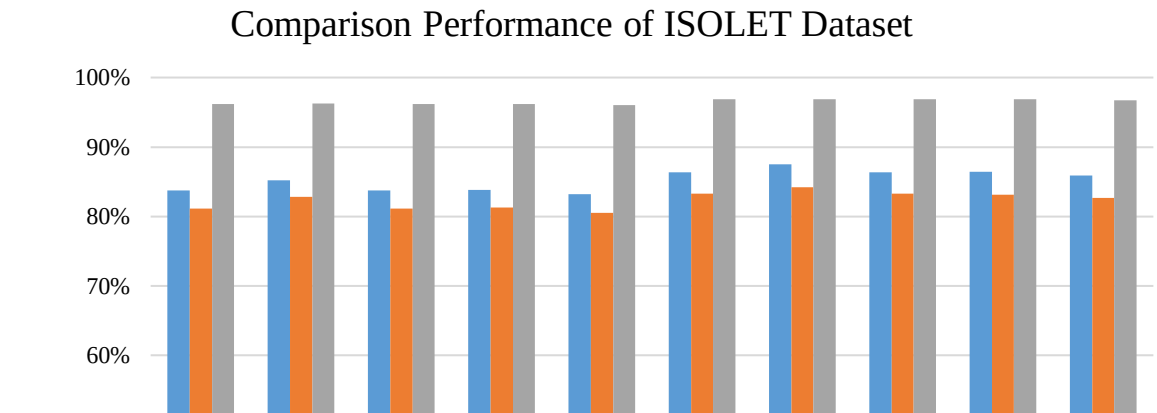
The study conducted tests using two methods of dimensional reduction, Principal Component Analysis, and Linear Discriminant Analysis, then classified using the K-NN method. The goal is to determine the effect of reduction methods on datasets with more than 100 attributes.



	Accuracy	Precision	Recall	F1	MCC	Accuracy	Precision	Recall	F1	MCC
	K=3					K=5				
KNN	63.24%	0.518	0.632	0.552	0.226	63.24%	0.458	0.632	0.521	0.174
PCA+KNN	63.24%	0.474	0.632	0.536	0.226	63.97%	0.495	0.640	0.539	0.221
LDA+KNN	98.53%	0.986	0.985	0.985	0.975	94.12%	0.908	0.941	0.923	0.898

Fig. 2 Comparison Performance of Arrhythmia Dataset

Based on figure 2, in the Arrhythmia dataset, the best performance results were obtained when the data was reduced using Linear Discriminant Analysis. At the same time, Principal Component Analysis method did not show such a significant difference.



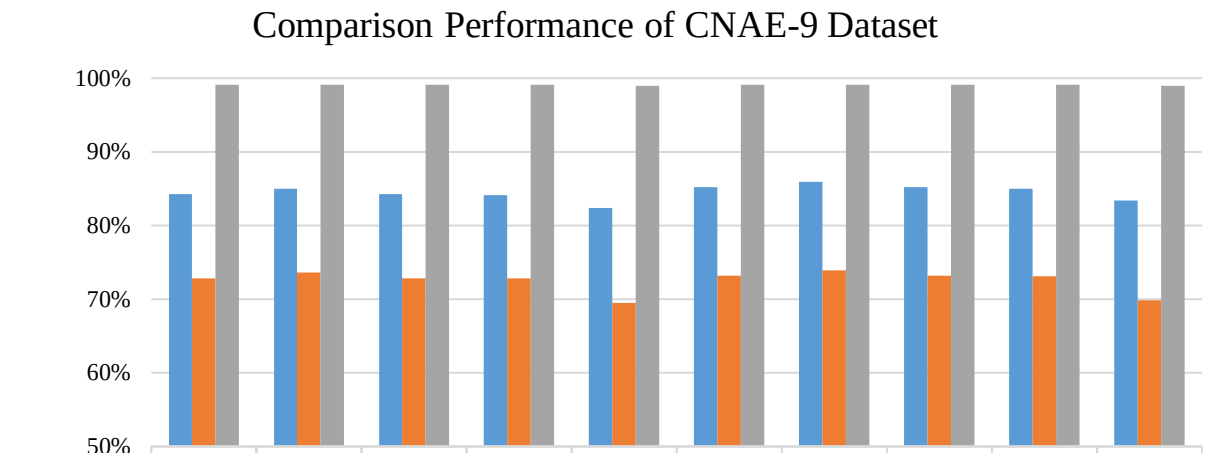
	Accuracy	Precision	Recall	F1	MCC	Accuracy	Precision	Recall	F1	MCC
	K=3					K=5				
KNN	83.76%	0.852	0.838	0.838	0.832	86.37%	0.875	0.864	0.864	0.859
PCA+KNN	81.15%	0.829	0.812	0.813	0.805	83.25%	0.842	0.832	0.831	0.826
LDA+KNN	96.15%	0.962	0.962	0.962	0.960	96.84%	0.969	0.968	0.968	0.967

Fig. 3 Comparison Performance of ISOLET Dataset

While in the ISOLET dataset, as shown in figure 3, it was obtained that the dataset has been reduced using Linear Discriminant Analysis is the best performance result. At the same time, the Principal Component Analysis method decreases compared to K-NN without the use of dimension reduction methods. The results of the performance of the CNAE-9 dataset can be seen in figure 4.

*name of corresponding author

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 6, Number 1, October 2021
DOI : https://doi.org/10.33395/sinkron.v6i1.11234
e-ISSN : 2541-2019
p-ISSN : 2541-044X



	Accuracy	Precision	Recall	F1	MCC	Accuracy	Precision	Recall	F1	MCC
	K=3					K=5				
KNN	84.26%	0.849	0.843	0.841	0.824	85.19%	0.859	0.852	0.850	0.834
PCA+KNN	72.84%	0.736	0.728	0.728	0.695	73.15%	0.739	0.731	0.731	0.699
LDA+KNN	99.07%	0.991	0.991	0.991	0.990	99.07%	0.991	0.991	0.991	0.990

Fig. 4 Comparison Performance of CNAE-9 Dataset

Figure 4 shows that LDA +K-NN also obtains the best performance.

CONCLUSION

Based on the study results, the conclusion that can be drawn from this study is that the performance of classifiers with LDA is better than with PCA with datasets that have more than 100 attributes. In the Arrhythmia dataset, which has 280 attributes, 452 data and 12 classes were able to improve performance on K-NN with values K = 3 and K = 5 and LDA + K-NN obtained the highest performance with a value of K = 3, which produced an accuracy value of 98.53%, precision of 0.986, recall 0.985, F1-SCORE 0.985 and MCC 0.975, while the lowest performance was in K-NN without reduction with a value of K = 3. For ISOLET dataset with 618 attributes, 7797 data, and 26 classes the best performance results are also obtained by data that has been reduced dimensions with Linear Discriminant Analysis but the best performance is obtained when the classification of K-NN with a value of K = 5, which produces an accuracy value of 96.84%, precision 0.969, recall 0.968, F1-SCORE 0.968 and MCC 0.967, while the lowest performance is in PCA + K-NN with a value of K = 3. And for the best performance of the CNAE-9 dataset, which has 857 attributes, 1080 data, and nine classes, is also achieved by LDA + K-NN, while the lowest performance is in PCA + K-NN with a value of K = 3.

This research can still be developed by comparing several other classification methods. More and more differences are seen when using dimension reduction methods, especially Principal Component Analysis and Linear Discriminant Analysis to improve performance on datasets with large attributes.

REFERENCES

Budiman, E., Santoso, E., & Afirianto, T. (2017). Pendeteksi Jenis Autis pada Anak Usia Dini Menggunakan Metode Linear Discriminant Analysis (LDA). *Jurnal Pengembangan Teknologi Informasi Dan Ilmu Komputer*, 1(7), 583–592.

Cahyani, S., Wiryasaputra, R., & Gustriansyah, R. (2018). Identifikasi Huruf Kapital Tulisan Tangan Menggunakan Linear Discriminant Analysis dan Euclidean Distance. *Jurnal Sistem Informasi Bisnis*, 8(1), 57. Retrieved from https://doi.org/10.21456/vol8iss1pp57-67

Hana, F. M. (2020). Perbandingan Algoritma Neural Network Dengan Linier Discriminant Analysis (Lda) Pada Klasifikasi Penyakit Diabetes, 1, 1541–1541.

Hariadi, F., Rambu, R., & Enda, H. (2019). Face Detection Using Linear Discriminant Analysis (Lda) Method and Support Vector Machine (Svm). *JOINCS (Journal of Informatics, Network, and Computer Science)*, 1(2), 1–5. Retrieved from https://doi.org/10.21070/joincs.v1i2.521

Hasdyna, N., Nababan, E., & Effendi, S. (2019). Dimension Reduction in Datasets Using Information Gain To Enhance K-NN Performance, 6, 379–383.

*name of corresponding author

e-ISSN : 2541-2019
p-ISSN : 2541-044X

Lubis, A., Sihombing, P., & Nababan, E. (2020). Analysis of Accuracy Improvement in K-Nearest Neighbor using Principal Component Analysis (PCA). *Journal of Physics: Conference Series*, 1566, 12062. Retrieved from https://doi.org/10.1088/1742-6596/1566/1/012062

Ma, F. A., & Wisesty, U. N. (2018). Analisis Pengaruh Metode Reduksi Dimensi Minimum Redundancy Maximum Relevance pada Klasifikasi Kanker Berdasarkan Data Microarray Menggunakan Classifier Support Vector Machine Analysis of The Influence of Minimum Redundancy Maximum Relevance as Dimensiona, 5(1), 1499–1506.

Mutawalli, L., Zaen, M. T. A., & Bagye, W. (2019). KLASIFIKASI TEKS SOSIAL MEDIA TWITTER MENGGUNAKAN SUPPORT VECTOR MACHINE (Studi Kasus Penusukan Wiranto). *Jurnal Informatika Dan Rekayasa Elektronik*, 2(2), 43. Retrieved from https://doi.org/10.36595/jire.v2i2.117

Rosadi, M. I., Sanjaya, C. B., & Hakim, L. (2018). Klasifikasi Diabetic Retinopathy Menggunakan Seleksi Fitur Dan Support Vector Machine. *Jurnal RESISTOR (Rekayasa Sistem Komputer)*, 1(2), 109–117. Retrieved from https://doi.org/10.31598/jurnalresistor.v1i2.312

Suyanto, S., Siregar, B., Nababan, E., & Fikri, H. (2020). Classification of Infection Type Based on Leukocytes Examination Results Using K-Nearest Neighbor. *Journal of Physics: Conference Series*, 1566, 12130. Retrieved from https://doi.org/10.1088/1742-6596/1566/1/012130

Syaliman, K. U., Nababan, E. B., & Sitompul, O. S. (2018). Improving the accuracy of k-nearest neighbor using local mean based and distance weight. *Journal of Physics: Conference Series*, 978(1). Retrieved from https://doi.org/10.1088/1742-6596/978/1/012047

Wibawa, M. S., & Novianti, K. D. P. (2017). Reduksi fitur untuk optimalisasi klasifikasi tumor payudara berdasarkan data citra FNA. *Konferensi Nasional Sistem & Informatika*, 73–78.

*name of corresponding author