Accredited SINTA 2 Ranking
Decree of the Director General of Higher Education, Research, and Technology, No. 158/E/KPT/2021
Validity period from Volume 5 Number 2 of 2021 to Volume 10 Number 1 of 2026

Published online at: **http://jurnal.iaii.or.id**

JURNAL RESTI
(Rekayasa Sistem dan Teknologi Informasi)
Vol. 9 No. 1 (2025) 175 - 184 e-ISSN: 2580-0760

PCA and t-SNE Implementation for KNN Hypertension Classification

Andi Aulia Cahyana Resky[1*], Jessica Crisfin Lapendy[2], Andi Akram Nur Risal[3*], Dewi Fatmarani Surianto[4], Abdul Wahid[5]
[1,2,3,4,5]Informatics and Computer Engineering, Engineering, Makassar State University, Makassar, Indonesia
[1]aauliacr@gmail.com, [2]jessica.c.lapendy@gmail.com, [3]akramandi@unm.ac.id, [4]dewifatmaranis@unm.ac.id, [5]wahid@unm.ac.id

Abstract

Hypertension is a condition that, if allowed to increase, can significantly injure internal organs due to high blood pressure. The objective of this study is to use the K-Nearest Neighbor (KNN) algorithm along with PCA and t-SNE to accurately identify four categories of Hypertension, Normal, Hypertension, Stage 1 Hypertension, and Stage 2 Hypertension. After establishing the scope, a dataset consisting of 7,794 samples was sourced from Labuang Baji Regional General Hospital, Makassar, and contained age, weight, and systolic and diastolic blood pressure parameters. The class distribution is Normal (36.3%), Hypertension (43.12%), Stage 1 Hypertension (8.29%), and Stage 2 Hypertension (12.31%). Experimental results show that the KNN base model achieved 99% accuracy, KNN with PCA reached 100%, and KNN with t-SNE attained 99%. Cross-validation was used to evaluate model generalization, yielding accuracies of 91%, 94%, and 91%, respectively. These findings suggest that KNN, particularly when integrated with t-SNE, is highly effective in visualizing and classifying non-linear data structures. Furthermore, this study demonstrates that incorporating dimensionality reduction techniques enhances the interpretability of classified hypertension data, which is crucial for informed decision-making by mental health committees.

Keywords: *Hypertension; KNN; Dimensionality Reduction; PCA; t-SNE*

How to Cite: A. A. Cahyana Resky, J. C. Lapendy, A. A. Nur Risal, D. F. Surianto, and A. Wahid, "PCA and t-SNE Implementation for KNN Hypertension Classification Visualization", *J. RESTI (Rekayasa Sist. Teknol. Inf.)* , vol. 9, no. 1, pp. 175 - 184, Feb. 2025..
DOI: https://doi.org/10.29207/resti.v9i1.6208

1. Introduction

Blood pressure is a force used by humans to be able to carry out the process of blood circulation to the arterial walls of the human body [1]. However, it is important to maintain the stability of human blood pressure. Blood pressure that is too high, or later referred to as hypertension, can be the cause of other more dangerous diseases, such as kidney failure, stroke and heart disease [2]. Hypertension itself is a condition where systolic and diastolic blood pressure increases significantly or abnormally [3]. Blood pressure is said to be normal if the results of systolic or diastolic blood pressure checks are in the range of 120/80 mmHg. Meanwhile, blood pressure categorized as hypertension is in the range of 140/90 mmHg [1].

This hypertensive condition can hinder the distribution of nutrients and oxygen supply throughout the body because it disrupts the blood circulation process that carries both of these things [3]. This can be caused by risk factors such as genetics, age, gender, stress, food consumption, obesity, lack of activity and so on [4]. The number of factors that affect this disease, causing hypertension to be named a global public health problem by the World Health Organization (WHO)[5]. The percentage of adults with hypertension reached 31.1%, or 1.39 billion people. Quoted from data from the Indonesian Ministry of Health, the population of hypertensive patients aged 18 years and over reached 31.7% [6].

WHO also says that uncontrolled hypertension is responsible for 7 million productive-age deaths and 64 million disabilities [7]. This is because the symptoms of hypertension are often invisible, even though the increase that occurs continuously for a long time can cause various complications of diseases and internal

Received: 04-12-2024 | Accepted: 18-02-2025 | Published Online: 28-02-2025

organs that lead to increased mortality [5]. The increase also causes patient data to increase but has not been maximally utilized to anticipate before the occurrence of hypertension in a person. In this regard, research related to predicting the type of hypertension disease can be a solution to reduce the number of people with hypertension by dealing with it according to the classification of its type.

Research that discusses classification with various methods has been done before. One of them is discussing the prediction of hypertension at Petra University using Machine Learning and the SMOTE model combined with the BrSmote method. The results of the study showed a classification accuracy value that reached 83.9%, specificity 85.1%, sensitivity 83.3% and AUC 89.6%. The study consisted of a comprehensive dataset of 31,500 patients divided into 12,658 hypertensive cases and 18,842 non-hypertensive cases [8]. Research classifying hypertension has also been conducted using the SVM Grid Search and SVM Genetic Algorithm methods. Based on the results of existing research, it was found that SVM Grid Search combined with RBF kernel has the best accuracy rate of 89.22%, compared to the same SVM method with Linear kernel, which only reaches 88.24%. Meanwhile, the accuracy obtained using the SVM Genetic Algorithm method only reached 88.24% for each kernel, both linear and RBF. The research dataset was taken from Padangsari Health Center from July to December 2021, with a total data of 510 hypertensive patients classified into two groups, namely patients with normal blood pressure and hypertensive patients [9].

The next research is related to the classification of microbiome data composition based on t-SNE and Aitchison distance and the comparison of Logistic Regression, SVM, and Decision Tree methods. The dataset in this study was obtained from two sources, namely, Mycoplasma pneumoniae (MP) infection data derived from a study of MP infection in 99 Chinese children, which was divided into 40 patients diagnosed with MP infection and 59 healthy children of the same age. The other source came from idiopathic central precocious puberty (ICCP) data, where ICPP data were fecal microbiota from 25 girls with idiopathic central precocious puberty and 23 healthy girls in China. This dataset was classified into 2 classes, namely controls and cases. The results obtained in this study are by using the optimal parameters, such as perplexity *per* = 30, maximum number of iterations = 2000 and number of neighbors k = 7 [10].

In addition, subsequent research on the use of t-SNE on transcriptomic single-cell data found that the shortcomings of t-SNE were that it could not read the global structure accurately, so development was carried out with PCA initialization, setting the learning rate, multi-scale similarity kernels and downsampling-based initialization. The dataset of this study was obtained from several sources of previous publications [11]. Another study that discusses the classification of heart disease by comparing the K-Nearest Neighbor (KNN) and Decision Tree methods using PCA techniques. The source of this research dataset comes from Kaggle, which contains 1025 samples of segmented heart disease data with 14 relevant clinical attributes. This research obtained the highest accuracy from the Decision Tree method combined with the PCA technique, which reached 100%, whereas previously, it was only 98.54% when relying only on the Decision Tree method. The accuracy with the KNN method combined with PCA only reaches 79.02%, which is even lower than the classification using only KNN, which is 80.49% [12].

There is also research that utilizes the KNN algorithm combined with the calculation of Euclidean and Manhattan distance metrics to classify student graduation into 2 classes. This research uses data on students of the Informatics Engineering Department in the 2014 and 2015 academic years at the Yogyakarta University of Technology by using 9 variables in determining student classification. The accuracy of this test reached 85.28% for both distance measurement metrics used with each determination of the k value of 7 [13]. Based on some of these studies, the KNN algorithm is considered to have good accuracy in the classification process, and the use of dimensionality reduction has also succeeded in improving accuracy in classification.

Through this research, a prediction model for the type of hypertension based on four classification features, namely, Age, Weight, Systolic and Diastolic blood pressure using the KNN method is proposed by comparing the use of PCA and t-SNE dimensional reduction. The use of dimensional reduction is important so that visualization can be displayed so that the distribution of the resulting data distribution can be presented more effectively than just looking at tables of numbers and raw data [14].

PCA (Principal Component Analysis) and t-SNE (t-Distributed Stochastic Neighbor Embedding) are two dimensionality reduction techniques that have unique advantages in multidimensional data analysis. PCA is known for its ability to perform linear transformations that are effective in reducing the dimensionality of data while still extracting important information from large data sets and thoroughly analyzing the variable structure [15]. PCA reduces the dimensionality, making complex data simple while preserving the main variance in the data, resulting in improved model accuracy and efficiency [16]. In contrast, t-SNE excels at visualizing data structure by capturing non-linear relationships and complex patterns. t-SNE excels at a good visualization of the arrangement of data points based on the similarity between data points in high-dimensional to low-dimensional space [17].

There are 4 stages of the proposed research, namely, data collection, KNN classification, dimensional reduction, and model validation and evaluation. This

research is expected to predict the type of hypertension based on the common features used and display visualizations that are easy to understand. This is expected to be an effort to increase awareness of the high number of people with hypertension who need treatment according to the classification of their type.

2. Research Methods

In this study, several stages of research were carried out to ensure that the research results were valid and reliable. The stages of research can be seen in Figure 1.

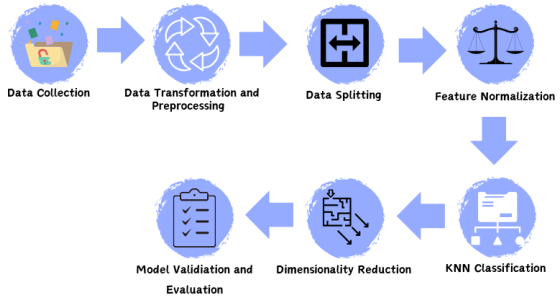

Figure 1. Research Stages

2.1 Data Collection

In this study, the dataset used is PPG-BP (photoplethysmography-blood pressure) data taken directly from Labuang Baji Regional General Hospital located in Makassar, South Sulawesi. The PPG-BP dataset used in this study comes from 7,794 people. The dataset consists of 8 features, namely gender, age, height (cm), weight (kg), blood pressure, pulse (bpm), and BMI (kg/m). As for the blood pressure feature, it will be separated into 2 different features, namely systolic pressure (mmHg) and diastolic pressure (mmHg), because these two features are important in determining the classification of hypertension disease types.

In this study, there are four classes used: Normal, Hypertension, Hypertension Stage 1, and Hypertension Stage 2. The Normal category is under 40 years old. Meanwhile, the Hypertension category is 30 years old and above, the Hypertension stage 1 category is 40 years old and above, and the Hypertension stage 2 category is 50 years old and above. As for the distribution of datasets from 7,794 samples, 36.3% of them are Normal data, 43.12% are Perhypertension data, 8.29% are stage 1 Hypertension data, and 12.31% are stage 2 Hypertension data, as shown in Figure 2. This means that the distributions of each class are imbalanced.

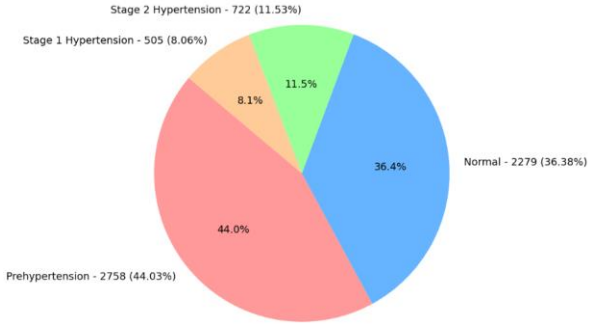

Figure 2. Dataset Distribution

2.2 Preprocessing Data

The dataset that has been collected will first pass the data transformation stage, which is to convert some features into numerical form so that the algorithm can be applied properly. Starting from the gender feature, by giving a value of 1 for men and a value of 2 for women. Furthermore, the hypertensive disease type category feature is also changed, namely Normal to 1, Perhypertension to 2, Stage 1 Hypertension to 3, and Stage 2 Hypertension to 4. The blood pressure feature is also separated into two features, namely Systolic Pressure and Diastolic Pressure, by removing the per sign (/) and using the first number as Systolic Pressure and the second number as Diastolic Pressure. Table 2 and Table 3 show the data before and after data transformation.

Table 2. Data Before Data Transformation

Gender	Age	High	Weight	Blood Pressure	Pulse	BMI	Category
Male	38	169	86	130/80	92	30,11099	Hypertension
Female	52	155	50	110/75	80	20,81165	Normal
Male	27	170	60	130/80	80	20,76125	Hypertension
Female	58	155	55	120/80	80	22,89282	Hypertension
Male	37	160	60	120/80	88	23,4375	Hypertension
...	...	...	...	...	...	...	...
Female	43	155	58	120/80	80	24,14152	Hypertension

Table 3. Data After Data Transformation

Gender	Age	High	Weight	Systolic Blood Pressure	Diastolic Blood Pressure	Pulse	BMI	Category
1	38	169	86	130	80	92	30,11099	2
2	52	155	50	110	75	80	20,81165	1
1	27	170	60	130	80	80	20,76125	2
2	58	155	55	120	80	80	22,89282	2
1	37	160	60	120	80	88	23,4375	2
...	...	...	...	...	...	...	...	...
2	43	155	58	120	80	80	24,14152	2

After data transformation, the dataset will be separated between features and targets. From eight features, there are 4 features only that will be used, such as "Age", "Weight", "Systolic Blood Pressure", and "Diastolic Blood Pressure", and the target is the four categories of hypertension disease types. The selection of features is based on several studies that prove that the risk of hypertension will increase with age due to changes that occur in the body, especially stiff arteries and decreased kidney function. [18]. The other four features were not used in the study because they generally have no direct effect on hypertension risk. According to the health research on Factors affecting hypertension in the community in Bedagai village, Pinang City, no influence was found between male or female gender on hypertension. This is not in line with research conducted by Aristotle, which states that women are more susceptible to hypertension due to hormonal differences. Based on the results of the research, the majority of people in Bedagai Kota Pinang village are male [19].

Other features, such as pulse, are not used because they do not have a direct effect on increasing the risk of hypertension. Based on research that discusses the relationship between blood pressure values and pulse frequency with the quality of life of hypertensive patients, pulse frequency is used to see the quality of life of a hypertensive patient, whether categorized as good or not [20]. As for the height feature, it is indirectly related to factors that increase the risk of hypertension significantly but do not directly affect the risk of hypertension itself. height, will affect a person's BMI, where BMI itself is the ideal body weight. An obese body condition can cause hypertension because it triggers greater pressure on the arterial walls due to the increased volume of blood circulating through the blood vessels. Excess weight that exceeds the BMI standard can be a contributing factor to the occurrence of various diseases, and one of them is hypertension [21]. However, we do not use this feature because the weight feature is sufficient. In addition, body weight is also an influential factor in hypertension because individuals with excess weight will have difficulty moving freely so the heart needs to pump blood and increase blood pressure [22].

Next, the feature data will pass the preprocessing stage by replacing nullable or empty values with average data (mean). Then, the data will be selected into training and testing data with a composition of 80% for training data and 20% for test data so that the total training data amounts to 6234 and test data amounts to 1559.

To normalize the feature data so that it is on the same scale in both training data and test data, feature normalization is carried out using the Robust Scaller method, which uses the median and Interquartile Range (IQR) in the normalization process [23]. The selection of the method is based on its advantages that are not affected by outliers [24]. The selection of this method is based on its advantages that are not affected by outliers [24], especially in the weight and systolic blood pressure features where some patients have quite extreme values.

2.3 KNN Classification

Next, a model is formed that will be used in data classification. In this case, it is a model for classifying hypertension using the K-Nearest Neighbor (KNN) method. The KNN algorithm begins with the selection of the K value by determining the number of nearest neighbors (K) that will be used in the classification process. This K value will affect the complexity of the model and can affect the performance of the model. Next, the calculation of the distance from the instance to be predicted with a commonly used distance metric is the Euclidian Distance or Manhattan Distance, but it can use other metrics depending on the case. The Euclidean Distance calculation can be formulated in Equation 1: [25]

$$d(x_1, x_2) = \sqrt{\sum_{i=1}^{n}(x_{2i} - x_{1i})^2} \quad (1)$$

In the formula, X1 refers to the value of each data used in the model training process, while X2 is the value of each data used in model testing. The variable n indicates the total amount of data used in the analysis, while i refers to the i data in the data set.

After the results of neighbor proximity are obtained, the values obtained will be sorted from largest to smallest, and then the score calculation is carried out. Before classification, the data used in the study was categorized first based on the type of hypertension according to the European Society of Cardiology (ESC) in 2018 [26] as shown in Table 1.

Table 1. Classification of Hypertension Disease Type

Classification	Systolic	Diastolic
Normal	< 120	< 80
Hypertension	120 - 139	80 - 89
Stage 1 hypertension	140 - 159	90 - 99
Stage 2 hypertension	≥ 160	≥ 100

2.4 Dimensionality Reduction

Dimensionality reduction is a process of removing or reducing a number of features contained in a document. This is done as a form of simplification of the complexity of large data, but important and specific information will still be retained so that the resulting conclusions will be more relevant [27]. This research will discuss the combination of the KNN method with two dimensionality reduction techniques, PCA and t-SNE, in the classification of hypertension disease predictions.

PCA (Principal Component Analysis) is one of the reduction techniques that is considered quite efficient in improving algorithm performance. This is because PCA can minimize the presence of multicollinearity or a strong relationship between two or more variables. Not only that, PCA is also able to remove the correlation that exists in the independent variables so as to cut off the possibility of correlated variables [28]. PCA

performs feature or dimension reduction by forming a matrix. After that, it is continued with the covariance matrix calculation process, where each vector is transformed into a new vector linearly. Based on the covariance matrix, the Eigenvalue and Eigenvector are then calculated. After that, the last stage is the calculation of the orthogonal transformation value [27].

t-SNE (t-Distributed Stochastic Neighbor Embedding). This technique belongs to unsupervised learning or non-linear techniques. Unlike PCA, which excels in data feature reduction, t-SNE will measure the similarity between pairs of data points in each dimension, through several types of measurements such as Euclidean or Gaussian [29] . In the t-SNE stage, it is first necessary to find the similarity between the closest points in the highest dimensional space. Next, each point in the highest dimensional space will be assigned to the lower dimensional space based on the pairwise relationship between points in the higher dimensional space. Then, gradient descent is applied to find a low-dimensional representation of the data. Finally, points that have similarities in the low-dimensional space will be calculated using the t-student distribution [30].

PCA affects the classification process by reducing the dimensionality of features based on the highest variance so that only the most relevant information is retained. This is one of the advantages of using PCA because it can reduce complexity and improve computational efficiency by accelerating machine learning algorithms [31]. Meanwhile, t-SNE helps in the better representation of non-linear data. This technique is very useful for visualizing the general structure and heterogeneity of data sets [32]. t-SNE excels at keeping similar samples close together and placing dissimilar samples at a greater distance. In addition, the ability of t-SNE to control local and global relationships between points usually results in more visually appealing clusters.

2.5 Model Validation and Evaluation

The model validation and evaluation stage is a stage carried out to see the success of the model in the previous classification process. Confusion matrix is one way that is usually done to see the accuracy, precision and recall of the system that has been tested, usually expressed in units of percent (%). The following are the steps for calculating the confusion matrix: Calculate the sum of the results of positive actual data and positive predicted data (TP); Calculate the number of results of positive actual data and negative predicted data (FN); Calculating the sum of the results of negative actual data and negative predicted data (TN); Calculate the sum of the results of negative actual data and positive predicted data (FP).

Accuracy is the ratio of the comparison between correct predictions, be it positive or negative predictions, with the entire data. Precision is the ratio of the comparison between positive correct predictions and the overall positive predicted results. The recall is the ratio of the comparison between positive correct predictions and the overall positive correct data. Equations 2, 3, and 4 are equations for calculating accuracy, precision, and recall using the confusion matrix [16]:

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN}\ x\ 100\% \quad (2)$$

$$\text{Precision} = \frac{TP}{TP+FP}\ x\ 100\% \quad (3)$$

$$\text{Recall} = \frac{TP}{TP+FN}\ x\ 100\% \quad (4)$$

In addition, model validation and evaluation are also carried out by comparing the classification results with the classification results using the KNN model.

Model validation, such as cross-validation, is also done to avoid the possibility of overfitting the data used. Cross-validation itself is a technique used to test model performance on different subsets of data [33]. This technique is widely used to evaluate the model by predicting and estimating how accurate a predictive model is if run in practice so that bias in the data can be eliminated [34]. In this research, one of the techniques of cross-validation is k-fold cross-validation with the use of k-fold = 5. The data will be split into K parts of the data set with the same size, then training and testing will be carried out as many as K [34]. The workflow of cross-validation can be seen in Figure 3.

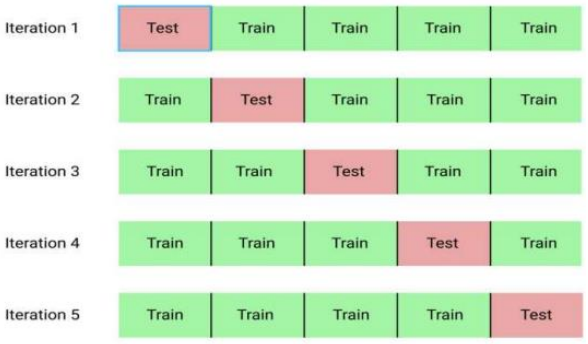

Figure 3. Workflow of Cross-Validation

3. Results and Discussions

This stage will present the classification results of the four hypertension categories through 3 experimental scenarios using the KNN algorithm. The data used is data that has been normalized at the data preprocessing stage. The following are the three test scenarios that have been carried out:

3.1 Test Scenario Using KNN

In this test, classification is carried out using the KNN method with test performance results from 1559 test data, which can be seen in Table 4. The model performance of this experiment can be seen in the confusion matrix shown in Figure 4.

It can be seen that the results of this test scenario are good, with a classification accuracy of 99%, so there are prediction errors in some classes due to the unbalanced amount of test data in each class. The explanation of

prediction errors according to the confusion matrix is as follows: In the Normal class there are 543 test data and the model successfully predicts 533 data as the Normal class while the other 10 data as the Hypertension class; In the Perhypertension class, there are 672 test data and the model successfully predicts 671 data as the Perhypertension class while the other 1 data is the Hypertension stage 1 class; In stage 2 Hypertension class, there are 220 test data and the model successfully predicts 216 data as stage 2 Hypertension class while 1 other data as Normal class, 2 other data as Perhypertension class, and 1 other data as stage 1 Hypertension.

In addition, to see the visualization of the classification pattern of the test data in the four categories, dimensional reduction is carried out using 2 methods, namely PCA [35] and t-SNE [36].

Table 4. KNN Classification Testing Result

Class	Precission	Recall	F1-Score	Accuracy
Normal	1.00	0.98	0.99	
Hypertension	0.98	1.00	0.99	
Stage 1 Hypertension	0.98	1.00	0.99	0.99
Stage 2 hypertension	1.00	0.98	0.99	

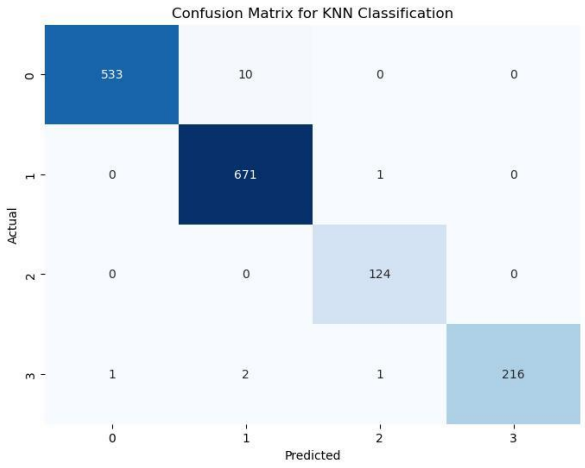

Figure 4. Confusion Matrix for KNN Classification

3.2 Testing Scenario Using KNN and PCA

In this test, classification is carried out using the KNN and PCA methods, which are two of the dimensionality reduction methods that are widely used in complex data [35]. The test performance results of 1559 test data in this scenario can be seen in Table 5.

Table 5. KNN and PCA Classification Test Result

Class	Precission	Recall	F1-Score	Accuracy
Normal	1.00	1.00	1.00	
Hypertension	0.99	1.00	1.00	
Stage 1 hypertension	0.99	1.00	1.00	1.00
Stage 2 hypertension	1.00	0.99	0.99	

With the dimensionality reduction, the data can be transformed into a 2-dimensional space, and the visualization of its classification on the four types of hypertension can be seen in Figure 5.

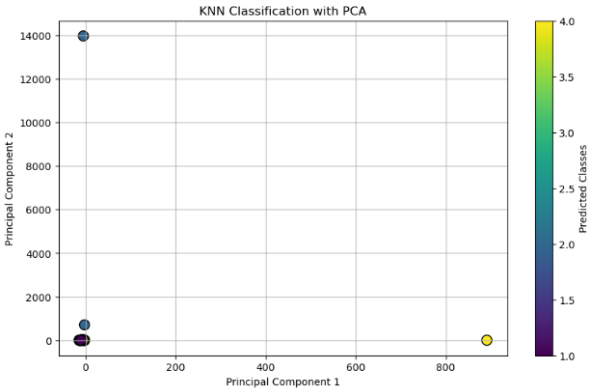

Figure 5. Visualization of KNN and PCA Classification

The results of PCA value interpretation on the top five test data can be seen in Table 6, where the value shows that the points are closer to each other, and this is in accordance with the classification visualization.

Table 6. PCA Value Interpretation Results

Age	Weight	Systolic Blood Pressure	Diastolic Blood Pressure	PC1	PC2
66	55	120	80	-9,8026	-0,70591
62	37	88	49	-6,56439	-0,38496
40	60	120	80	-8,49313	0,896607
58	61	158	93	-8,8305	-0,24106
28	42	95	75	-10,2786	-1,53962

The model performance of this experiment can be seen in the confusion matrix shown in Figure 6.

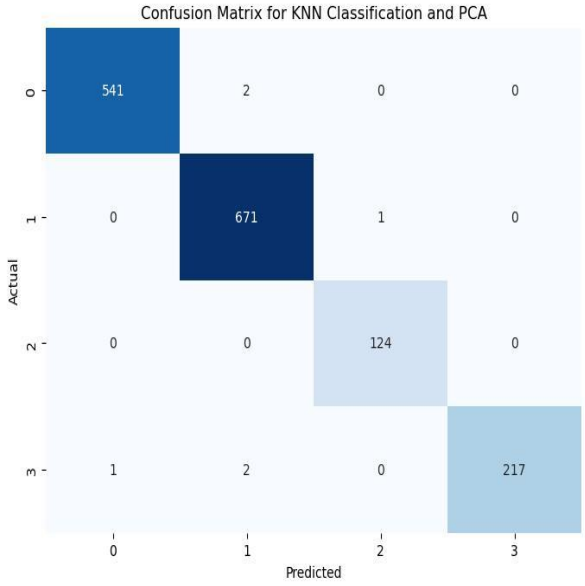

Figure 6. Confussion Matrix for KNN Classification and PCA

It can be seen that the results of this test scenario are superior with a classification accuracy obtained of

100%. However, there are still prediction errors in the same 3 types of classes as in the first test. This is because the research data has a non-linear relationship, as shown in Table 6, so the model is difficult to provide accurate predictions. The following is an explanation of the prediction error according to the confusion matrix above: In the Normal class, there are 543 test data and the model successfully predicts 541 data as the Normal class while the other 2 data as the Hypertension class; In the Perhypertension class, there are 672 test data and the model successfully predicts 671 data as the Perhypertension class while the other 1 data is the Hypertension stage 1 class; In stage 2 Hypertension class, there are 220 test data and the model successfully predicts 217 data as stage 2 Hypertension class while 1 other data as Normal class and 2 other data as Perhypertension class.

The superior accuracy in this test scenario is because PCA dimension reduction focuses more on the most important features based on the highest variance [37] of the four features used so as to reduce noise from less relevant features. However, there are drawbacks to this test scenario, namely the poor visualization results because it is a linear method, so it cannot capture non-linear relationships that exist in the data [38]. Table 7 shows seven examples of data used in the study that have non-linear relationships.

Table 7. Data with Non-Linier Relationship

Gender	Age	Height	Weight	Systolic Blood Pressure	Diastolic Blood Pressure	Pulse	BMI	Categories
1	38	169	86	130	80	92	30,11099	2
2	31	157	54	110	70	80	21, 90758	1
1	55	175	75	114	74	105	24,4898	1
2	58	155	55	120	80	80	22,89282	2
1	22	150	50	120	80	80	22,22222	2
1	54	155	65	184	107	95	27,05515	4
1	34	150	86	174	123	94	38,22222	4

It is said to be non-linear because in the "Age" and "Weight" features, there are changes in one data that are not always the same as changes in other data. For example, the first data, when compared to the fourth data, has a quite different age and weight range even though it is in the same category, namely category 2 or Hypertension. Therefore, a dimension reduction method is needed that can capture non-linear relationships in the data well, one of which is the t-SNE method [38].

3.3 Testing Scenario Using KNN and t-SNE

In this test, classification is carried out using the KNN and t-SNE methods, which are dimension-reduction techniques based on the concept of probability distribution and similarity between data points. [39]. The test performance results of 1559 test data in this scenario can be seen in Table 8.

Table 8. KNN and t-SNE Classification Testing Results

Class	Precission	Recall	F1-Score	Accuracy
Normal	1.00	0.98	0.99	
Hypertension	0.98	1.00	0.99	
Stage 1 hypertension	0.98	1.00	0.99	0.99
Stage 2 hypertension	1.00	0.98	0.99	

The visualization of the classification of the four types of hypertension can be seen in Figure 6 with a color description, and the amount of test data used is shown in Table 9.

The visualizations displayed by t-SNE in Figure 7 represent the results of reducing data from high dimensions to lower dimensions (2-dimensional or 3-dimensional). The X and Y axes seen in the visualization depiction show the relative relationship between data, and do not depict the original dataset variables. Different colors indicate classification labels, with the scale in the color bar helping to correlate the similarity of certain classes in the dataset. For example, in both the color bar scale and the data distribution pattern, the yellow and green classes are closely spaced. This shows that the yellow and green classes have some similar features compared to other classes. t-SNE basically simplifies the relationship between data, where similar data will be placed close together and very different data will be placed far apart.

Table 9. Classification Visualization Description

Class	Color	Number
Normal	Purple	536
Hypertension	Blue	681
stage 1 hypertension	Green	126
stage 2 hypertension	Yellow	216

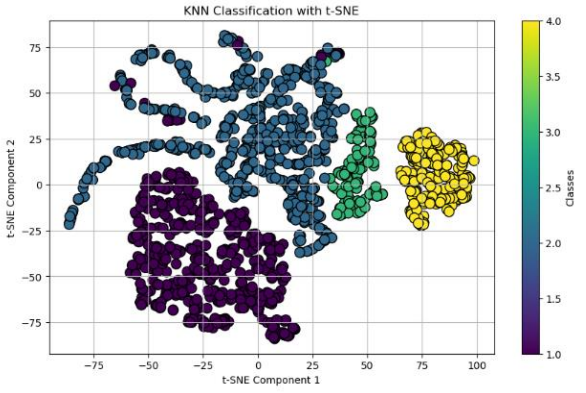

Figure 7. Visualization of KNN and t-SNE Classification

The model performance of this experiment can be seen in the confusion matrix shown in Figure 8.

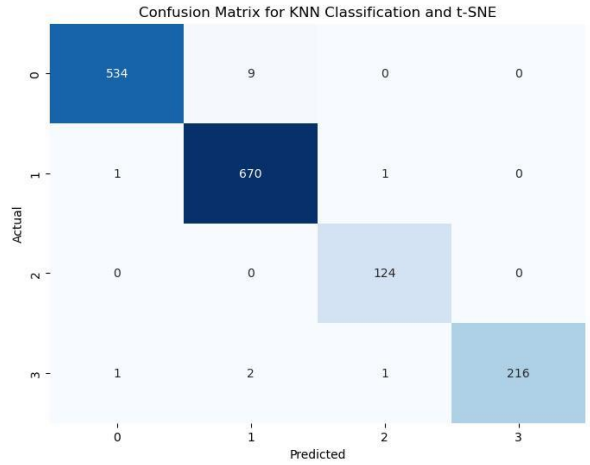

Figure 8. Confusion matrix for KNN Classification and t-SNE

It can be seen that there are still prediction errors in 3 types of classes, just like in the first and second test scenarios, so it can be said that these errors are related to the non-linear relationship that exists in the data. The following is an explanation of the prediction error according to the confusion matrix above: In the Normal class there are 543 test data and the model successfully predicts 541 data as the Normal class while the other 2 data as the Hypertension class; In the Perhypertension class, there are 672 test data and the model successfully predicts 671 data as the Perhypertension class while the other 1 data is the Hypertension stage 1 class; In stage 2 Hypertension class, there are 220 test data and the model successfully predicts 217 data as stage 2 Hypertension class while 1 other data as Normal class and 2 other data as Perhypertension class. In this test scenario, the classification accuracy obtained is 99%, which indicates that the accuracy is lower when compared to PCA. This is because t-SNE focuses on preserving local information in the data [37] thus ignoring some global information. In addition, the visualization of data classification in PCA and t-SNE is very different. This is because t-SNE is superior in capturing non-linear relationships between data so that it is effective in visualizing structures that PCA fails to capture. In addition, t-SNE is also better able to capture the fine details within each class, while PCA only retains the global data variance. [40]. Therefore, the visualization results of data classification using t-SNE dimension reduction show a clearer separation of the four types of hypertension, although there are some areas in which data with different colors overlap to indicate misclassification. This is in line with the model performance results shown previously, which did not reach 100%.

Based on the three test scenarios that have been carried out, the KNN classification model and t-SNE dimension reduction are used. Table 10 shows a comparison of the initial classification results with the classification results using the model on five test data. In these five data, there are 4 test data with the same prediction results, and there is also 1 test data with different prediction results marked with yellow highlights. The initial classification results were obtained based on the Systolic Blood Pressure and Diastolic Blood Pressure features described by ESC in 2018. The model classification results are obtained based on the features of Age, Weight, Systolic Blood Pressure and Diastolic Blood Pressure.

Table 10. Prediction Results of KNN and t-SNE Classification

Age	Weight	Systolic Blood Pressure	Diastolic Blood Pressure	Initial Classification	Model Classification
66	55	120	80	1	1
62	37	88	49	4	4
40	60	120	80	2	2
58	61	158	93	2	2
45	35	116	79	1	2

3.4 Testing Scenario Using K-Fold Cross-Validation

All three test scenarios obtained excellent results; even in the KNN and PCA test scenarios, the model prediction accuracy obtained was 100%. Meanwhile, the test scenario using KNN and the test scenario using KNN combined with t-SNE both obtained 99% accuracy. The high accuracy obtained from the three types of scenarios used is a promising result, but the possibility of overfitting and bias in the data is something to be aware of. Overfitting is a condition where a model that fits the training data too well may fail to perform well on new data. To overcome this problem, one of the cross-validation methods is used to evaluate the generalization ability of the model more comprehensively. The testing model used is K-Fold Cross-validation, where the model will test on various subsets of data based on the specified K value. In this test, the k value used is k = 5. The results of cross-validation performance on the KNN, KNN and PCA, and KNN and t-SNE models can be seen in Tables 11, 12 and 13, respectively.

In addition to seeing the accuracy of Cross Validation, the train and test accuracy are checked again, whether there is a significant difference between the two which proves the existence of overfitting data. The train accuracy and test accuracy results obtained are 94% and 91% for each test scenario performed. The results obtained prove that the performance of the model on training data and test data is almost the same; in this case, the model can be said to be generalized. Where the model not only memorizes training data but can also work well on new data. The small difference between the two proves that although the accuracy is not perfect, the model is stable in testing the data.

Table 11. KNN Cross-Validation Test Result

Class	Precission	Recall	F1-Score	Accuracy
Normal	0.98	0.90	0.94	
Hypertension	0.87	0.98	0.92	
Stage 1 hypertension	0.83	0.67	0.74	0.91
Stage 2 hypertension	0.98	0.87	0.92	

Table 12. KNN and PCA Cross-Validation Test Result

Class	Precission	Recall	F1-Score	Accuracy
Normal	0.99	0.93	0.96	
Hypertension	0.90	0.99	0.94	
Stage 1 hypertension	0.86	0.73	0.79	0.94
Stage 2 hypertension	0.99	0.89	0.94	

Table 13. KNN and t-SNE Cross-Validation Test Result

Class	Precission	Recall	F1-Score	Accuracy
Normal	0.97	0.90	0.94	
Hypertension	0.88	0.97	0.92	
Stage 1 hypertension	0.82	0.70	0.75	0.91
Stage 2 hypertension	0.95	0.89	0.92	

4. Conclusions

This investigation exemplified the accuracy of the K-Nearest Neighbor (KNN) method in identifying four classes of hypertension—Normal, Hypertension, Hypertension Stage 1, and Hypertension Stage 2—using the PPG-BP dataset that included 7,794 samples. Dimensionality reduction methods PCA and t-SNE were used for classification performance improvement and effective data visualization. As a result, KNN and PCA together had an unbeatable accuracy (100%) in the context of the analysis because the method was focused on linear relationships and, consequently, removing noise from irrelevant features. But the linear approach of PCA failed to represent non-linear data relationships, which led to the visualization of the data becoming less than perfect. Nevertheless, the joint operation of KNN and t-SNE resulted in a 99%-accuracy, however, it stood out in the visualization of the non-linear nature of the dataset. t-SNE effectively gathered similar data points, meanwhile, drawing inter-class transitions which gave a better picture of the data structure, although the classification accuracy was a little lower than PCA. These findings give attention to the length of the dimensionality reduction technique and its purpose, which guides if the emphasis is laid on the level of accuracy versus quality of visualization. The combination of KNN with dimensionality reduction enables medical decision-makers to enhance classification performance as well as interpretability by improving their medical practice which in turn leads to better treatment and timely detection of high blood pressure.

Acknowledgements

We, as the authors are deeply grateful to Makassar Labuang Baji Hospital for its valuable role in providing data for this research.

References

[1] P. Purwono, P. Dewi, S. K. Wibisono, and B. P. Dewa, "Model Prediksi Otomatis Jenis Penyakit Hipertensi dengan Pemanfaatan Algoritma Machine Learning Artificial Neural Network," *Insect (Informatics Secur. J. Tek. Inform.*, vol. 7, no. 2, pp. 82–90, 2022.
https://doi.org/10.33506/insect.v7i2.1828

[2] B. L. Yudha, L. Muflikhah, and R. C. Wihandika, "Klasifikasi Risiko Hipertensi Menggunakan Metode Neighbor Weighted K- Nearest Neighbor (NWKNN)," *J. Pengemb. Teknol. Inf. dan Ilmu Komput. Univ. Brawijaya*, vol. 2, no. 2, pp. 897–904, 2018.
https://j-ptiik.ub.ac.id/index.php/j-ptiik/article/view/998.

[3] I. Agustinus, E. Santoso, and B. Rahayudi, "Klasifikasi Risiko Hipertensi Menggunakan Metode Learning Vector Quantization (LVQ)," *J. Pengemb. Teknol. Inf. dan Ilmu Komput.*, vol. 2, no. 8, pp. 2947–2955, 2018.
https://j-ptiik.ub.ac.id/index.php/j-ptiik/article/view/1725.

[4] A. Insani and H. A. Ramadhani, "Determinan Kejadian Hipertensi Berdasarkan Pola Konsumsi: Model Prediksi Dengan Sistem Skoring," *Qual. J. Kesehat.*, vol. 16, no. 1, pp. 9–20, 2022.
https://ejournal.poltekkesjakarta1.ac.id/index.php/adm/article/view/399

[5] E. Martinez-Ríos, L. Montesinos, M. Alfaro-Ponce, and L. Pecchia, "A review of machine learning in hypertension detection and blood pressure estimation based on clinical and physiological data," *Biomed. Signal Process. Control*, vol. 68, no. May, p. 102813, 2021.
https://doi.org/10.1016/j.bspc.2021.102813

[6] A. Yonata, A. Satria, and P. Pratama, "Arif Satria Putra Pratama dan Ade Yonata | Hipertensi sebagai Faktor Pencetus Terjadinya Stroke Majority," *J. Major.*, vol. 5, no. 3, p. 17, 2016, [Online].
http://repository.lppm.unila.ac.id/id/eprint/22420

[7] L. Muflikhah, N. Hidayat, and D. J. Hariyanto, "Prediction of hypertention drug therapy response using K-NN imputation and SVM algorithm," *Indones. J. Electr. Eng. Comput. Sci.*, vol. 15, no. 1, pp. 460–467, 2019.
http://doi.org/10.11591/ijeecs.v15.i1.pp460-467

[8] Y. Sakka, D. Qarashai, and A. Altarawneh, "Predicting Hypertension using Machine Learning: A Case Study at Petra University," *Int. J. Adv. Comput. Sci. Appl.*, vol. 14, no. 3, pp. 586–591, 2023.
https://doi.org/10.14569/IJACSA.2023.0140368

[9] F. O. Awalullaili, D. Ispriyanti, and T. Widiharih, "Klasifikasi Penyakit Hipertensi Menggunakan Metode Svm Grid Search Dan Svm Genetic Algorithm (Ga)," *J. Gaussian*, vol. 11, no. 4, pp. 488–498, 2023.
https://doi.org/10.14710/j.gauss.11.4.488-498

[10] X. Xu, Z. Xie, Z. Yang, D. Li, and X. Xu, "A t-SNE Based Classification Approach to Compositional Microbiome Data," *Front. Genet.*, vol. 11, no. December, pp. 1–10, 2020.
https://doi.org/10.3389/fgene.2020.620143

[11] D. Kobak and P. Berens, "The art of using t-SNE for single-cell transcriptomics," *Nat. Commun.*, vol. 10, no. 1, 2019.
https://doi.org/10.1038/s41467-019-13056-x

[12] Al Danny Rian Wibisono, Syahrul Hidayat, Humam Maulana Tsubasanofa Ramadhan, and Eva Yulia Puspaningrum, "Comparison of K-Nearest Neighbor and Decision Tree Methods using Principal Component Analysis Technique in Heart Disease Classification," *Indones. J. Data Sci.*, vol. 4, no. 2, pp. 90–100, 2023.
https://doi.org/10.56705/ijodas.v4i2.70

[13] N. Hidayati and A. Hermawan, "K-Nearest Neighbor (K-NN) algorithm with Euclidean and Manhattan in classification of student graduation," *J. Eng. Appl. Technol.*, vol. 2, no. 2, pp.

86–91, 2021.
http://dx.doi.org/10.21831/jeatech.v2i2.42777
[14] I. G. I. Sudipa *et al.*, *Teknik Visualisasi Data*. PT. Sonpedia Publishing Indonesia, 2023.
https://books.google.co.id/books?id=LjC4EAAAQBAJ&lpg=PA16&dq=visualisasi%20data%20penting&lr&pg=PA15#v=onepage&q=visualisasi%20data%20penting&f=false
[15] R. Rianti, R. Andarsyah, and R. M. Awangga, "Penerapan PCA dan Algoritma Clustering untuk Analisis Mutu Perguruan Tinggi di LLDIKTI Wilayah IV," *Nuansa Inform.*, vol. 18, no. 2, pp. 67–77, 2024.
https://doi.org/10.22146/ijccs.65176
[16] P. S. Rao, D. N. Malleswari, K. S. Rao, B. S. Babu, and K. Saikumar, "The Impact of PCA and t-SNE on the Predictive Accuracy of k- NN , Naive Bayes , and LDA : A Study Using the Legal Medicine Legal Medicine Dataset," vol. 27, no. 2, pp. 68–80, 2024.
https://ijmtlm.org/index.php/journal/article/view/168
[17] M. C. Cieslak, A. M. Castelfranco, V. Roncalli, P. H. Lenz, and D. K. Hartline, "t-Distributed Stochastic Neighbor Embedding (t-SNE): A tool for eco-physiological transcriptomic analysis," *Mar. Genomics*, vol. 51, p. 100723, Jun. 2020.
https://doi.org/10.1016/j.margen.2019.100723
[18] T. Unger *et al.*, "2020 International Society of Hypertension Global Hypertension Practice Guidelines," *Lippincott Williams & Wilkins*, vol. 75, no. 6, pp. 1334–1357, 2020.
https://doi.org/10.1161/HYPERTENSIONAHA.120.15026
[19] M. Rahmadhani, "Faktor-Faktor Yang Mempengaruhi Terjadinya Hipertensi Pada Masyarakat Di Kampung Bedagai Kota Pinang," *J. Kedokt. STM (Sains dan Teknol. Med.*, vol. 4, no. 1, pp. 52–62, 2021.
https://doi.org/10.30743/stm.v4i1.132
[20] T. K. S. Jaya, "Hubungan nilai tekanan Darah dan Frekuensi Nadi dengan Kualitas Hidup Penderita Hipertensi," *Univ. Muhammadiyyah Surakarta*, vol. 1, p. 8, 2021.
http://eprints.ums.ac.id/id/eprint/93232
[21] G. Melliya Sari, V. Eko Kurniawan, E. Puspita, and S. Devi Amalia, "Hubungan Indeks Massa Tubuh Dengan Tekanan Darah Pada Penderita Hipertensi Di Poli Jantung Rumah Sakit Husada Utama Surabaya," *Prima Wiyata Heal.*, vol. 4, no. 1, pp. 47–63, 2023.
https://doi.org/10.60050/pwh.v4i1.39
[22] F. Fantin, A. Giani, E. Zoico, A. P. Rossi, G. Mazzali, and M. Zamboni, "Weight loss and hypertension in obese subjects," *Nutrients*, vol. 11, no. 7, 2019.
https://doi.org/10.3390/nu11071667
[23] Z. Thakker and B. Harshadkant, "Effect of Feature Scaling Pre-processing Techniques on Machine Learning Algorithms to Predict Particulate Matter Concentration for Gandhinagar, Gujarat, India," *Int. J. Sci. Res. Sci. Technol.*, pp. 410–419, 2024.
https://doi.org//10.32628/IJSRST52411150
[24] D. B. G. N. Singh and A. Bandyopadhyay, "Robust estimation strategy for handling outliers," *Commun. Stat. - Theory Methods*, vol. 0, no. 0, pp. 1–20, 2023.
https://doi.org/10.1080/03610926.2023.2218567
[25] A. R. Isnain, J. Supriyanto, and M. P. Kharisma, "Implementation of K-Nearest Neighbor (K-NN) Algorithm For Public Sentiment Analysis of Online Learning," *IJCCS (Indonesian J. Comput. Cybern. Syst.*, vol. 15, no. 2, p. 121, 2021.
https://doi.org/10.22146/ijccs.65176
[26] J. Bergler-Klein, "What's new in the ESC 2018 guidelines for arterial hypertension: The ten most important messages," *Wien. Klin. Wochenschr.*, vol. 131, no. 7–8, pp. 180–185, 2019.
https://doi.org/10.1007/s00508-018-1435-8
[27] F. Ardiansyah, F. Hamdan, S. Sugiyanto, and I. Wahyu Siadi, "Klasifikasi Customer Relationship Management Menggunakan Dataset KDD Cup 2009 dengan Teknik Reduksi Dimensi," *Komputika J. Sist. Komput.*, vol. 11, no. 2, pp. 193–202, 2022.
https://doi.org/10.34010/komputika.v11i2.6498
[28] D. N. Aini, B. Oktavianti, M. J. Husain, D. A. Sabillah, S. T. Rizaldi, and M. Mustakim, "Seleksi Fitur untuk Prediksi Hasil Produksi Agrikultur pada Algoritma K-Nearest Neighbor (KNN)," *J. Sist. Komput. dan Inform.*, vol. 4, no. 1, p. 140, 2022.
https://doi.org/10.30865/json.v4i1.4813
[29] M. Rizky Adriansyah, M. Reza Faisal, A. Gafur, R. Adi Nugroho, I. Budiman, and M. Muliadi, "Implementasi Reduksi Fitur t-SNE Pada Clustering Gambar Head shape Nematoda," *J. Komputasi*, vol. 10, no. 1, pp. 54–64, 2022.
https://doi.org/10.23960/komputasi.v10i1.2963
[30] B. Firmanto, H. Soekotjo, and H. Suyono, "Perbandingan Kinerja Algoritma Promethee Dan Topsis Untuk Pemilihan Guru Teladan," *J. Penelit. Pendidik. IPA*, vol. 2, no. 1, 2016.
https://www.academia.edu/download/113421197/31.pdf
[31] B. M. S. Hasan and A. M. Abdulazeez, "A Review of Principal Component Analysis Algorithm for Dimensionality Reduction," *J. Soft Comput. Data Min.*, vol. 2, no. 1, pp. 20–30, 2021.
https://doi.org/10.30880/jscdm.2021.02.01.003
[32] A. Platzer, "Visualization of SNPs with t-SNE," *PLoS One*, vol. 8, no. 2, 2013, doi: 10.1371/journal.pone.0056883.
https://doi.org/10.1371/journal.pone.0056883
[33] H. Hafid, "Penerapan K-Fold Cross Validation untuk Menganalisis Kinerja Algoritma K-Nearest Neighbor pada Data Kasus Covid-19 di Indonesia," *J. Math.*, vol. 6, no. 2, pp. 161–168, 2023, [Online].
https://doi.org/10.35580/jmathcos.v6i2.53043
[34] H. Azis, P. Purnawansyah, F. Fattah, and I. P. Putri, "Performa Klasifikasi K-NN dan Cross Validation Pada Data Pasien Pengidap Penyakit Jantung," *Ilk. J. Ilm.*, vol. 12, no. 2, pp. 81–86, 2020.
https://doi.org/10.33096/ilkom.v12i2.507.81-86
[35] S. Dewi and M. A. I. Pakereng, "Implementasi Principal Component Analysis Pada K-Means Untuk Klasterisasi Tingkat Pendidikan Penduduk Kabupaten Semarang," *JIPI (Jurnal Ilm. Penelit. dan Pembelajaran Inform.*, vol. 8, no. 4, pp. 1186–1195, 2023.
https://doi.org/10.29100/jipi.v8i4.4101
[36] S. Abimanyu, N. Bahtiar, and E. Adi Sarwoko, "Implementasi Metode Support Vector Machine (SVM) dan t-Distributed Stochastic Neighbor Embedding (t-SNE) untuk Klasifikasi Depresi," *J. Masy. Inform.*, vol. 14, no. 2, pp. 146–158, 2023.
https://doi.org/10.14710/jmasif.14.2.59513
[37] D. D. W, "Dimensionality Reduction : LDA, PCA, t-SNE," *Medium*, 2021.
https://medium.com/analytics-vidhya/dimensionality-reduction-pca-vs-lda-vs-t-sne-681636bc686
[38] Sachsoni, "Mastering t-SNE(t-distributed stochastic neighbor embedding)," *Medium*, 2024.
https://medium.com/@sachinsoni600517/mastering-t-sne-t-distributed-stochastic-neighbor-embedding-0e365ee898ea
[39] Unknown, "t-SNE and PCA: Two powerful tools for data exploration," *Fabrizio Musacchio*, 2023. https://www.fabriziomusacchio.com/blog/2023-06-12-tsne_vs_pca/ (accessed May 14, 2024).
[40] A. Chawla, "The Ultimate Comparison Between PCA and t-SNE Algorithm," *Daily Dose of Data Science*, 2023. https://blog.dailydoseofds.com/p/the-ultimate-comparison-between-pca.

