Recursive Journal of Informatics
Vol. 3, No. 2, September 2025
p-ISSN 2963-5551 https://journal.unnes.ac.id/journals/index.php/rji e-ISSN 2986-6588

Implementation of the K-Nearest Neighbor Algorithm (KNN) with Principal Component Analysis to Diagnose Tuberculosis

Yuliana Putri[1], Alamsyah[2]
[1, 2]Computer Science Department, Faculty of Mathematics and Natural Sciences,
Universitas Negeri Semarang, Indonesia

Abstract.
Purpose: Tuberculosis (TB) is an infectious disease that attacks the respiratory organs, the lungs, and some can attack organs outside the lungs. Indonesia is one of the largest contributors to TB cases with around 320,000 new cases every year. Delays in diagnosing TB disease can cause a higher number of deaths due to errors in the treatment of sufferers. This makes the early diagnosis of TB disease important as early as possible. The research carried out aims to implement machine learning techniques to help diagnose TB disease.
Methods: The research was carried out using the *K-Nearest Neighbor* (KNN) classification algorithm which was optimized with the *Principal Component Analysis* (PCA) feature selection technique. The dataset used consists of 577 data with 12 attributes labeled patients with tuberculosis and patients who do not have tuberculosis.
Result: From the research that has been conducted, models that implement the KNN algorithm with PCA produce models with better performance than models that only implement KNN. The model that only uses KNN gets an accuracy of 92.528%, while the model that uses KNN and PCA gets an accuracy of 98.85%. This shows that the implementation of KNN and PCA is able to produce a good tuberculosis diagnosis model and can be used to assist in the early diagnosis of tuberculosis.
Novelty: Using PCA in the feature selection process can reduce unnecessary attributes. It is a PCA that helps reduce the dimensionality, simplifies the visualization and interpretation of complex data sets. The use of PCA has been proven to be able to optimize the performance of the KNN algorithm for the detection of tuberculosis.

Keywords: Machine Learning, Tuberculosis, Data Mining, Disease Diagnosis, KNN

Received May 2024 / **Revised** September 2025 / **Accepted** September 2025

INTRODUCTION

Tuberculosis (TB) is an infectious disease caused by a bacterial infection called the mycobacterium tuberculosis virus. This disease attacks the respiratory organs of the lungs, and some can attack organs outside the lungs, such as lymph nodes and the lining of the brain [1]. Based on the WHO global tuberculosis report 2022, Indonesia is one of the countries contributing the largest number of TB cases. Indonesia has a high burden of tuberculosis, with approximately 320,000 new cases each year. However, there has been progress in efforts to prevent and treat tuberculosis in Indonesia, such as increasing access to diagnosis and treatment, as well as the use of more effective combination therapy. Tuberculosis (TB) in Indonesia is third after India and China, with a total of 824 thousand cases and 93 thousand deaths per year or the equivalent of 11 deaths per hour. Based on the 2022 Global Tuberculosis Report, the highest number of TB cases is in the productive age group, especially those aged 25 to 34 years.

TB disease has symptoms that are almost the same as other lung diseases. Therefore, errors can occur in providing a diagnosis for sufferers. Therefore, self-examination and treatment are needed as soon as possible, if someone often has contact with TB sufferers and experiences the symptoms they experience. If it is too late to treat TB disease, things like this can cause an even higher number of deaths due to errors in treating sufferers. Therefore, it is important to make an early diagnosis of tuberculosis so that it can be treated immediately according to good health procedures.

[1]*Corresponding author.
Email addresses: anayuliputri@students.unnes.ac.id (Putri)
DOI: 10.15294/rji.v3i2.5235

The level of precision in diagnosing the disease is very important and is needed. One of them is tuberculosis, which is a problem in the world of health. The combination of technology and data has changed the way we understand and interact with the world. The development of data today means that data is increasingly being used to solve problems, one of which is problems in the health sector. The development of the amount of data stored in large databases cannot be separated from information technology, which is increasingly global. The large amount of data that exists requires appropriate data processing methods so that the data can be used optimally, one of which is data mining techniques. Data mining can be applied in various fields that have quite large data. Data mining itself is widely used to diagnose diseases, one of which is using a machine learning approach. Machine learning itself includes classification, clustering, prediction, and finding associations [2].

Data mining algorithms such as classification grouping, and so on, can be used to find patterns to determine future business movements [3]. One algorithm that is often used to solve diagnostic problems is the *K-Nearest-Neighbor* (KNN) classification algorithm. Classification is a type of supervised machine learning in which the labeled data in advance [4]. The KNN algorithm is a generalized algorithm with nearest-neighbors rules. The inductive offset is the class label of the *k* samples with the class label to be tested that is most similar to the closest one [5]. KNN can also be interpreted as an algorithm used to classify objects based on mining data that has the closest distance to testing data [6]. This method requires a distance measure to determine the proximity of the object. An algorithm that groups new data whose class is not yet known by selecting the *k* data that has the closest value to the new data. Predictions are made based on the class of test data that appears most frequently among *k* neighbor [7]. In the context of inductive offset, KNN also considers the class labels of the k-nearest samples, making it more contextual in determining predictions. Therefore, KNN is not just a classification algorithm, but also an analysis tool that understands the context of the data considering the existence of nearest neighbor.

There have been many previous studies that have applied machine learning approaches to diagnose diseases. One of them is the research conducted by Zalloum et al. in 2022. This research uses machine learning techniques to diagnose breast cancer. This research uses machine learning algorithms, namely k-nearest neighbor (KNN), logistic regression, decision tree, random forest, and support vector machine (SVM), to classify breast cancer (BC) data and PCA is applied as feature extraction. The final results of this research were that KNN obtained 96% accuracy, logistic regression obtained 97% accuracy, Decision Tree obtained 96% accuracy, random forest obtained 96% accuracy, and SVM obtained 98% accuracy [8].

Other research was also conducted by Refaat et al. in 2022. This research discusses the use of machine learning techniques to diagnose brain tumors. This research was carried out using machine learning algorithms, namely k-nearest neighbor (KNN), general regression neural network (GRNN), dan support vector machine (SVM), to classify with PCA to extract features from the brain tumor dataset. The final results of this research showed that SVM achieved an accuracy of 0.9624%, GRNN accuracy of 94.7%; KNN accuracy 97%. The KNN algorithm was concluded to provide the best diagnostic accuracy, because the diagnostic accuracy was 97% and the error rate was 3%; the minimum observed objective reached 0.059 at 30 function evaluations; and the KNN computing time is 172.01[9].

From several studies that have been carried out, it is known that the KNN algorithm is capable of producing models with good performance when using the PCA feature selection technique. Principal Component Analysis (PCA), also called the Karhunen loeve transformation, is a technique used to simplify data, by transforming the data linearly to form a new coordinate system with maximum variance [10]. In implementing PCA, variables that are initially large in number can be reduced to a smaller number of principal components, but still retain as much information as possible from the dataset. This is done by transforming the data into smaller dimensions, which act as feature summaries [11]. The most important reason for using PCA is to create high-quality feature groups and reduce data size [12]. By reducing data dimensions, PCA can help identify critical attributes that contribute significantly to class formation in a data object. Therefore, the integration of PCA with the KNN algorithm can create a more holistic and efficient approach.

Based on the background of the problems described previously, this research was carried out with the aim of building a tuberculosis diagnosis model using a machine learning approach. The research will be carried out using the KNN classification algorithm and also the PCA feature selection technique. By combining the power of classification algorithms such as KNN and dimensionality reduction techniques such as PCA,

the horizon in disease diagnosis can be increased, and precision and efficiency in TB treatment can be increased.

METHODS

Dataset

This research data uses secondary data in the form of the "Tuberculosis Dataset" dataset. The data source in this study was collected from information available on Kaggle such as tuberculosis symptoms data. This data set has 12 variables, namely sex, age, cough for two weeks, night sweats, weight loss, appetite loss, TB contact history, cough with phlegm, coughing blood, BCG results appear fast, lumps that appear around the armpits and neck, and results.

Table 1. Description of the attributes of the tuberculosis data set

Attribute	Description	Attribute Value	Attribute Type
Sex	Gender	Male, Female	Nominal
Age	Age (years)	[10,…,70]	Numeric
cough for two weeks	History of cough for two weeks	0 = No, 1 = Yes	Nominal
night sweats	History of excessive	0 = No, 1 = Yes	Nominal
weight loss	History of weight loss	0 = No, 1 = Yes	Nominal
loss of appetite	History of loss of appetite	0 = No, 1 = Yes	Nominal
TB contact history	History of contact with someone who has been infected with Tuberculosis	0 = No, 1 = Yes	Nominal
cough with phlegm	History of phlegm cough with phlegm	0 = No, 1 = Yes	Nominal
coughing blood	History of phlegm cough with phlegm	0 = No, 1 = Yes	Nominal
BCG results appear fast	History of the BCG (Bacillus Calmette-Guérin) vaccine	0 = No, 1 = Yes	Nominal
lumps that appear around the armpits and neck	History of lumps appearing around the armpits and neck	0 = No, 1 = Yes	Nominal
Result	Tuberculosis results	0 = No, 1 = Yes	Nominal

The target attribute represents a value of 1 for cases infected with tuberculosis and 0 for cases not infected with tuberculosis. This data set has the.csv format as in Figure 1.

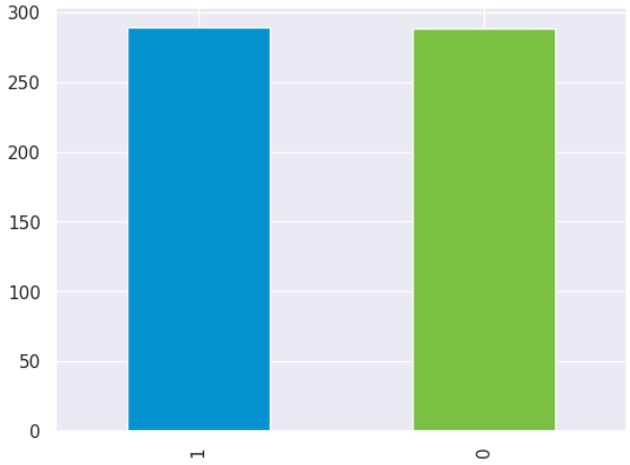

Figure 1. Visualization of target attribute data balance on the dataset

Data Preprocessing

Before the data classification process is carried out, it is necessary to go through a preprocessing process so that the data can be read by the programming language, and this process can also increase the accuracy value of the data mining model. The pre-processing stage consists of the following stages:

Data type conversion

In the tuberculosis dataset, data type conversion and label encoding stages were carried out to prepare the data before analysis. Some attributes will be converted from integer to object data types to facilitate a better analysis process. In addition, the 'sex' attribute will undergo a label encoding process, where this column is separated into two separate columns: 'male' and 'female'. A value of 1 indicates the presence of the appropriate gender, while a value of 0 indicates the opposite. This stage helps in preparing data for further analysis.

Cleaning data

Cleaning data is used to identify and improve the data to be researched. Data correction is carried out because raw data is often not ready for mining, such as missing values and duplicate data in the dataset used. Missing values in the tuberculosis dataset usually come from data that has no value or information on its attributes. Meanwhile, duplicate data occurs due to repetition of the same entries in the dataset.

Data standardization

In the data standardization process, the main goal is to change the values in the dataset so that the distribution has a mean (average) of zero and a standard deviation of one. This means that each value in the dataset will be changed relative to the mean and spread (variability) of the data. By standardizing the data set, data will be obtained that has a uniform scale, where different values will not have very large differences. This helps ensure that the influence of each variable in data analysis is balanced and not too dominated by large scale variables. Thus, data that has been standardized will be easier to interpret and process in data analysis.

Split Data

The Split Data stage is the stage of dividing the data into training data and testing data. Training data is used for classification to form a model. Meanwhile, testing data is used to test the extent to which classification can be carried out correctly. In this study, a data proportion of 70:30 was used. Dividing data for training and testing with a ratio of 70:30 is the best testing ratio to get the best performance of the machine learning model. Additionally, it is known that the percentage of data in the training dataset increases (Nguyen et al., 2021). Data distribution is carried out in consistent randomization (random state) with the aim that each time the calculation is carried out the value is constant.

Feature Selection With PCA

Data sets that have gone through the cleaning process will then undergo a PCA feature selection process that aims to inherit the variations that exist in the data set using a number of small factors. Each principal component (PC) is a combination of the original attributes, which maintain some correlation between each attribute, and mutually orthogonal PCs (Granato et al., 2018).

Step 1: The tuberculosis dataset will be standardized or normalized. The goal is to ensure that each feature in the dataset has a mean value of zero and a standard deviation of one, so that variables with different scales make a balanced contribution to the PCA analysis.

Step 2: Calculate the covariance matrix, search for values and eigenvectors from the matrix, and project the data into a new feature space using the principal components found.

Step 3: Find the eigenvalues and eigenvectors of the matrix. Eigenvalues represent the variance of each principal component, while eigenvectors indicate the direction of each principal component in feature space.

Step 4: The dataset will project the data into a new feature space consisting of the main components that have been discovered previously. The most significant principal components are chosen to represent the main dimensions of the data, so that the original data can be reduced to lower dimensions.

Step 5: Evaluate the results of data projection into the new feature space. Once the data is projected into lower dimensions based on the selected principal components, it is important to evaluate how well this representation explains the variation in the original data.

The PCA can be seen in Figure 2.

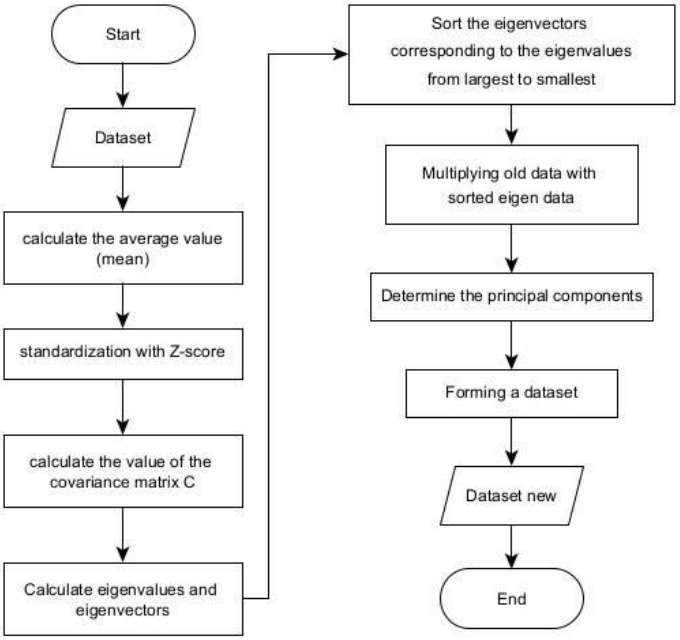

Figure 2. Flowchart of the PCA feature selection technique

Implementation of the KNN Classification Algorithm

The KNN classification stages of the research can be seen in Figure 3.

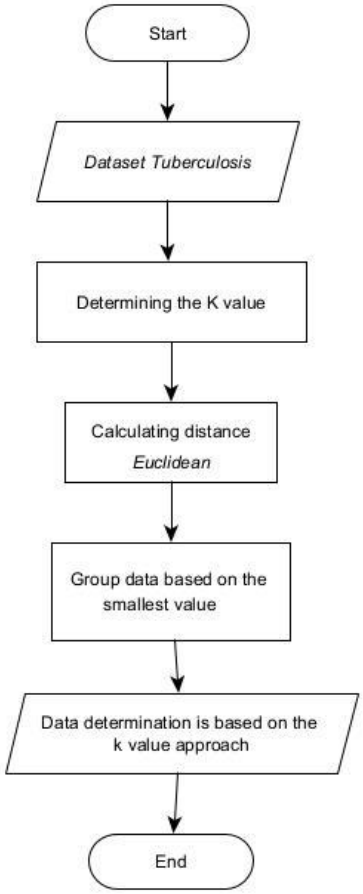

Figure 3. Flowchart of the KNN classification algorithm

The KNN algorithm begins by preparing a tuberculosis dataset as input. Then, it will be processed by determining the *k* value of the data to be classified. Next, calculate the Euclidean distance using a predetermined formula and group it based on the smallest value of the results obtained. Then, the final results are based on determining the data set using the *k* value approach.

KNN algorithm classification with PCA

At this stage we will combine the PCA method with the KNN algorithm to determine the resulting level of accuracy. The flowchart of the combination of methods used in this research can be seen in Figure 4.

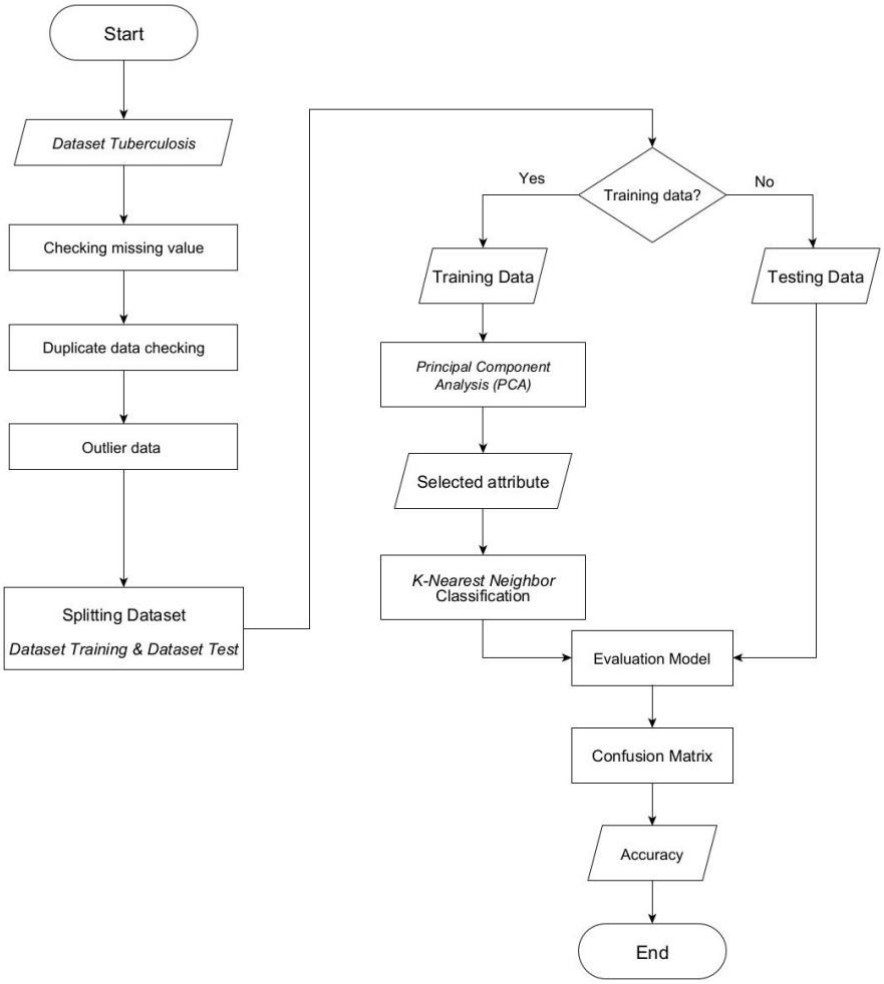

Figure 4. Flowchart Algorithm KNN with PCA

1) Prepare the dataset. The dataset used in this research is the tuberculosis dataset taken from Kaggle with a total of 577 data and 12 attributes with 1 class label.
2) Carrying out pre-processing stages consisting of data type conversion and data cleaning.
3) Divide the dataset into training data and test data using the split data method with a proportion of 70% for training data and 30% for testing data.
4) Carry out PCA feature selection, so that the best attributes are obtained which will be continued in the classification process.
5) Then carry out the classification process using the KNN algorithm to produce a new KNN classification model.
6) Next is to test the testing data.
7) Based on this test, analysis of the results can be carried out by calculating the accuracy obtained using a confusion matrix. So that the accuracy and accuracy of the KNN algorithm can be obtained by applying PCA feature selection to the tuberculosis dataset.

Model Performance Evaluation

The model performance evaluation stage is carried out using a confusion matrix. The confusion matrix is a method that can be used to measure the performance of a classification method. Confusion matrix is also one way of visualizing system learning results; the visualization displayed contains two or more categories [13]. The confusion matrix performs 4 calculations, namely recall, precision, accuracy, and error rate. The table below is an example of confusion matrix results predicting two classes. The table of the confusion matrix can be seen in Table 1.

Table 1. Confusion matrix table

Classification		Predicted Values	
		1 (Positive)	0 (Negative)
Actual Values	1 (Positive)	True Positive (TP)	False Negative (FN)
	0 (Negative)	False Positive (FP)	True Negative (TN)

Accuracy calculations carried out by the confusion matrix based on the table above can use the following equation.

Calculating the accuracy value is stated in Equation 1.

$$Accuracy = \frac{\text{TP+TN}}{\text{TP+FN+FP+TN}} \times 100\% \quad (1)$$

Calculating the recall value is expressed in Equation 2.

$$Recall = \frac{\text{TP}}{\text{(TP+FP)}} \quad (2)$$

Calculating the precision value is expressed in Equation 3.

$$Precision = \frac{\text{TP}}{\text{(TP+FN)}} \quad (3)$$

Calculating the error rate value is expressed in Equation 4.

$$Error\ rate = \frac{\text{FP+FN}}{\text{TP+TN+FP+FN}} \quad (4)$$

RESULT AND DISCUSSION

Data Preprocessing

Data type conversion results

At this stage, change the column data type cough for two weeks, weight loss, loss of appetite, night sweats, TB contact history, cough with phlegm, coughing blood, BCG results appear fast, lumps that appear around the armpits and neck, and results from integers to objects is a very important step in ensuring the integrity and accuracy of the data which will be shown in Table 2.

Table 2. Visualization before data type conversion

No	Attribute	Type Attribute
1.	Sex	Object
2.	Age	Integer
3.	Cough for Two Weeks	Integer
4.	Night Sweats	Integer
5.	Weight Loss	Integer
6.	Loss of Appetite	Integer
7.	TB Contact History	Integer
8.	Cough with Phlegm	Integer
9.	Coughing Blood	Integer
10.	BCG Results Appear Fast	Integer
11.	Lumps That Appear Around the Armpits and Neck	Integer
12.	Result	Object

The steps implemented to manage the conversion of this data type are key in ensuring more representative analysis results. Therefore, it is important to check each value in that column and convert it to a more appropriate object data type, which can be visualized through the relevant processes that can be seen in Table 3.

Table 3. Visualization after data type conversion

No	Attribute	Type Attribute	
		Before	After
1.	Sex	Object	Object
2.	Age	Integer	Integer
3.	Cough for Two Weeks	Integer	Object
4.	Night Sweats	Integer	Object
5.	Weight Loss	Integer	Object
6.	Loss of Appetite	Integer	Object
7.	TB Contact History	Integer	Object
8.	Cough with Phlegm	Integer	Object
9.	Coughing Blood	Integer	Object
10.	BCG Results Appear Fast	Integer	Object
11.	lumps that appear around the armpits and neck	Integer	Object
12.	Result	Object	Object

The next step is to separate the 'sex' column into two separate columns, namely 'male' and 'female'. The 'male' column will contain the value 1 if the gender is male, and 0 if not, while the 'female' column will contain the value 1 if the gender is female, and 0 if not. These stages can be seen in Table 4 and Table 5.

Table 4. Value of the sex attribute before separation

Attribute	Value
Sex	0
	1

Table 5. Value of the sex attribute after separation

Attribute	Value
Female	0
	1
Male	0
	1

Checking missing values

The steps implemented to handle missing values are key in ensuring more representative analysis results. Therefore, it is important to check each attribute of the data for the presence of missing values, which can be visualized through Figure 5.

```
perempuan                                  0
laki-laki                                  0
age                                        0
cough for two weeks                        0
night sweats                               0
weight loss                                0
loss of appetite                           0
TB contact history                         0
cough with phlegm                          0
coughing blood                             0
BCG results appear fast                    0
lumps that appear around the armpits and neck    0
result                                     0
dtype: int64
```

Figure 5. Visualization of missing value checking

Figure 4 shows a visualization of the results of the missing value check process on the tuberculosis dataset. You can see a clear picture of the distribution of missing values before and after treatment, that the values in the dataset do not have missing values.

Duplicate data check
The results of the verification of duplicate data can be seen in Figure 6.

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 386 entries, 6 to 576
Data columns (total 13 columns):
 #   Column                                      Non-Null Count  Dtype
---  ------                                      --------------  -----
 0   perempuan                                   386 non-null    uint8
 1   laki-laki                                   386 non-null    uint8
 2   age                                         386 non-null    int64
 3   cough for two weeks                         386 non-null    object
 4   night sweats                                386 non-null    object
 5   weight loss                                 386 non-null    object
 6   loss of appetite                            386 non-null    object
 7   TB contact history                          386 non-null    object
 8   cough with phlegm                           386 non-null    object
 9   coughing blood                              386 non-null    object
 10  BCG results appear fast                     386 non-null    object
 11  lumps that appear around the armpits and neck  386 non-null  object
 12  result                                      386 non-null    int64
dtypes: int64(2), object(9), uint8(2)
memory usage: 36.9+ KB
```

Figure 6. Visualization of checking duplicate data

Figure 5 shows a visualization of the double data check results of the process in the tuberculosis dataset. A clear picture of the distribution of duplicates before and after handling can be seen, and these duplicate data are retained. From these results, it shows that this dataset is still suitable to proceed to the next stage in the classification process.

Label encoding results
This phase aims to change the form of data from categorical (in the form of letters) to numeric (in the form of numbers) because data in categorical form cannot be read by the model created. Each categorical data will be changed to numeric which depends on the number of values that differ from one another. The data will be converted into numbers 0 and 1. The attribute whose data will be changed from categorical to numeric is result. Table 6 shows the values of the result attribute before label encoding is carried out.

Table 6. Value attribute results before label encoding

Attribute	Value
Result	Negative
	Positive

Next, the data values in categorical form will be replaced in numerical form, where negative values are changed to 0 while positive values are changed to 1. Table 7 shows the values of the result attribute after label encoding.

Table 7. Value attribute results after label encoding

Attribute	Value
Result	0
	1

The results of the label encoding process for the tuberculosis dataset can be seen in Figure 7.

TB contact history	cough with phlegm	coughing blood	BCG results appear fast	lumps that appear around the armpits and neck	result
0	1	0	0	0	0
0	1	0	0	0	0
0	1	0	0	0	0
1	1	0	0	0	1
0	1	0	0	0	0

Figure 7. Example of a tuberculosis dataset after label encoding

Data transformation results

In this research, the data transformation process plays an important role in improving the quality of the tuberculosis dataset before in-depth analysis is carried out. Data transformation involves a series of steps to change the format or structure of data to make interpretation and analysis easier. At this stage, data standardization is carried out using a standard scaler, especially for variable x (predictor variable) or all attributes other than the result (target) in the tuberculosis dataset. Visualization of data standardization can be seen in Figure 8.

```
          0         1         2    3         4    5         6         7
0 -1.642778  1.642778  1.637285  0.0 -1.766757  0.0  0.65979 -0.552657
1  0.608725 -0.608725 -0.737390  0.0  0.566009  0.0  0.65979 -0.552657
2  0.608725 -0.608725  0.823111  0.0  0.566009  0.0  0.65979 -0.552657
3  0.608725 -0.608725  0.958807  0.0  0.566009  0.0  0.65979 -0.552657
4  0.608725 -0.608725  1.094502  0.0  0.566009  0.0  0.65979 -0.552657

          8         9         10        11
0  0.234189 -0.793690 -0.102508 -0.102508
1  0.234189  1.259938 -0.102508 -0.102508
2  0.234189  1.259938 -0.102508 -0.102508
3  0.234189  1.259938 -0.102508 -0.102508
4  0.234189  1.259938 -0.102508 -0.102508
```

Figure 8. Visualization of data standardization

Figure 8 presents a visualization of the results of the data transformation process, especially data standardization on variable x or result (target) in the tuberculosis dataset. The graph or diagram displayed depicts changes in the distribution of variable values after going through the standardization process. This process increases the consistency and interpretability of the dataset, facilitating further analysis of factors that contribute to tuberculosis disease.

Feature Selection Results with PCA

This standardization is important in the context of health data analysis because it ensures that each attribute has a balanced contribution to the analysis, allowing for more accurate PCA results. The data standardization process is carried out using the scaler standardization method. The stages of the standardization process begin with finding the average value for each attribute of the data set used, followed by calculating the standard deviation for each attribute. The results obtained by calculating the mean value for each attribute are shown in Table 8.

Table 8. Mean value for each attribute

Attribute	Mean Value	Attribute	Mean Value
Female	0.724	Loss of Appetite	0.712
Male	0.2754	TB Contact History	0.240
Age	34.739	Cough with Phlegm	0.945
Cough for Two Weeks	1.0	Coughing Blood	0.384
Night Sweat	0.744	BCG Results Appear Fast	0.0099
Weight Loss	1.0	Lumps That Appear Around the Armpits and Neck	0.0099

Next, the standard deviation of each feature is also calculated to assess the distribution of the data in each dimension. The results of the calculation of the standard deviation for each attribute are shown in Table 9.

Table 9. Sample standard deviation calculation

Attribute	Standard deviation
Female	1.0000000000000002
Male	1.0
Age	1.0
Cough for Two Weeks	0.0
Night Sweats	1.0
Weight Loss	0.0
Loss of Appetite	0.9999999999999999
TB Contact History	1.0000000000000002
Cough with Phlegm	1.0000000000000002
Coughing Blood	1.0
BCG Results Appear Fast	0.9999999999999999
Lumps That Appear Around the Armpits and Neck	0.9999999999999999

Next, look for the Z-score value using the mean and standard deviation that was calculated previously. The results of the normalization of the data are shown in Table 10.

Table 10. Sample normalization results

No	Female	Age	…	Night Sweats	TB contact history
0.	-1.641353	1.635866	…	-1.765225	-0.552178
1.	0.608197	-0.736751	…	0.565518	-0.552178
2.	0.608197	0.822397	…	0.565518	-0.552178
3.	0.608197	0.957975	…	0.565518	-0.552178
4.	0.608197	1.093554	…	0.565518	-0.552178
…	…	…	…	…	…
572	0.608197	0.144507	…	-1.765225	-0.552178
573	-1.641353	-0.397805	…	0.565518	1.807872
574	-1.641353	-0.804540	…	0.565518	-0.552178
575	-1.641353	-0.533383	…	0.565518	-0.552178
576	-1.641353	-0.668962	…	0.565518	-0.552178

After normalizing using the standard deviation, the covariance matrix is calculated. The following are the calculation of the results of the covariance matrix that are shown in Table 11.

Table 11. Sample results of calculating the covariance matrix

No	1	2	3	…	9	10	11
0	1.0	-1.0	0.28	…	0.0	-0.13	0.20
1	-1.0	1.0	-0.28	…	0.0	0.13	-0.20
2	0.28	-0.28	1.0	…	0.0	0.35	0.013
3	0.0	0.0	0.0	…	0.0	0.0	0.0
4	-0.13	0.13	-0.35	…	0.0	1.0	-0.26
5	0.0	0.0	0.0	…	0.0	0.0	0.0
6	0.2	-0.20	0.01	…	0.0	-0.26	1.0
7	0.08	-0.08	-0.38	…	0.0	0.22	0.36
8	-0.14	0.14	-0.15	…	0.0	0.41	-0.15
9	0.20	-0.20	-0.22	…	0.0	0.32	0.52
10	0.06	-0.6	-0.15	…	0.0	-0.18	0.06
11	0.06	-0.6	-0.15	…	0.0	0.18	0.06

After calculating the covariance matrix, the next step is to calculate the eigenvalues and eigenvectors of the covariance matrix. The eigenvalue and the eigenvector contain important information from the tuberculosis dataset to determine the principal components (PC) formed. In this investigation, the main component, PC is selected based on the percentage of variance explained. The variance explained shows how much information is explained by the selected main components.

The number of PCs selected for classification using KNN is determined by the parameter n. Researchers conducted trials with several n values, from n = 12 to n = 1. The quality of each experiment was assessed using the KNN algorithm, and the n value that provided optimal accuracy was selected. From the results of this experiment, the n value that provides the best quality will be selected. carried out, the number n = 8 produces the most optimal accuracy. The accuracy results for each n can be seen in Table 12.

Table 12. Accuracy results for each n in PCA

n	Accuracy	Total Variance	n	Accuracy	Total Variance
12	0.150	1.0	6	0.124	0.9997
11	0.147	1.0	5	0.132	0.9995
10	0.167	1.0	4	0.136	0.9989
9	0.165	1.0	3	0.166	0.9978
8	0.236	1.0	2	0.166	0.9966
7	0.122	0.9999	1	0.166	0.9945

Of the 11 components that have been sorted, several components were selected that retained information or explained the variance of at least 70% of the data set. Based on the data from Table 12, 12 PC components were selected. The variance resulting from these 8 components can be seen in Table 13.

Table 13. Variance explained PC

	PC	Eigenvalue	Explained Variance	Cumulative Variance
0				
1	PC1	2.48913	0.2484820	0.2484820
2	PC2	2.16776	0.2164011	0.4648832
3	PC3	1.98167	0.1978238	0.6627071
4	PC4	1.27148	0.1269282	0.7896354
5	PC5	0.85977	0.0858287	0.8754641
6	PC6	0.67757	0.0676397	0.9431039
7	PC7	0.40083	0.0400140	0.9831180
8	PC8	0.16911	0.0168819	1.0

Based on the results of this explanation, the results of the explained variance and eigenvalue of each main component (PC) in the PCA analysis are presented. Based on these results, it was decided to use 8 main components, with the selected attributes being male, female, age, night sweats, loss of appetite, TB contact history, cough with phlegm, coughing blood, BCG results appear fast, and lumps around armpits and neck, of the main components aim to retain most of the information in the tuberculosis dataset. The accuracy results of the model using 8 main components show good performance, with an accuracy level of 23.6% and a total variance of 1.0. This decision is based on the variance explained ratio explained by each main component. By using 8 components, data analysis becomes more effective and efficient in understanding patterns and relationships in the tuberculosis dataset.

Implementation of the KNN Classification Algorithm

The initial step is to apply the train-test split method to separate the dataset into two parts, namely training data and test data. In this study, the tuberculosis data set was divided in a 70:30 ratio, where 70% of the data was used as training data and 30% as test data. First, the k value is set to determine the number of neighbors that will be compared with the test data to classify the data as a result label in the tuberculosis data set. The k-value used in this research is 10. After that, to calculate the similarity between the test data and the training data, the Euclidean distance calculation method is used. Here, modelling is carried out using only the KNN algorithm, first using the attributes of the original dataset. The results of the model prediction can be seen in the confusion matrix table in Table 14.

Table 14. Confusion matrix table of KNN model prediction results

	Predicted		
Actual	Positive	Negative	Amount
Positive	76	85	161
Negative	11	2	13
Amount	87	87	174

$$\text{Accuracy} = \frac{\text{TP+TN}}{\text{TP+FN+FP+TN}} \times 100\% \quad (1)$$

$$\text{Accuracy} = \frac{76+85}{76+2+11+85} \times 100\%$$

$$\text{Accuracy} = \frac{161}{174} \times 100\%$$

$$\text{Accuracy} = 92{,}528\%$$

The next process carried out in this research is classification using the KNN algorithm using attributes/features that have been selected based on PCA. At this stage, several experiments were carried out using a different number of attributes based on the previous stage when selecting features using the PCA method. This experiment was carried out based on the values of n or main component that were determined during feature selection using PCA, which ranged from $n = 1$ to $n = 12$. The accuracy results of each experiment in this research can be seen in Table 15.

Table 15. Sample accuracy results of KNN experiments with PCA

No	*n*	*k*	Accuracy	No	*n*	*k*	Accuracy
1.	1	1	0.977011	5.	1	8	0.954023
2.	1	2	0.982759	6.	1	12	0.948276
3.	1	3	0.965517	7.	1	14	0.936782
4.	1	4	0.965517	8.	8	10	0.988506

In this research, it was found that a good classification model accuracy value was obtained using the KNN algorithm with PCA with a *n* value limit of 8 with a value of k = 10 which produced an accuracy value of 98.8506%. Based on the results obtained, it shows that applying PCA feature selection in the classification process using the KNN algorithm on the tuberculosis dataset can work as expected.

CONCLUSION

From the results of the research that has been carried out, it is known that the PCA feature selection technique can improve the performance of the KNN model for the diagnosis of TB disease. The accuracy results from the first classification process used the KNN algorithm with all the attributes from the dataset, namely an accuracy value of 92.528%, while in the second classification process, the KNN algorithm was used using selected attributes based on the *n* value. Several experiments were carried out with the best accuracy value obtained by the classification model that uses the value $n = 8$, the accuracy value is 98.85%. Based on the experiments that have been carried out, the application of PCA feature selection in the classification process using the KNN algorithm can increase the accuracy value by 6.33%. These results show that the performance of the KNN model with PCA is able to provide accurate diagnoses, so that it can be used to help diagnose TB disease for early prevention.

REFERENCES

[1] A. Nurjannah, F. Y. Rahmalia, H. R. Paramesti, and L. A. Laily, "Determinan Sosial Tuberculosis di Indonesia," vol. 3, no. 1, pp. 65–76, 2022.

[2] S. Maleki, B. Seyed, and H. Khasteh, "A survey on data mining techniques used in medicine," *J. Diabetes Metab. Disord.*, pp. 2055–2071, 2021, doi: 10.1007/s40200-021-00884-2.

[3] S. Maggo, A. Gupta, S. Jamwal, P. Setia, and S. Rathee, "Prediction of Tuberculosis disease using Data Mining Algorithms", doi: 10.4108/eai.27-2-2020.2303126.

[4] M. Gera, "Data Mining - Techniques , Methods and Algorithms : A Review on Tools and their Validity," vol. 113, no. 18, pp. 22–29, 2015.

[5] W. Xing and Y. Bei, "Medical Health Big Data Classification Based on KNN Classification Algorithm," vol. 8, 2020.

[6] N. Mariana, "Penerapan Algoritma k-NN (Nearest Neighbor) Untuk Deteksi Penyakit (Kanker Serviks)," 2016.

[7] S. Zhang and S. Member, "Challenges in KNN Classification," *IEEE Trans. Knowl. Data Eng.*, vol. 34, no. 10, pp. 4663–4675, 2022, doi: 10.1109/TKDE.2021.3049250.

[8] H. N. Zalloum, S. Al Zeer, A. Manassra, M. R. A. Sara, and J. H. Alkhateeb, "Breast Cancer Grading using Machine Learning Approach Algorithms," *J. Comput. Sci.*, vol. 18, no. 12, pp. 1213–1218, 2022, doi: 10.3844/JCSSP.2022.1213.1218.

[9] F. M. Refaat, M. M. Gouda, and M. Omar, "Detection and Classification of Brain Tumor Using Machine Learning Algorithms," *Biomed. Pharmacol. J.*, vol. 15, no. 4, pp. 2381–2397, 2022, doi: 10.13005/bpj/2576.

[10] R. Koenker, V. Chernozhukov, X. He, and L. Peng, *Handbook of quantile regression*. 2017. doi: 10.1201/9781315120256.

[11] J. Lever, M. Krzywinski, and N. Altman, "Principal component analysis," *Nat. Publ. Gr.*, vol. 14, no. 7, pp. 641–642, 2017, doi: 10.1038/nmeth.4346.

[12] S. S. Bali, "Reduksi Fitur Untuk Optimalisasi Klasifikasi Tumor Payudara Berdasarkan Data Citra FNA," pp. 73–78, 2017.

[13] M. F. Rahman, D. Alamsah, M. I. Darmawidjadja, and I. Nurma, "Klasifikasi Untuk Diagnosa Diabetes Menggunakan Metode Bayesian Regularization Neural Network (RBNN)," *J. Inform.*, vol. 11, no. 1, p. 36, 2017, doi: 10.26555/jifo.v11i1.a5452.

[14] A. Alamsyah and T. Fadila, "Increased accuracy of prediction hepatitis disease using the application of principal component analysis on a support vector machine," *J. Phys. Conf. Ser.*, vol. 1968, no. 1, 2021, doi: 10.1088/1742-6596/1968/1/012016.

[15] Q. H. Nguyen *et al.*, "Influence of Data Splitting on Performance of Machine Learning Models in Prediction of Shear Strength of Soil," vol. 2021, 2021.

[16] I. Gupta, V. Sharma, S. Kaur, and A. K. Singh, "PCA-RF: An Efficient Parkinson's Disease Prediction Model based on Random Forest Classification," 2022, [Online]. Available: http://arxiv.org/abs/2203.11287

[17] D. Granato, J. S. Santos, G. B. Escher, B. L. Ferreira, and R. M. Maggio, "Use of principal component analysis (PCA) and hierarchical cluster analysis (HCA) for multivariate association

between bioactive compounds and functional properties in foods: A critical perspective," *Trends Food Sci. Technol.*, vol. 72, no. October 2017, pp. 83–90, 2018, doi: 10.1016/j.tifs.2017.12.006.
[18] L. Ibiapina *et al.*, "Application of Data Mining Algorithms for Dementia in People with HIV / AIDS," vol. 2021, 2021.
[19] D. R. Fakhma, "Diagnosis of TBC Disease Using SVM and Feedforward Backpopagation," vol. 4, no. April, pp. 77–86, 2022.
[20] M. T. Khan, A. C. Kaushik, L. Ji, S. I. Malik, S. Ali, and D. Q. Wei, "Artificial neural networks for prediction of tuberculosis disease," *Front. Microbiol.*, vol. 10, no. MAR, pp. 1–9, 2019, doi: 10.3389/fmicb.2019.00395.
[21] H. Saiyar, "Aplikasi Diagnosa Penyakit Tuberculosis Menggunakan Algoritma Naive Bayes," *Jurikom)*, vol. 5, no. 5, pp. 498–502, 2018, [Online]. Available: http://ejurnal.stmik-budidarma.ac.id/index.php/jurikom%7CPage%7C498
[22] R. Vanitha and K. Perumal, "Hybrid Model of KNN and PCA to Pre-processor of Thyroid Dataset using Machine Learning," *J. Pharm. Negat. Results*, vol. 13, no. SO3, pp. 1091–1096, 2022, doi: 10.47750/pnr.2022.13.s03.170.
[23] M. A. Ullah, S. H. Afrin, K. M. Nazib, R. Roy, and L. E. Ali, "Unravelling Parkinson's Disease Prediction: An Evaluation of Feature Selection Techniques with a Focus on PCA and KNN Performance," *Rev. Comput. Eng. Stud.*, vol. 10, no. 2, pp. 20–27, 2023, doi: 10.18280/rces.100201.
[24] J. Svensson and O. P. Guillen, "WHAT IS DATA AND WHAT CAN IT BE USED FOR ? KEY QUESTIONS IN THE AGE OF BURGEONING DATA-ESSENTIALISM," 2003.
[25] I. T. Jolliffe, J. Cadima, and J. Cadima, "Principal component analysis : a review and recent developments Subject Areas : Author for correspondence :," 2016.
[26] N. Lounes, H. Oudghiri, R. Chalal, and W.-K. Hidouci, "From KDD to KUBD: Big Data Characteristics Within the KDD Process Steps," in *Trends and Advances in Information Systems and Technologies*, Á. Rocha, H. Adeli, L. P. Reis, and S. Costanzo, Eds., Cham: Springer International Publishing, 2018, pp. 931–937.
[27] Walid, A., and Alamsyah, I. U., "Recurrent neural network for forecasting time series with long memory pattern," in *J. Phys.: Conf. Ser.*, vol. 824, no. 1, p. 012038, 2017.