Journal of
Information Systems Engineering
and Business Intelligence

Vol.5, No.1, April 2019
Available online at: http://e-journal.unair.ac.id/index.php/JISEBI

Decision Support System for Classification of Early Childhood Diseases Using Principal Component Analysis and K-Nearest Neighbors Classifier

Damar Dananjaya[1)], Indah Werdiningsih[2)*], Rini Semiati[3)]

[1,2,3)]Universitas Airlangga, Indonesia
Kampus C Mulyorejo, Surabaya
[1)]damar.dananjaya-13@fst.unair.ac.id, [2)]indah-w@fst.unair.ac.id, [3)]rini-s@fst.unair.ac.id

Article history:

Received 7 September 2018
Revised 29 December 2018
Accepted 5 January 2019
Available online 28 April 2019

Keywords:

Classification
Disease
Early childhood
K-NN
PCA

Abstract

Background: Data on early childhood disease collected in clinics has accumulated into big data. Those data can be used for classification of early childhood diseases to help medical staff in diagnosing diseases that attack early childhoods.
Objective: This study aims to apply Principal Component Analysis (PCA) and K-Nearest Neighbor (K-NN) Classifier for the classification of early childhood diseases.
Methods: Data analysis was performed using PCA to obtain variables that had a major influence on the classification of early childhood diseases. PCA was done by observing the correlation between variables and eliminating variables that have little influence on classification. Furthermore, data on early childhood disease was classified using the K-Nearest Neighbor Classifier method.
Results: The results of system evaluation using 150 test data indicated that the classification system by applying PCA and KNN Classifier had an accuracy value of 86%.
Conclusion: PCA can be used to reduce the number of variables involved so that it can improve system performance in terms of efficiency. In addition, the application of PCA and KNN can also improve accuracy in the classification of early childhood diseases.

I. INTRODUCTION

Early childhood are assets of the nation that need good education and care for improving the quality of the nation [1]. Early childhood mortality rate each year is about 12.4 million [2]. Improving children's health is one of the most important programs of the government.

During the growth period, children are susceptible to various diseases. The diseases are mostly caused by germs or viruses that experience direct contact with children. The symptoms that often accompany early childhood diseases are fever, cough, and diarrhea [3][4][5][6]. Proper handling is needed for children to be healthy again without any side effects.

Children health clinic provides services and treatments for early childhood. Every patient who comes is recorded in the medical record. The medical record data is collected in clinical storage, so it becomes a big data. This causes the medical staff difficult in processing the patient data stack. The medical staff diagnose the disease and give treatment based on Manajemen Terpadu Balita Sakit (MTBS).

MTBS is a comprehensive program that deal with sick children come for basic health care service [7]. MTBS is an embodiment of the international program of WHO and UNICEF to integrate services for early childhood diseases, initially called MTBS. MTBS classifies several classes of early childhood disease based on complaints of cough, diarrhea and fever. Each class has a list of symptoms suffered by children.

Symptoms suffered by children can be processed, so that medical staff can determine the type of disease and type of treatment [8] [9]. In addition, there are other factors that also affect the disease that children suffer. Medical record data generally includes data on body weight, height, age, body temperature and so on. Collection of data in medical records can be used to extract information [8], so that the types of diseases suffered by patients can be known [10].

Many studies have been conducted on the diagnosis of early childhood disease, one of which is [9] who states that forward chaining can be used to diagnose early childhood disease. However, this study only uses symptom

* Corresponding author

ISSN 2443-2555 (online) 2598-6333 (print)
doi: http://dx.doi.org/10.20473/jisebi.5.1.13-22

factors and ignores other factors, such as height, weight, age, sex, and body temperature. These factors can be used to diagnose disease [11].

The patient's medical record consists of various features. Features reduction is very important to identify the most significant risk factors associated with disease [12]. Feature reduction is an efficient data pre-processing technique in data mining to reduce data dimensions [13] [14] [15].

PCA is one of the well-known statistical techniques that aims to reduce the dimensions of data without losing important information in data [16] [17]. PCA basically converts and decomposes a large number of correlated variables into a small number of uncorrelated variables and can reduce the dimensions of data [18]. PCA has several advantages such as reducing data redundancy, reducing complexity, reducing database size, and reducing noise. PCA can be used to determine correlations between variables [18].

Classification is a technique used to find unknown data classes [19]. Various methods are known for classification, such as decision trees, rule based, K-NN classifier, and others [20]. K-NN classifier is a method used to classify new objects based on their nearest neighbors. K-NN became well known among the data mining techniques because of its simplicity and relatively high speed of convergence. K-NN is also called lazy learning because it did not go through training phase and memory based classification because the training sample must be in memory while the process was running [11].

Jabbar [11] implemented K-NN with Feature Subset Selection to determine variables that contribute more to disease prediction. This method can indirectly reduce the number of tests that must be taken by patient. This prediction model can help doctors in an efficient decision-making process with fewer variables to diagnose heart disease.

In those studies, the accuracy of the diagnosis depends on the features used in the diagnosis of early childhood disease. Therefore, it is very important to develop a systematic scheme that is able to determine the most representative features to maximize the accuracy of diagnoses in early childhood diseases. In this paper we investigate the application of K-NN with feature reduction to classify early childhood diseases.

II. METHODS

A. Data collection

The data collection methods used in this study were direct interviews with sources and literature studies. It aims to gain knowledge about the assessment and classification of early childhood diseases. Interviews were conducted on three doctors and two nurses. Patients were all children diagnosed with the disease at the children's health clinic where data was collected. Data was obtained from patient data at one of the children's health clinics in Surabaya for 3 years, from 2015 to 2017.

B. Data analysis

The data obtained was divided into two parts, i.e. training data (70%) and testing data (30%). PCA analysis was used to know the correlation between variables data. K-NN classifier was used to determine the classification of early childhood diseases. Based on the reference of MTBS, the early childhood diseases were classified into 16 diagnoses and 26 symptoms. Data analysis consisted of two analysis, i.e. Principal Component Analysis (PCA), and K-Nearest Neighbor (K-NN) Classifier.

1) Principal Component Analysis (PCA)

PCA analysis was used to discover the correlation between data variables. Data variables were age, weight, height, body temperature, sex, and 26 symptoms. The symptoms were cough, diarrhea, fever, inability to drink or suckle, vomiting, unconsciousness, fast breathing, breathing difficulty, Stridor, liquid or soft defecating, hollowed eyes, poor abdominal skin turgor, fussiness/ irritability, abnormal thirst, nausea, diarrhea of 14 days or more, blood in feces, stiff neck, rash, red eyes, turbidity on the cornea, mouth ulcer, festering eyes, fever for 2 to 7 days, high and continuous sudden fever. 16 diagnosis were cough, pneumonia, severe pneumonia, diarrhea, mild dehydration diarrhea, severe dehydration diarrhea, persistent diarrhea, severe diarrhea, dysentery, common fever, severe fever, measles, measles with severe complication, measles with complication, fever may be Dengue Hemorrhagic Fever (DHF), DHF, and fever isn't DHF.

By looking at the correlation between these variables would be obtained the factors that influence the early childhood diseases. The PCA steps were used as follows:

1. Calculation of the variance

Calculation of the variance used (1).

$$var(X) = \frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{(n-1)} \quad (1)$$

2. Calculation of the covariance
Calculation of covariance used (2). After that, the covariance matrix was generated.

$$cov(X,Y) = \frac{\sum_{i=1}^{n}(X_i-\bar{X})(Y_i-\bar{Y})}{(n-1)} \quad (2)$$

3. Calculation of the eigenvalues and eigenvectors
Calculation of the eigenvalues and eigenvectors used (3). After the eigenvalues of the covariance matrix was known, the eigenvectors of each eigenvalue was calculated by using (4).

$$(A-\lambda I)X = 0 \quad (3)$$

$$det\ (A-\lambda I) = 0 \quad (4)$$

4. Calculation of the principal component.
After the principal component was known, the correlation between the principal variables and components will be calculated using (5). Reducing variables by eliminating the low components.

$$r_{y_{ik},X_i} = \frac{\hat{e}_{ik}\sqrt{\lambda_k}}{s_{ii}} \quad (5)$$

Notation:
A = Matrix n x n
λ = Eigenvalue
I = Identity Matrix
r_{y_{ik},x_j} = Correlation between sample X_i and principal component y_{ik}
$\hat{e}_{ik}$ = Eigenvektor
s_{ii} = Covariance matrix

2) K-Nearest Neighbor (K-NN) Classifier

K-NN classifier was used to determine the classification of early childhood diseases. Variables used for classification were variables obtained from PCA. KNN classifier used weighted voting to calculate the weight of each class. Steps of KNN classifier were as follows:
1. Variables obtained from PCA were variables that influence classification.
2. The dataset was normalized using the min-max normalization in (6).

$$X' = \frac{X-\min(X)}{\max(X)-\min(X)} \quad (6)$$

3. Calculating the distance by using Euclidean distance in (7).

$$d_{euclidean}\,(X,Y) = \sqrt{\sum_{i=1}^{n}(X_i-Y_i)^2} \quad (7)$$

4. Generating the class of testing data.
The weighted voting function was used to calculate the weight of each class. The class with the greatest value was included in its class. Weighted voting used (8).

$$Weighted\ voting = \sum_{i=1}^{n}(\frac{1}{d_j^2}) \quad (8)$$

C. System Design

The system design displayed the flow of the application of Principal Component Analysis and k-Nearest Neighbor for classification of early childhood diseases in the form of flowchart diagrams. Flowcharts were made for two systems, i.e. the early childhood diseases classification system by using PCA and KNN Classifier, and the early childhood diseases classification system by using KNN Classifier.

D. System Implementation

At this stage, the system would be built based on system design and implemented into a web-based system using PHP programming language. MySQL Database was used as a storage for early childhood diseases data.

E. System Testing

System testing was required to find out whether the system is working properly and correctly. System testing was done by black box testing techniques using data obtained from the Children Clinic.

F. System Evaluation

System evaluation compared the output of system by applying PCA and K-NN classifier, with K-NN classifier without PCA. Each system output was compared against the original data to obtain their respective

accuracy. After obtaining the accuracy of each output, it would be compared to know which one was better between classification accuracy using PCA and K-NN classifier, or using K-NN classifier.

III. RESULTS

A. Data and Information Collection

Data collection techniques used in this study were interviews and literature studies. The results of the interview were information about how to classify early childhood diseases based on symptoms and complaints suffered by patients, and how to obtain patient data. The PCA and KNN classification literature studies were obtained from books in libraries, e-books, and scientific journals. A literature study was also conducted to find out more about the MTBS which can be found in the MTBS modules and related scientific journals.

B. Data Analysis

The data used were 500 data. The data was divided into two parts, namely 350 training data (70%) and 150 testing data (30%). Training data was used to determine the variables that influence the classification process using PCA. After getting the influencing variable, the classification process was carried out. K-NN was used for the classification process using testing data.

1) Principal Component Analysis (PCA)

PCA analysis was used to determine the correlation between data variables. Data analyzed using PCA was only training data, as a system knowledge base. The steps in PCA were as follows:

1. Determination of the variables to be analyzed.

Variables used were weight, height, sex, age, body temperature, and 26 Symptoms. Variables name code were used to ease the programming process, i.e. age = X_1, weight = X_2, height = X_3, body temperature = X_4, sex = X_5, cough = X_6, diarrhea=X_7, fever=X_8, inability to drink or suckle =X_9, vomiting=X_{10}, seizures=X_{11}, unconsciousness=X_{12}, fast breathing=X_{13}, breathing difficulty=X_{14}, Stridor=X_{15}, liquid or soft defecating=X_{16}, hollowed eyes=X_{17}, poor abdominal skin turgor =X_{18}, fussiness/ irritability =X_{19}, abnormal thirst=X_{20}, nausea=X_{21}, diarrhea of 14 days or more=X_{22}, blood in feces=X_{23}, stiff neck =X_{24}, rash=X_{25}, red eyes=X_{26}, turbidity on the cornea=X_{27}, mouth ulcer =X_{28}, festering eyes =X_{29}, fever for 2 to 7 days =X_{30}, high and continuous sudden fever=X_{31}.

2. Creation of covariance matrix.

Covariance matrix could be created by calculating the covariance values between data variables. For each variable, value of the relationship between the variables themselves as well as other variables were calculated according to Eq. 2. For example, cov (X_1, X_1) was the covariance of variable X_1 against X_2, cov (X_1, X_2) was the covariance of variable X_1 against X_2. The same formula used for X_2, X_3, X_4……. X_{31}. Example of covariance calculation for cov (X_1, X_2):

$$Cov(X_1, X_2) = \frac{(20 - 23.41)(10 - 10.89) + \cdots + (12 - 23.41)(9 - 10.89)}{(500 - 1)} = 42.334$$

3. Calculation of eigenvalues and eigenvectors

After the covariance matrix was formed, the eigenvalue and vector was calculated. The eigenvalue and eigenvector could be completed using the website: www.comnuan.com. Calculation of eigenvalues and eigenvectors for X1, X2, X3, and X4 could be seen in Table 1.

TABLE 1
EIGENVALUES AND EIGENVECTORS

Eigenvalues	515.587	38.747	2.6145	0.6913
Eigenvectors	-0.6582	-0.7415	-0.1295	0.0146
	-0.12	-0.0675	0.9855	-0.0995
	-0.7432	0.6675	-0.0454	-0.0065
	-0.0072	0.0085	0.1002	0.9949

4. Determination of Principal Components

The eigenvector above was used to determine the Principal Components. From these calculations could be made four Principal Components formed by multiplying the eigenvector with X variable.

$$Y1 = (-0.6582).X1 - 0.12.X2 - 0.7432.X3 - 0.0072.X4$$
$$Y2 = (-0.7415).X1 - 0.0675.X2 + 0.6675.X3 + 0.0085.X4$$
$$Y3 = (-0.1295).X1 + 0.9855.X2 - 0.0454.X3 + 0.1002.X4$$
$$Y4 = (0.0146).X1 - 0.0995.X2 - 0.0065.X3 + 0.9949.X4$$

5. Find the correlation between variables and Principal Components.

Each Principal Component formed was correlated with all variables in accordance with (8). The following was an example of correlation calculation for Y_1 and X_1.

$$r_{y_{ik},x_i} = \frac{(-0.6582)\ x\ \sqrt{515.5873}}{244.714} = -0.06107315$$

The solution of the above equation was converted into a matrix form, so that a correlation matrix was generated. By observing the matrix, variables that were less influential for the classification process were known. The example of the correlation matrix could be seen in Table 2 below.

TABLE 2
CORRELATION MATRIX

	X1	X2	X3	X4
Y1	-0.06107	0.023	0.391	0.380
Y2	-0.45480	0.291	0.703	-0.209
Y3	-0.00069	0.158	0.619	0.078
Y4	0.01639	0.535	-0.045	0.674

Based on the matrix above, the variables chosen were whose value is above 0.5. The age variable (X_1) was not used for classification variables. Variables used for classification were weight (X_2), height (X_3), and body temperature (X_4).

Based on the results of experiments with the same calculation above using training data, then the variables obtained were 18 variables, i.e. weight, sex, cough, flu, diarrhea, fever, vomiting, inability to drink or suckle, seizures, breathing difficulty, unconsciousness, stridor, blood in the feces, hollowed eyes, poor abdominal skin turgor, fussiness / irritability, diarrhea 14 days or more, turbidity on the cornea. This variable would be used for the classification process.

2) K-Nearest Neighbor (K-NN) Classifier

The way the KNN classifier works was to calculate the distance between data testing against training data. The steps of the KNN Classifier were as follows:

1. Data Normalization

Data normalization was carried out on training data and testing data. The Sex variable has two values, i.e. "Male" and "Female", normalized to "1" and "0". "1" for male and "0" for female. For Symptom variables, the values "Yes" and "None" are normalized to be "1" and "0". The value "1" for normalizing the value "Yes" and "0" for the value "None". The equation min-max normalization was used for the variable weight, height, and body temperature. The equation used the highest value and the smallest value of training data for each variable.

Example for calculation of min-max normalization data of patient "Y" who had 20 months of age, weight = 10 kg, height = 79 cm and body temperature 37.5 ° C. The Maximum and Minimum values of training data could be seen in Table 3.

TABLE 3
MAXIMUM AND MINIMUM VALUES

	Weight	Height	Body Temperature
Max	21	120	40
Min	4.4	21	36

$$X_2{}' = \frac{X_2 - \min(X_2)}{\max(X_2) - \min(X_2)} = \frac{10 - 4.4}{21 - 4.4} = 0.33735$$

$$X_3{}' = \frac{X_3 - \min(X_3)}{\max(X_3) - \min(X_3)} = \frac{79 - 21}{120 - 21} = 0.58586$$

$$X_4{}' = \frac{X_4 - \min(X_4)}{\max(X_4) - \min(X_4)} = \frac{37.5 - 36}{40 - 36} = 0.375$$

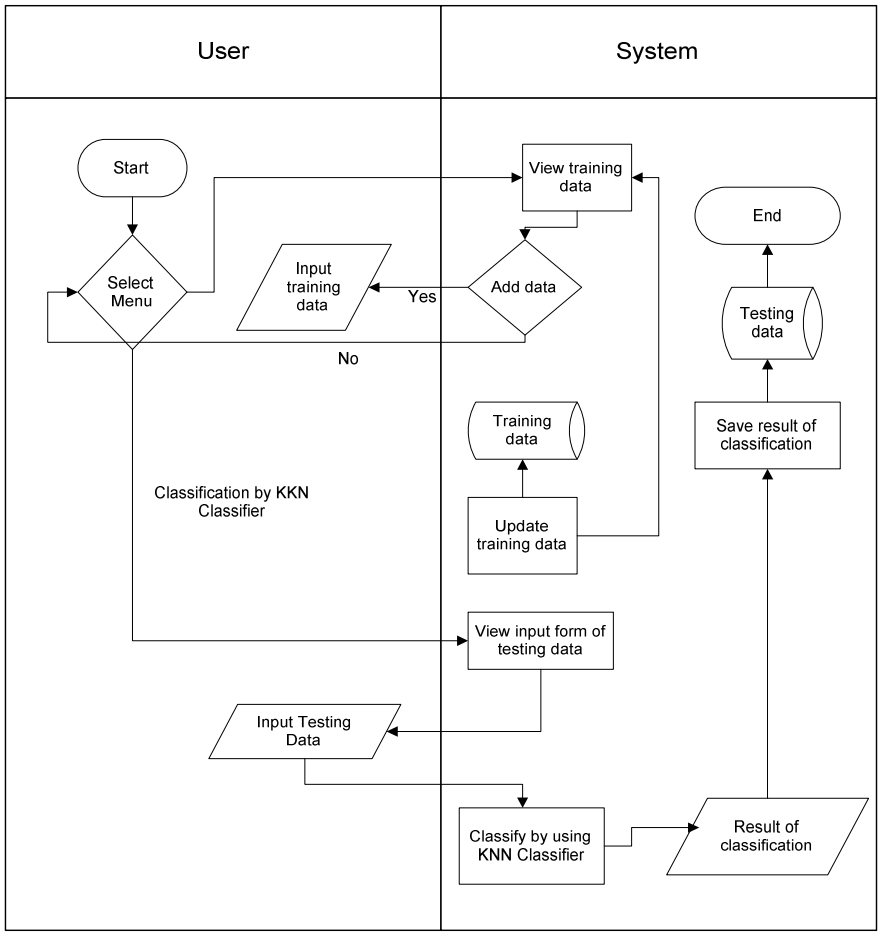

Fig. 1 System flow diagram for classification using K-NN

After normalizing the min-max normalization, the "Y" patient had a value of weight = 0.33735, height = 0.58586 and body temperature = 0.375.

2. Calculation of Euclidean distance

Similarity level of data to a class was determined by its distance. The smaller the distance, the greater the resemblance to a class. Euclidean Distance equation was used to measure the similarity distance. The distance for each testing data to training data was calculated.

Example of calculating Euclidean distance for "Z" patient and "Y" patient by applying (7).

$$d(Z,Y) = \sqrt{(1-0)^2 + (0.2048-0.3373)^2 + \cdots + (0-0)^2} = 1.76339$$

The distance values for "Z" and "Y" was 1.76339. The same formula was used to calculate the distance of "Z" patient to another patient.

3. Weighted Voting

Errors in predicting classes could occur even though the proximity distance had been calculated. Those was because there was noise or data that deviates far from the original class.

Weighted Voting was needed to determine the class location from new data. Voting was done by weighting each class. By calculating the total value of Euclidean Distance for each class, the final result of the classification process can be determined.

The output of this study were 16 classes of early childhood diseases, i.e. cough, pneumonia, severe pneumonia, diarrhea, mild dehydration diarrhea, severe dehydration diarrhea, persistent diarrhea, severe persistent diarrhea, dysentery, fever, fever with common risk sign, measles, measles with severe complication, measles with complication, fever may be DHF, DHF, fever isn't DHF. The calculation below was an example of weighted voting for the "Cough" class. There are 45 data that had "cough" diagnosis from 400 training data. Equation (8) was applied to get the "Cough" class score from the "Z" patient.

$$voting\ (Cough) = \left(\frac{1}{1.5413^2}\right) + \cdots + \left(\frac{1}{1.05003^2}\right) = 146.8976$$

Other early childhood diseases classes were also applied the same calculation with different values. The results of weighted voting for all classes of early childhood diseases were shown in Table 4. The "Cough" class had a score of "146.8976105", which was the highest score compared to other classes. So it could be concluded that "Z" patient suffer "Cough".

TABLE 4
WEIGHTED VOTING SCORE

Early Childhood Diseases	Score
Cough	146.8976
Pneumonia	5.364010
Severe Pneumonia	3.265106
Diarrhea	3.651260
Mild Dehydration Diarrhea	1.000566
Severe Dehydration Diarrhea	0.848560
Persistent diarrhea	0.121155
Severe diarrhea	0.220797
Dysentery	0.461117
Common fever	5.183737
Severe fever	4.284291
Measles	5.061522
Measles with severe complication	0.420538
Measles with complication	4.064063
fever may be DHF	1.383750
DHF	0.661613
Fever isn't DHF	0.317643

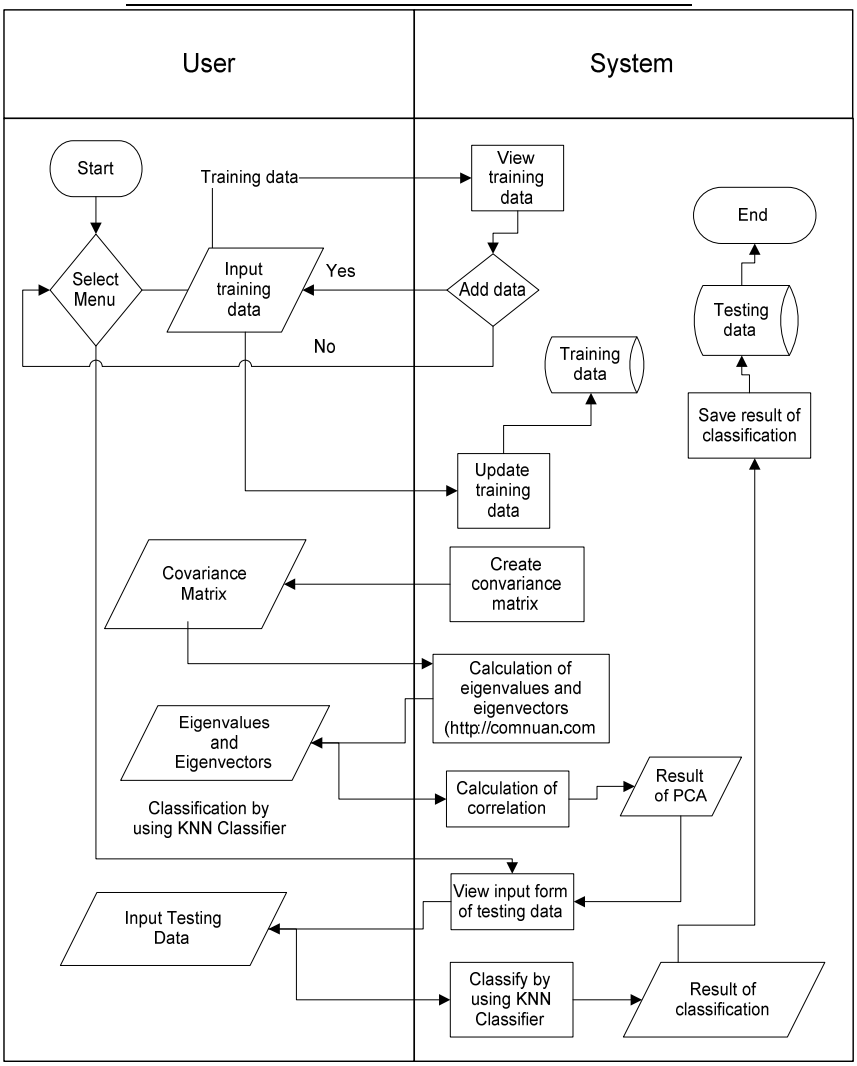

Fig. 2 System flow diagram for classification using PCA and K-NN

C. System Design

The system design of the PCA and K-NN Classifier for the classification of early childhood diseases was described using System Flow Diagram. Two scenarios are applied to prove that PCA can reduce features and improve the accuracy of classification. The first scenario, the user can use two features of the system, which is entering new data and viewing the results of testing. The new data entered will be training data. Users can see the results of testing by entering testing data. The system will process the testing data and display the results of the classification. The System Flow Diagram for the system classification for early childhood disease using K-NN can be seen in Fig.1. The second scenario is the same as the first, but the testing data is managed using feature reduction. The System Flow Diagram for the system classification for early childhood by using K-NN and PCA can be seen in Fig. 2. The form that displays the classification results can be seen in Fig. 3.

D. System Implementation

System implementation was a web-based application. This system was built using Hypertext Preprocessor (PHP) programming language, Hypertext Markup Language (HTML), JavaScript, and MySQL as a database management system. The tools used to build the system were Adobe Dreamweaver, Bracket, and XAMPP.

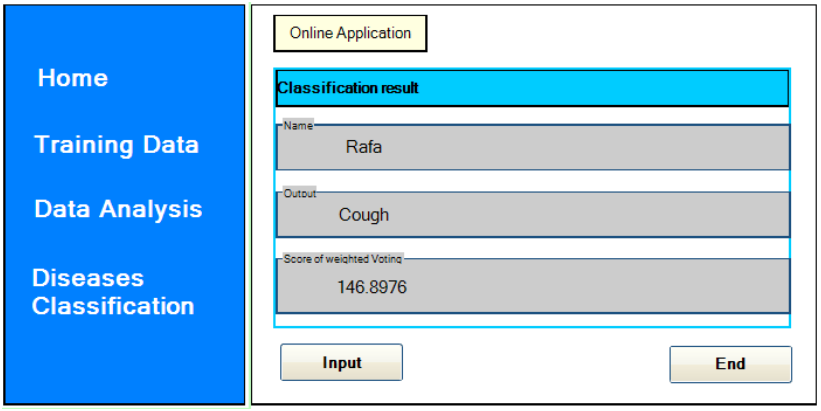

Fig. 3 The form that displays the classification results

E. System Testing

Black Box Testing was done for system testing to test how well the performance of the application made. Black Box testing was done by comparing the expected results with the results issued by the system. Testing was carried out by applying PCA and KNN Classifier for classification of early childhood diseases. The form of KNN classifier result could be seen in Fig. 3.

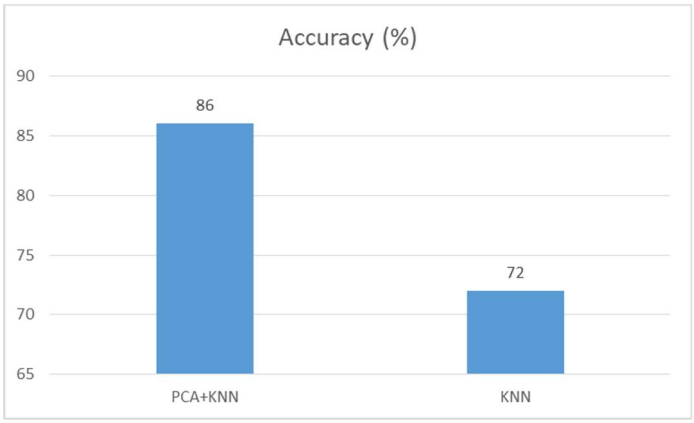

Fig. 4 The result of system evaluation

F. System Evaluation

System evaluation was carried out on the two systems built, i.e. classification system using PCA and K-NN classifier and classification system using K-NN classifier. 500 data obtained were processed for both systems, 350 data were used as training data and 150 data were used as testing data. Based on the results of system testing, 14 data were labelled as true for classification using PCA and KNN, while 18 data were labelled as true for classification using KNN without PCA. Result accuracy was calculated by dividing the number of correct data to all data, then multiplied it by 100%. The result was correct if the expected result and the result issued by

system had the same result. Result Accuracies obtained were 86 % for early childhood classification using PCA and K-NN classifier and 72% for using K-NN Classifier without PCA. The result of system evaluation could be seen in Fig. 4.

IV. DISCUSSION

Based on these results, reducing variables using PCA could improve classification results because the PCA eliminated variables that had low principal component values [17]. In the classification process, only variables that had a large principal component values were used, so that the variables used were only the most influential on the classification process.

Variables used for classification were 18 variables, i.e. weight, sex, cough, flu, diarrhea, fever, vomiting, inability to drink or suckle, seizures, breathing difficulty, unconsciousness, stridor, blood in the feces, hollowed eyes, poor abdominal skin turgor, fussiness / irritability, diarrhea 14 days or more, turbidity on the cornea. Those symptoms variables were symptoms often experienced by early childhoods having fever, cough, and diarrhea [3][4][5][6].

The accuracy of the results obtained in this research was unsatisfactory because KNN did not conduct base learning, but directly classified the existing database, and could not classify data outside the database [21]. Rule based system is a system able to do base learning inside the database.

The future work in this study is rule-based classification. By using a new method, hopefully it can improve the accuracy gained. If the system has better accuracy, it can be used to help medical staff to be able to diagnose early childhood disease earlier

V. CONCLUSIONS

After applying the Principal Component Analysis and K-Nearest Neighbor Classifier for the early childhood classification, it can be concluded that the variables that influence the classification of early childhood diseases by using PCA were 18 variables, i.e. weight, sex, cough, flu, diarrhea, fever, vomiting, inability to drink or suckle, seizures, breathing difficulty, unconsciousness, stridor, blood in the feces, hollowed eyes, poor abdominal skin turgor, fussiness / irritability, diarrhea 14 days or more, turbidity on the cornea. The system evaluation result showed that the accuracy for the classification system using PCA and KNN Classifier was 86% and KNN Classifier without PCA was 72%. From the two systems built, it could be concluded that implementing PCA could be used to reduce variables. PCA and KNN Classifier had greater accuracy in the classification of early childhood diseases.

REFERENCES

[1] A. Fauzi, "Penerapan Forward Chaining dalam Sistem Pakar Untuk Mendiagnosis Penyakit Pada Anak," *Techno Xplore,* vol. 1, no. 1, pp. 11-16, 2016.

[2] W. Bank, "World Development Report: Investing in Health," Oxford Univ. Press, Oxford, 2012.

[3] M. Garenne, C. Ronsmans and H. Campbel, "The Magnitude of Mortality From Acute Respiratory Infections in Children Under 5 Years in Developing Countrie," *World Health Stat Q,* vol. 45, no. 2-3, p. 180–191, 1992.

[4] J. Snyder and M. Merson, "The magnitude of the global problem of acute diarrhoeal disease: A review of active surveillance data," *Bull World Health Organ,* vol. 60, no. 4, pp. 605-13, 1982.

[5] B. C, M. J, d. Z. I and G. RI, "The magnitude of the global problem of diarrhoeal disease: a ten-year update," *Bull World Health Organ,* vol. 70, no. 6, pp. 705-714, 1992.

[6] O. WA, M. LE, A. WL and H. AR, "Worldwide Measles Prevention," *Israel Joumal of Medical Science,* vol. 30, no. 5-6, pp. 469-481, 1994.

[7] Kementrian Kesehatan Republik Indonesia, "Manajemen Terpadu Balita Sakit (MTBS)," Jakarta, 2012.

[8] P. K. Patra, D. P. Sahu and n. Mandal, "An Expert System for Diagnosis of Human Diseases," *International Journal of Computer Applications,* vol. 1, no. 13, pp. 71-73, 2010.

[9] B. F. Yanto, I. Werdiningsih and E. Purwanti, "Perancangan Sistem Pakar Diagnosa Penyakit Pada Anak Bawah Lima Tahun Menggunakan Metode Forward Chaining," *Journal of Information Systems Engineering and Business Intelligence,* vol. 3, no. 1, pp. 61-67, 2017.

[10] E. Kurniawan, I. K. E. Purnama and S. Sumpeno, "Analisa Rekam Medis untuk Menentukan Pola Kelompok Penyakit Menggunakan Klasifikasi dengan Decision Tree J48," in *Pasca Sarjana Teknik Elektro ITS*, Surabaya, 2011.

[11] M. A. Jabbar, B. Deekshatulu and P. Chandra, "Heart Disease Classication Using Nearest Neighbor Classifier With Feature Subset Selection," *Anale. Seria Informatică,* vol. 11, no. 1, pp. 47-54, 2013.

[12] D. Jain and V. Singh, "Feature selection and classification systems for chronic disease prediction: A review," *Egyptian Informatics Journal,* vol. xxx, pp. xxx-xxx, 2018.

[13] I. E. Kaya, A. A. Pehlivanl, E. G. Zekiskarde and T. Ibrikci, "PCA Based Clustering For Brain Tumor Segmentation of T1w MRI Image," *Computer Methods and Programs in Biomedicine,* vol. 140 , no. C, pp. 19-28, 2017.

[14] M. Shardlow, "An Analysis of Feature Selection Techniques," in *The University of Mancheste*, Manchester, 2016.

[15] M. Dash and L. H, "Feature selection for classification," *Intelligent Data Analysis,* vol. 1, no. 1-4, pp. 131-156, 1997.

[16] M. G. Hendro, A. T. Bharata, S. and N. Akhmad, "Penggunaan Metodologi Analisa Komponen Utama (PCA) untuk Mereduksi FaktorFaktor yang Mempengaruhi Penyakit Jantung Koroner," in *Seminar Nasional Science, Engineering and Technology*, Yogyakarta, 2012.

[17] J. Han, M. Kamber and J. Pei, Data Mining Concepts and Techniques, Waltham: Morgan Kaufmann, 2012.

[18] J. Tang, S. Alelyani and H. Liu, "Feature Selection for Classification: A Review," in *Data Classification: Algorithms and Applications*, New York, CRC Press, 2014, p. 37.

[19] S. Beniwal and J. Arora, "Classification and Feature Selection Techniques in Data Mining," *International Journal of Engineering Research & Technology,* vol. 1, no. 6, pp. 1-6, 2012.

[20] D. T. Larose, Discovering Knowledge in Data, Second Edition, New Jersey: John Wiley & Sons, 2014.

[21] Ahn, H., & Kim, K.-j, Bankruptcy prediction modeling with hybrid case-based reasoning and genetic algorithms approach, Elsevier, 599-607, (2009)