# Expert system for diagnosing diphtheria with k-nearest neighbor method

Chavid Syukri Fatoni [a,1,*], Emma Utami [a,2], Ferry Wahyu Wibowo [a,3]

[a] Teknik Informatika, Universitas AMIKOM Yogyakarta

[1] fatoni.work@gmail.com*; [2] emma@nrar.net; [3] ferry.w@amikom.ac.id

* corresponding author

ARTICLE INFO

ABSTRACT

Outbreak Diphtheria is now a special concern for the Indonesian government and is noted as an extraordinary case (KLB) in 2017. Diphtheria is an infectious disease and can cause complications of dangerous and deadly diseases if not addressed immediately, so far the community often underestimates often feels like a sore throat, flu, fever. Similarity of symptoms Diphtheria with common diseases and complications such as myocarditis, airway obstruction, Acute Kidney Injury (AKI) makes Diphtheria difficult to treat due to viral infections that spread so quickly. Some complications of diphtheria can cause death if not treated immediately and there must be identification at the beginning for diphtheria, an expert system is needed that can help the community and government in diagnosing diphtheria. This study uses the K-Nearest Neighbor (KNN) method in which this method calculates the similarity value of each Diphtheria symptom so that it can provide an initial diagnosis for Diphtheria before complications occur. The output of this study was in the form of diagnosis of diphtheria based on the symptoms experienced with an accuracy test of 93.056%, as well as providing an initial diagnosis so as not to be late in the treatment of Diphtheria

## 1. Introduction

Outbreak Diphtheria is now a special concern for the Indonesian government and recorded as a case of extraordinary events (KLB) in 2017 [1]. Diphtheria is an infectious disease and can cause complications of dangerous and deadly diseases if not treated immediately, so far people often underestimate diseases that are often felt such as sore throat, flu, fever. Similarity of symptoms Diphtheria with common diseases and complications such as Myocarditis, Airway Obstruction, Acute Kidney Injury (AKI) makes Diphtheria difficult to treat due to viral infections that spread so quickly. Some complications of Diphtheria that can cause death if not treated immediately and there must be identification at the beginning for Diphtheria, an expert system is needed that can help the community and government in diagnosing Diphtheria.

This study uses the K-Nearest Neighbor (KNN) method in which this method calculates the similarity value of each Diphtheria symptom so that it can provide an initial diagnosis for Diphtheria before complications occur. This study focuses on processing data mining with Diphtheria disease symptom data sets obtained from previous studies, the concept of Expert Systems can help in processing data on the knowledge base to get a diagnosis of a disease. Data that has been obtained then over time will increase and increase so as to improve accuracy in providing a diagnosis of

Diphtheria. In this study the authors limit to only using 148 data of Diphtheria disease from previous studies, and using the KNN method with a similarity value of 0.8 to calculate the distance similarity from one case to a specific diagnosis whether cases including Diphtheria cases are classified as Myocarditis, Airway Obstruction, Acute Kidney Injury (AKI). The test in this study uses Cross Validation with a random K value from 1 to 10 to produce what percentage of the accuracy of the method applied in this study.

The first research is about how K-Nearest Neighbor can identify efficiency in the selection of documents based on the categories taken in the document. It was concluded in the study that the model used to process and select similar documents was based on words, but the use of words to process document content still caused two problems, namely semantic confusion, words that implied that training documentation was inappropriate, shared the same words that we want to classify and then choose can increase noise, in contrast to cases of disease diagnosis that clearly the attributes used are the results of expert analysis and from medical record data [2]. The second study applied 60 medical record data on intestinal diseases taken from RSUD dr. Soetrasno Rembang with a scenario of 40 source cases and 20 target cases, it can be seen that the accuracy of the system diagnosis is 95%, but in this study there is no greater amount of data, and the attributes used are all related. symptoms that have not entered the data Characteristics, of 60 data classified into 5 underlying intestinal diseases, not included in complications of the disease [3]. The third study analyzed the limited number of mechanics and user knowledge which resulted in various difficulties and treatments in dealing with damage to the machine. Creating an artificial intelligence system with a case-based reasoning concept can facilitate the initial diagnosis of a case, but this study does not mention the amount of training data used in the study and accuracy is not stated [4].

The fourth study is about the automatic diagnosis of determination of non-critical diseases or in cases where doctors may not be available to diagnose in remote locations. By learning different data mining techniques, we found techniques suitable for making medical diagnoses using KNN and achieving accuracy above 84% [5]. In the study did not mention how much data used in the experiment other than that there was no information about the class of Diphtheria disease. The fifth study discusses the diagnosis of heart disease with accuracy reaching between 94% and 97.4%, respectively, with a K value between one and 13 [6]. This study does not determine how much data is used and only focuses on heart disease for the research case. The sixth study discusses developing a powerful method for diagnosis of heart disease, which is needed to introduce a meta-heuristic optimization algorithm to select the most relevant data from the heart disease dataset that will be used for the classification of heart disease [7]. In this study discusses K-means in classification and integration in KNN with genetic algorithms that improve accuracy classification for many datasets, not discuss accuracy.

## 2. The Proposed Method

### 2.1. Data Mining

Data mining is a method used to find knowledge that can be used to describe data in a database. Data mining has a number of processes that are quite a lot starting from the process of data sorting, data normalization and retrieving information from large amounts of data [8]. Mining data is very close to the analysis process, interpreting data that has a large source of information, and has meaningful patterns and rules. Some important things in data mining include the following [9].

1. Automation in a data has patterns and rules.

2. Automation on large data.

3. Automation in data mining to get meaningful improvement of patterns and references for indications and useful information.

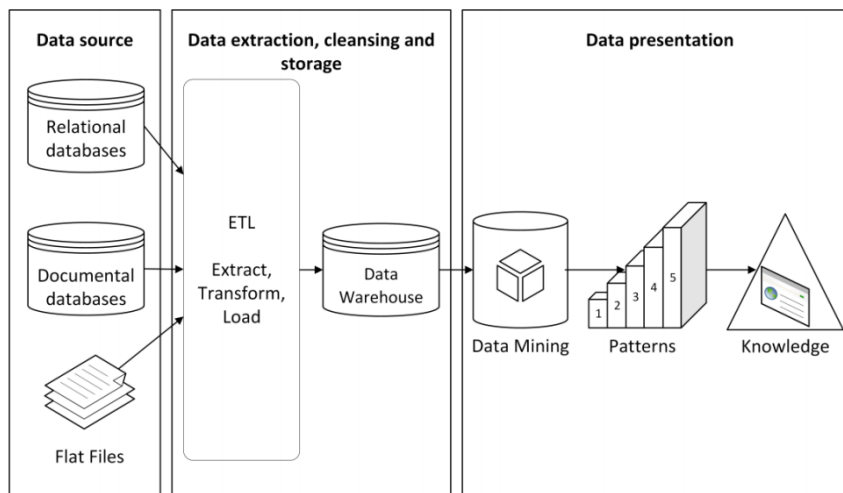Below we explain the process of mining data described in Figure 1 [10].

**Fig. 1.** Data mining process

Data mining has several process steps ranging from data sources in the form of relational and documental databases, as well as flat files which then go into the process of extracting, cleaning and storing data, here before becoming data warehouses that were previously irregular or widely modified and extracted, and loaded become a data warehouse. The last process from the data warehouse can be processed into data that has patterns and rules that are packaged in data mining, so that it can provide useful information and knowledge. The data presentation illustrates how data can be easily accessible sources of knowledge, previously from very large data sources and without benefits because temporary use can become data mining that can provide useful information [10].

## 2.2. Expert System

Expert systems are part of artificial intelligence, besides this system has been developed for a long time. Expert systems are closely related to knowledge summarized in large data, the source of this knowledge is the knowledge of an expert or expert in solving problems so that they get data patterns that can produce information to their users through a computerized system. Expert systems identify human knowledge and then create patterns or rows in large datasets and data entered into computers to solve problems based on cases or problems that are usually resolved by experts or experts [11]. In general, expert systems are very influential on the automation of a knowledge that is processed in a database, then the data will provide information needed by the user to solve problems that arise. The following is the Expert System work concept described in Figure 2 [12].



**Fig. 2.** Expert system proses
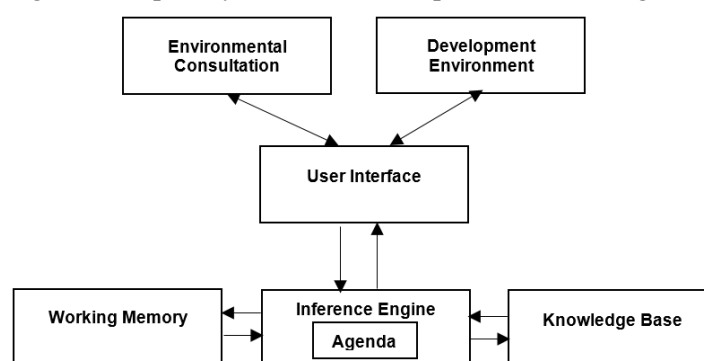
Expert systems are in accordance with the explanation in Figure 2. Including environmental consultations, the development of a human knowledge environment is compiled in a database, with the help of a user interface that will bridge scheduled and interconnected inference machines with knowladge bases and working memory used to solve problems that are usually resolved by experts [13].

### 2.3. K-Nearest Neighbor (KNN)

The KNN method is a data classification algorithm or object that has proximity to one another [14]. KNN can classify new data that has no known similarities to the data already owned, by selecting the number of its closest neighbors to find the similarity value with the new data. The object that has the closest similarity value to the k number is specified, then the new data object is selected. The specified k value uses odd number rules to get different distances in the data classification process [15]. Similarity is a way to calculate the value of the similarity of new data with existing data. The equation used in the calculation to find the value of similarity to a new object with the old object is described in equation (1) [16]:

$$Similarity(A, B) = \left( \sum_{x=1}^{z} f(A_x, B_x) \times w_x \right) : w_x \qquad (1)$$

Explanation of Equations :

- $A$  : New object
- $B$  : Old object in the data
- $z$  : Total attributes in each object
- $x$  : Individual attributes between 1 s.d. $z$
- $f$  : The function looks for the similarity of attributes between object $A$ and object $B$
- $w$  : The weight given between to $x$ attribute

The steps to calculate the K Nearest Neighbor method include [17]:

1. Determine the parameter $K$ (the nearest number of neighbors).
2. Calculate the square of the Euclid distance (example query) of each object against the given sample data using equation (1).
3. Then sort objects into groups that have the smallest Euclid distance.
4. Collect categories $Y$ (Nearby Neighbor Classification).
5. Using the Nearest Neighbor category, the majority can predict the sample query values that have been calculated.

## 3. Method

This study has several stages ranging from literature studies to experimental methods and through the process of testing in the KNN method to test the similarity symptoms of Diphtheria with data comparison. In this study applying the working concept of the Expert System and data mining processing from the knowledge base to get the results of early diagnosis of Diphtheria quickly so that it has not entered into complications of the disease that causes death.
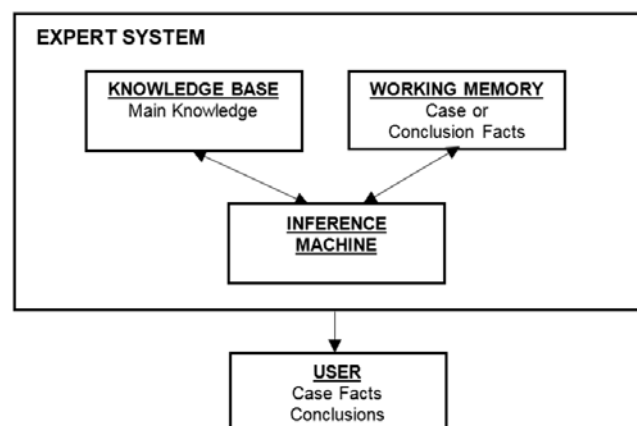


**Fig. 3.** Expert system flow

Figure 3 describes the main components of the Expert System structure including the Knowledge Base, Inference Engine, Working Memory, and Connected User Interface. The knowledge base

system structure and working memory are processed by the inference engine so that it gets conclusions or facts to the user.

## 4. Results and Discussion

The process of system design analysis in accordance with the steps in the flow of expert system creation ranging from data collection, interviews, calculations and testing methods, and the flow of the system design analysis process is presented in Figure 4 [18].

**Fig. 4.** Flow of System Design Analysis Process

The data collection process related to Diphtheria can use data from various valid sources that discuss Diphtheria, from previous research data on Diphtheria disease and its characteristics can be used for trials in this study. Expert interviews are needed to obtain information and validation of data collected about the disease. Specialists are the first reference in gathering information and knowledge about Diphtheria. After all data has been collected, then enter the diphtheria analysis process using the KNN method to find the value of similarity in the case that has occurred, and if there is a symptom input in a new case whether the data has a high similarity value of 8 0.8 and classified according to the class of diphtheria . The following is a flow chart that can describe the KNN work process from the start of symptom input until the diagnosis is stored in memory, the following is described in Figure 5.
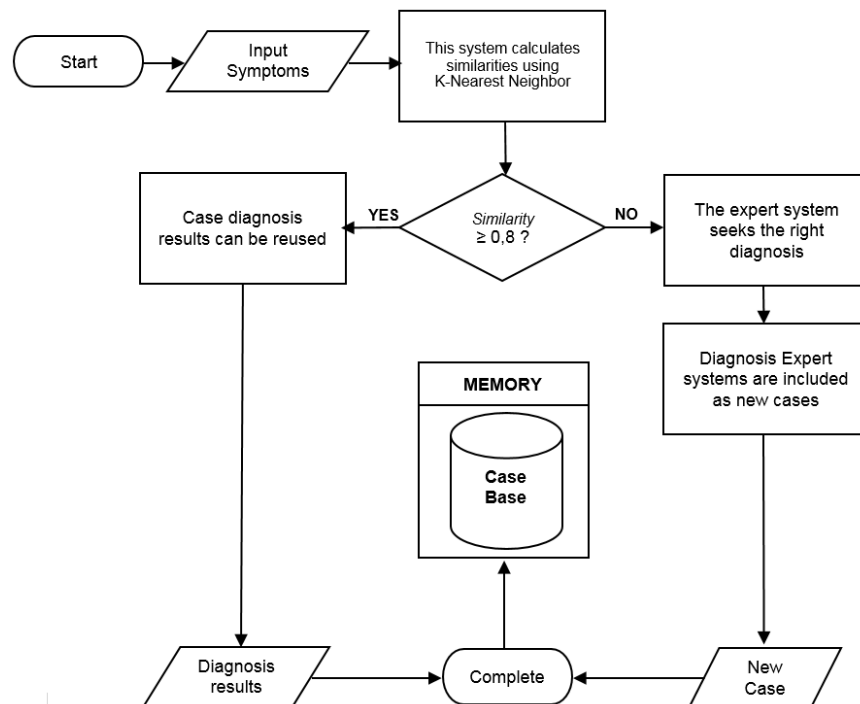
**Fig. 5.** System flow chart

The flow chart above explains that the first process that starts from symptom input will immediately go into the process of calculating the similarity value in the case using the KNN method. If the symptom result is $\geq 0.8$, the system will accompany the case diagnosis and can be reused to multiply the case or data on the knowledge base. If the value of the similarity in the resulting symptoms is $\leq 0.8$, the system will look for the right diagnosis so that the data will be a new case that can be stored in the database. The Knowledge Base will increase in number and get better if more new cases enter and provide a diagnosis of conclusions based on similarity values [19].

### A. Data Collection

Collecting data in this study is taking data in previous studies sourced from medical record data from RSUD Dr. Soetomo Surabaya. The data used are data on the disease and their symptoms and doctor's diagnosis of suffering from Diphtheria patients. The following are the data that has been collected and presented in Table 1 and Table 2 [20].

Table 1. Data Diagnosis of 148 Diphtheria Patients

| Diagnosis | Total Cases | Percentage |
|---|---|---|
| Age (years) means ± SD (range) | 5,11 ± 0,24 | 61,5%, 31,8% 6,7% |
| Gender (Female, Men) | 79, 69 | 1,14:1 |
| Before MRS (days) | | |
| a.   Means | 4,96 ± 2,41 | |
| b.   0 – < 5 | 77 | 52% |
| c.   ≥ 5 | 71 | 48% |
| City of residence | | |
| a.   Surabaya | 84 | 56,8% |
| b.   Outside Surabaya | 64 | 43,2% |
| Size of Diphtheria | | |
| a.   Light Casing | 2 | 1,4% |
| b.   Medium Casing | 62 | 41,9% |
| c.   Heavy Casing | 84 | 56,7% |
| DPT Immunization | | |
| a.   Without Vaccines | 23 | 15,5% |
| b.   Get a Vaccine | 125 | 84,5% |
| Nutrition Size | | |
| a.   Bad | 6 | 4,1% |
| b.   Lack or Good | 142 | 95,9% |

Diagnostic data in 148 patients with Diphtheria found that the sex census consisted of 79 men (53.4%) and 69 (46.6%) women. There were 91 (61.5%) patients aged less than 5 years, 47 (31.8%) aged 5 to 10 years, and 10 (6.7%) aged more than 10 years, with an average age of 5.11 year. Mostly, 77 (52%) sufferers 2-5 days after the new symptoms went to the hospital. After this other data that has been collected about the clinical diagnosis of Diphtheria patient cases is presented in Table 2.

Table 2. Clinical Diagnosis Data for Diphtheria Patients

| Clinical Symptoms | Total Cases | Percentage |
|---|---|---|
| Membrane location | | |
| a.   Unilateral tonsils | 19 | 12,8% |
| b.   Bilateral tonsils | 125 | 84,4% |
| c.   Pharynx | 6 | 4,0% |
| d.   Larynx | 3 | 2,0% |
| e.   Eye | 1 | 0,06% |
| Clinical Info | | |
| a.   Fever | 147 | 99,3% |
| b.   Throat pain | 92 | 62,2% |
| c.   Raucous | 29 | 19,6% |
| d.   Stridor | 59 | 39,9% |
| e.   Bullneck | 44 | 29,7% |
| Positive Culture | | |
| a.   Corynebacterium diphtheria | 33 | 22,3% |
| Complications | | |
| a.   Myocarditis | 17 | 11,5% |
| b.   Airway obstruction | 6 | 4,0% |
| c.   Acute Kidney Injury (AKI) | 6 | 4,0% |
| Died | 11 | 7,4% |

Analysis of clinical diagnostic data of Diphtheria patients as many as 125 of 148 with a percentage of 84.4% of patients complaining of pseudomembranes in bilateral tonsils and unilateral tonsil remains, pharynx, larynx and conjunctiva. Patients who go to the hospital come with hot clinical symptoms, but only 29 people with a percentage of 19.6% with clinical hoarseness. All patients who performed Corynebacterium diphtheriae culture received positive results as many as 34

patients with a percentage of 22.9%. Complications that often arise and are obtained are myocarditis in 17 patients from 148 patients with a percentage of 11.5%. Six patients with a 4% percent dose had tracheostomy therapy. Eleven patients with a percentage of 7.4% died, with the cause of death for myocarditis as many as 9 patients from 81.8% of patients who experienced complications.

### B. Expert Interview

Interviews with experts are mandatory in the expert system, in this case Specialist Doctors are the most dominant source for interviews regarding Diphtheria. Doctor at RSUD Dr. Soetomo Surabaya as the speaker's choice in previous research, included a discussion of diagnostic data obtained until later in the expert system workflow on research to disperse the problems raised in this study. In addition to analyzing attribute data and analysis of attribute weight values for Diphtheria which will be processed by the KNN method, the following are presented in table 3.

Table 3. Attribute and Weight Analysis

| No | Object | A | B | C | D | E | F | G | H | I | J |
|----|--------|---|---|---|---|---|---|---|---|---|---|
| 1 | Myocarditis | | | | | | | | | | |
| | < 5 | 1 | | | | | | | | | |
| | 5 – 10 | 0,5 | | | | | | | | | |
| | Male | | 1 | | | | | | | | |
| | Female | | 0,5 | | | | | | | | |
| | 0 – < 5 | | | 1 | | | | | | | |
| | Surabaya | | | | 1 | | | | | | |
| | Outside Surabaya | | | | 0,5 | | | | | | |
| | Medium diphtheria | | | | | 0,75 | | | | | |
| | Severe diphtheria | | | | | 1 | | | | | |
| | Not vaccinated | | | | | | 1 | | | | |
| | Vaccinated | | | | | | 0,5 | | | | |
| | Bad | | | | | | | 1 | | | |
| | Lack or Good | | | | | | | 0,75 | | | |
| | Unilateral tonsils | | | | | | | | 0,75 | | |
| | Bilateral tonsils | | | | | | | | 1 | | |
| | Pharynx | | | | | | | | 0,75 | | |
| | Larynx | | | | | | | | 0,5 | | |
| | Eye | | | | | | | | 0,25 | | |
| | Fever | | | | | | | | | 1 | |
| | Pain in the throat | | | | | | | | | 0,75 | |
| | Stridor | | | | | | | | | 0,25 | |
| | Bullneck | | | | | | | | | 0,5 | |
| | Corynebacterium diphtheria | | | | | | | | | | 1 |
| 2 | Airway Obstruction | | | | | | | | | | |
| | < 5 | 1 | | | | | | | | | |
| | > 10 | 0,5 | | | | | | | | | |
| | Female | | 1 | | | | | | | | |
| | ≥ 5 | | | 0,5 | | | | | | | |
| | Surabaya | | | | 1 | | | | | | |
| | Mild diphtheria | | | | | 0,5 | | | | | |
| | Medium diphtheria | | | | | 0,75 | | | | | |
| | Vaccinated | | | | | | 0,5 | | | | |
| | Bad | | | | | | | 1 | | | |
| | Lack or Good | | | | | | | 0,75 | | | |
| | Bilateral tonsils | | | | | | | | 1 | | |
| | Pharynx | | | | | | | | 0,75 | | |
| | Fever | | | | | | | | | 1 | |
| | Pain in the throat | | | | | | | | | 0,75 | |
| | Raucous | | | | | | | | | 0,25 | |
| | Bullneck | | | | | | | | | 0,5 | |
| | N/A | | | | | | | | | | 0,25 |

Table 3. Attribute and Weight Analysis (Continued)

| No | Object | A | B | C | D | E | F | G | H | I | J |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 3 | Acute Kidney Injury (AKI) | | | | | | | | | | |
| | < 5 | 1 | | | | | | | | | |
| | 5 – 10 | 0,5 | | | | | | | | | |
| | Male | | 1 | | | | | | | | |
| | 0 – < 5 | | | 1 | | | | | | | |
| | Surabaya | | | | 1 | | | | | | |
| | Mild diphtheria | | | | | 0,5 | | | | | |
| | Severe diphtheria | | | | | 1 | | | | | |
| | Not vaccinated | | | | | | 1 | | | | |
| | Bad | | | | | | | 1 | | | |
| | Unilateral tonsils | | | | | | | | 0,75 | | |
| | Bilateral tonsils | | | | | | | | 1 | | |
| | Larynx | | | | | | | | 0,5 | | |
| | Eye | | | | | | | | 0,25 | | |
| | Fever | | | | | | | | | 1 | |
| | Stridor | | | | | | | | | 0,75 | |
| | Corynebacterium diphtheria | | | | | | | | | | 1 |

Keterangan :
A : Age
B : Gender
C : Before MRS
D : City of residence
E : Size of Diphtheria
F : DPT immunization
G : Size of Nutrition
H : Membrane layout
I : Clinical info
J : Positive Culture

Analysis of the weight value obtained from the diagnosis data of the disease and clinical symptoms of patients with Diphtheria with the rule of weight values from 0 to 1, and references in giving their own weight to analysis of training data in 148 patient data and input from doctors, so that the data will be used as dataset in this study.

### C. Calculation method

The calculation process in the KNN method to get a diagnosis in patients with beru. The KNN method will process data from new patients with existing diagnoses and be adjusted to the attribute data that has been made, then adjusted to the weight of each attribute. Experiments were carried out on new patients who experienced symptoms that had indications for Diphtheria, the following are shown in Table 4.

Table 4. Diagnostic data for new patients

| New patient | | | |
|---|---|---|---|
| Lack or Good | 0,75 | < 5 | 1 |
| Unilateral tonsils | 0,75 | Male | 1 |
| Larynx | 0,5 | 0 – < 5 | 1 |
| Fever | 1 | Surabaya | 1 |
| Bullneck | 0,5 | Medium diphtheria | 0,75 |
| T / A | 0,25 | Unvaccinated | 1 |

Case 1 looks for similarity values with the old case which took 2 case examples that were most similar to the new case. The following is an example taken from research data which then looks at the value of similarity with new patients in table 4. Next in table 5.

Table 5. Weight values in the case of new patients and old patients

**Old Complications of Myocarditis Patients**

| Symptom diagnosis | Weight Value | | Similarity Value | |
|---|---|---|---|---|
| < 5 | $x_1$ | 1 | $y_1$ | 1 |
| Female | $x_2$ | 0,5 | $y_2$ | 0 |
| 0 – < 5 | $x_3$ | 1 | $y_3$ | 1 |
| Surabaya | $x_4$ | 1 | $y_4$ | 1 |
| Mild diphtheria | $x_5$ | 0,5 | $y_5$ | 0 |
| Vaccinated | $x_6$ | 0,5 | $y_6$ | 0 |
| Lack or Good | $x_7$ | 0,75 | $y_7$ | 1 |
| Bilateral tonsils | $x_8$ | 1 | $y_8$ | 0 |
| Pharynx | $x_9$ | 0,75 | $y_9$ | 0 |
| Eye | $x_{10}$ | 0,25 | $y_{10}$ | 0 |
| Fever | $x_{11}$ | 1 | $y_{11}$ | 1 |
| Pain in the throat | $x_{12}$ | 0,75 | $y_{12}$ | 0 |
| Bullneck | $x_{13}$ | 0,5 | $y_{13}$ | 1 |
| Corynebacterium diphtheria | $x_{14}$ | 1 | $y_{14}$ | 0 |

The data described in Table 5 are old data consisting of patient data, attributes, weights and attribute values of proximity of the similarity values in each symptom that will be processed in the calculation.

$$Similarity = \frac{\begin{array}{c}(x_1 \times y_1) + (x_2 \times y_2) + \\ (x_3 \times y_3) + (x_4 \times y_4) + \\ (x_5 \times y_5) + (x_6 \times y_6) + \\ (x_7 \times y_7) + (x_8 \times y_8) + \\ (x_9 \times y_9) + (x_{10} \times y_{10}) + \\ (x_{11} \times y_{11}) + (x_{12} \times y_{12}) + \\ (x_{13} \times y_{13}) + (x_{14} \times y_{14})\end{array}}{\begin{array}{c}y_1 + y_2 + y_3 + \\ y_4 + y_5 + y_6 + \\ y_7 + y_8 + y_9 + \\ y_{10} + y_{11} + y_{12} + \\ b_{13} + b_{14}\end{array}}$$

$$Similarity = \frac{\begin{array}{c}(1 \times 1) + (0.5 \times 0) + \\ (1 \times 1) + (1 \times 1) + \\ (0.5 \times 0) + (0.5 \times 0) + \\ (0.75 \times 1) + (1 \times 0) + \\ (0.75 \times 0) + (0.25 \times 0) + \\ (1 \times 1) + (0.75 \times 0) + \\ (0.5 \times 1) + (1 \times 0)\end{array}}{\begin{array}{c}1 + 0 + 1 + \\ 1 + 0 + 0 + \\ 1 + 0 + 0 + \\ 0 + 1 + 0 + \\ 1 + 0\end{array}}$$

$$Similarity = \frac{\begin{array}{c}1 + 0 + 1 + \\ 1 + 0 + 0 + \\ 0.75 + 0 + 0 + \\ 0 + 1 + 0 + \\ 0.5 + 0\end{array}}{\begin{array}{c}1 + 0 + 1 + \\ 1 + 0 + 0 + \\ 1 + 0 + 0 + \\ 0 + 1 + 0 + \\ 1 + 0\end{array}}$$

$$Similarity = \frac{\lfloor 5,25 \rfloor}{6}$$
$$Similarity = 0,875$$

The second experiment with the case of the second long patient then looked at the similarity values and described in table 6.

Table 6. Analysis of Attribution Analysis and Weight Value

**Old Patients Complicated with Airway Obstruction**

| Symptom diagnosis | Weight Value | | Similarity Value | |
|---|---|---|---|---|
| 5 – 10 | $a_1$ | 0,5 | $b_1$ | 0 |
| Fimale | $a_2$ | 0,5 | $b_2$ | 1 |
| ≥ 5 | $a_3$ | 0,5 | $b_3$ | 0 |
| Surabaya | $a_4$ | 1 | $b_4$ | 1 |
| Mild diphtheria | $a_5$ | 0,5 | $b_5$ | 0 |
| Vaccinated | $a_6$ | 0,5 | $b_6$ | 0 |

| Lack or Good | $a_7$ | 0,75 | $b_7$ | 1 |
|---|---|---|---|---|
| Pharynx | $a_8$ | 0,75 | $b_8$ | 0 |
| Larynx | $a_9$ | 0,5 | $b_9$ | 1 |

Table 6. Analysis of Attribution Analysis and Weight Value (Continued)

| Old Patients Complicated with Airway Obstruction | | | | |
|---|---|---|---|---|
| **Symptom diagnosis** | **Weight Value** | | **Similarity Value** | |
| Eye | $a_{10}$ | 0,25 | $b_{10}$ | 0 |
| Pain in the throat | $a_{11}$ | 0,75 | $b_{11}$ | 0 |
| Raucous | $a_{12}$ | 0,25 | $b_{12}$ | 0 |
| Stridor | $a_{13}$ | 0,75 | $b_{13}$ | 0 |
| N / A | $a_{14}$ | 0,25 | $b_{14}$ | 1 |

The data presented in Table 6 are old data consisting of patient data, attributes, weights and attribute values of the proximity of the similarity values in each symptom that will be processed in the calculation.

$$Similarity = \frac{\begin{bmatrix} (a_1 \times b_1) + (a_2 \times b_2) + \\ (a_3 \times b_3) + (a_4 \times b_4) + \\ (a_5 \times b_5) + (a_6 \times b_6) + \\ (a_7 \times b_7) + (a_8 \times b_8) + \\ (a_9 \times b_9) + (a_{10} \times b_{10}) + \\ (a_{11} \times b_{11}) + (a_{12} \times b_{12}) + \\ (a_{13} \times b_{13}) + (a_{14} \times b_{14}) \end{bmatrix}}{\begin{matrix} b_1 + b_2 + b_3 + \\ b_4 + b_5 + b_6 + \\ b_7 + b_8 + b_9 + \\ b_{10} + b_{11} + b_{12} + \\ b_{13} + b_{14} \end{matrix}}$$

$$Similarity = \frac{\begin{bmatrix} (0,5 \times 0) + (0,5 \times 1) + \\ (0,5 \times 0) + (1 \times 1) + \\ (0,5 \times 0) + (0,5 \times 0) + \\ (0,75 \times 1) + (0,75 \times 0) + \\ (0,5 \times 1) + (0,25 \times 0) + \\ (0,75 \times 0) + (0,25 \times 0) + \\ (0,75 \times 0) + (0,25 \times 1) \end{bmatrix}}{\begin{matrix} 0 + 1 + 0 + \\ 1 + 0 + 0 + \\ 1 + 0 + 1 + \\ 0 + 0 + 0 + \\ 0 + 1 \end{matrix}}$$

$$Similarity = \frac{\begin{bmatrix} 0 + 0,5 + 0 + \\ 1 + 0 + 0 + \\ 0,75 + 0 + 0,5 + \\ 0 + 0 + 0 + \\ 0 + 0,25 \end{bmatrix}}{\begin{matrix} 0 + 1 + 0 + \\ 1 + 0 + 0 + \\ 1 + 0 + 1 + \\ 0 + 0 + 0 + \\ 0 + 1 \end{matrix}}$$

$$Similarity = \frac{|3|}{5}$$
$$Similarity = 0,6$$

Using the diagnosis of the previous case to calculate the value of similarity in each symptom. If the minimum similarity value ranges from a value of 0.8 to 1, then the diagnosis of the old case will be reused for the new case. The new patient data will be tested sequentially with all the old patient data until the similarity values of each old patient data are obtained, then from the similarity data obtained, the highest one is obtained to get the desired results according to equation (1). The results of the above study indicate that the new patient data with the old patient data get the equation 0.875 and 0.6 so that it can be concluded that the new patient is more similar to the old patient data in the first trial which has a diagnosis of complications from myocarditis. In this study the new data obtained will continue to be stored in the database, so that the training data we get will be better and can be used as an initial diagnostic reference for the characteristics of Diphtheria.

*D.* **Testing**

The testing process on 148 records of Diphtheria patients, using the KNN method and 30 data testing using Cross Validation obtained the results of manual calculation details with 94% accuracy. The training data provided is 148 data from medical records. The test data used from the collected data is 80% of the total data, namely 148 training data, and the data used for the experiment is 20% of the 30 data data, then the trial process will be carried out 10 times with different data. from 148 data. The following data test results are presented in Figure 6.
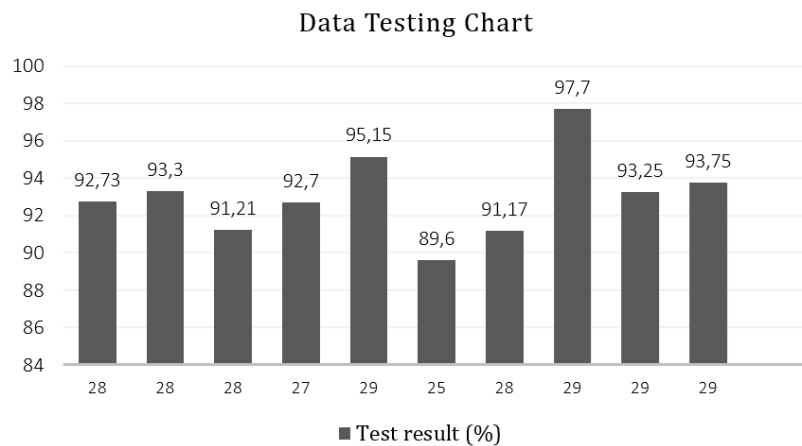
Figure 6. Data Testing Graph

The results of testing the data obtained by the highest results are estimated to reach 97.7% and then averaged with the total test results. The following detailed results from the tests are presented in table 7.

Table 7. Experiment data on new cases

| No | Number of Experimental Data | Suitability | Not Available | Level of Accuracy |
|----|----|----|----|----|
| 1 | 30 | 28 | 2 | 92,73% |
| 2 | 30 | 28 | 2 | 93,3% |
| 3 | 30 | 28 | 2 | 91,21% |
| 4 | 30 | 27 | 3 | 92,7% |
| 5 | 30 | 29 | 1 | 95,15% |
| 6 | 30 | 25 | 5 | 89,6% |
| 7 | 30 | 28 | 2 | 91,17% |
| 8 | 30 | 29 | 1 | 97,7% |
| 9 | 30 | 29 | 1 | 93,25% |
| 10 | 30 | 29 | 1 | 93,75% |
| | | | Rata-rata | 93,056% |

The results of testing new case data shows the results of testing medical record data with the number of data testing 30 new case data against 148 training data, obtained an accuracy rate of 93.056%.

## 5. Conclusion

The conclusion of this study that the KNN method can be used as an alternative to calculate distance and similarity of cases, such as the diagnosis of diphtheria. KNN can provide a fairly accurate diagnosis with 93.056% accuracy from 148 training data and 10 trials with 30 data tests. Expert System Analysis Diagnosis of Diphtheria with K-Nearest Neighbor Method can be done by modifying the method to get better accuracy and adding concepts to store data base knowledge for each new case with less similarity value.

## Acknowledgment

## References

[1] Kementerian Kesehatan Republik Indonesia , 2017, Profil Kesehatan Indonesia http://www.depkes.go.id/article/view/17121200002/menkes-difteri-menular-berbahaya-dan-mematikan-namun-bisa-dicegah-dengan-imunisasi.html, 01 June 2018.

[2] Barigou, F., 2016, Improving K-Nearest Neighbor Efficiency For Text Categorization, Neural Network World, Page. 45-65.

[3] Vedayoko, L.G., Sugiharti, E., Muslim, M.A., 2017, Expert System Diagnosis of Bowel Disease Using Case Based Reasoning with Nearest Neighbor Algorithm, Scientific Journal of Informatics, Vol. 4, No. 2, Page. 134-142.

[4] Kosasi, S., 2015, Pembuatan Aplikasi Diagnosa Kerusakan Mesin Sepeda Motor Matic dengan Case-Based Reasoning, Citec Journal, Vol. 2, No. 3, Page. 192-206.

[5] Punjani, Z., dkk., 2016, Automated Medical Diagnosis using K-Nearest Neighbor Classification, IMPACT FACTOR, Vol. 6, No. 4, ISSN 2231-4334.

[6] Shouman, M., Turner, T., Stocker, R., 2012, Applying k-Nearest Neighbour in Diagnosing Heart Disease Patients, International Journal of Information and Education Technology, Vol. 2, No. 3.

[7] Aro, T.O., Bajeh, A..O, Olatinwo, I.S., Omokanye, S.O., 2016, A Review On Prediction Of Heart Disease Using Techniques In Data Mining, International Conference of IEEE Nigeria Computer Chapter In collaboration with Dept. of Computer Science.

[8] Sulastri, H., Gufroni, A.I., 2017, Penerapan Data Mining dalam Pengelompokan Penderita Thalassaemia, Jurnal Teknologi dan Sistem Informas, Vol. 03, No. 02, ISSN 2476-8812.

[9] Listriani, D., Setyaningrum, A.H., Eka, F.M.A., 2016, Penerapan Metode Asosiasi Menggunakan Algoritma Apriori Pada Aplikasi Analisa Pola Belanja Konsumen ( Studi Kasus Toko Buku Gramedia Bintaro), Jurnal Teknik Informatika, Vol. 9 No. 2 Okt. 2016.

[10] Guzman, L.S., Dkk., 2015, Literature Review Of Data Mining Applications In Academic Libraries, The Journal of Academic Librarianship, Page. 499–510. http://doi.org/10.1016/j.acalib.2015.06.007.

[11] Suriyanti, 2013, Aplikasi Sistem Pakar Pendeteksian Kerusakan Printer dengan Case Based Reasoning, Pelita Informatika Budi Darma, Vol.V. No.3. ISSN: 2301-9425, Medan.

[12] Aribowo, A.S., Khomsah, S., 2011, Sistem Pakar Dengan Beberapa Knowledge Base Menggunakan Probabilitas Bayes Dan Mesin Inferensi Forward Chaining, semnasIF, ISSN: 1979-2328.

[13] Turban, Efraim, J.E. Aronson and Ting Peng Liang, 2005, Decision Support Systems and Intelligent Systems, Andi Offset, Yogyakarta.

[14] Putri, R.E., Suparti, Rahmawati, R., 2014, Perbandingan Metode Klasifikasi Naïve Bayes dan K-Nearest Neighbor Pada Analisis Data Status Kerja di Kabupaten Demak Tahun 2012, JURNAL GAUSSIAN, Vol. 3, No. 4, Page. 831–838, ISSN: 2339-2541.

[15] Rivki, M., Bachtiar, M., Dkk., 2017, Implementasi Algoritma K-Nearest Neighbor Dalam Pengklasifikasian Follower Twitter Yang Menggunakan Bahasa Indonesia, Journal of Information System, Vol.13, Issue 1.

[16] Kusrini, 2009, Algoritma Data Mining, Penerbit Andi, Yogyakarta.

[17] Johar, A., Yanosma, D., Anggriani, K., 2016, Implementasi Metode K-Nearest Neighbor (KNN) dan Simple Additive Weighting (SAW) Dalam Pengambilan Keputusan Seleksi Penerimaan Anggota Paskibraka, Jurnal Pseudocode, Vol.3, No.2, ISSN 2355-5920.

[18] Fatoni, C.S., Noviandha, F.D., 2018, Case Based Reasoning Diagnosis Penyakit Difteri dengan Algoritma K-Nearest Neighbor, Citec Journal, Vol. 4, No. 3, Page. 220-232.

[19] Jain, N., Srivastava, V., Husada, D., 2013, Data Mining Techniques: A Survey Paper, IJRET, Vol. 02, Issue 11, eISSN: 2319-1163.

[20] Puspitasari, D., Supatmini, E., Husada, D., 2012, Gambaran Klinis Penderita Difteri Anak Di RSUD Dr. Soetomo, Jurnal Ners, Vol. 7, No. 2, Page. 136–141.