Comparison of The Results of The Jaccard Similarity and K- Nearest Neighbor Algorithms Using The Case Based Reasoning(CBR) Method on An Expert System for Diagnosing Pediatric Diseases

Altundri Wahyu Hidayatullah [1], Dian Palupi Rini [2*], Osvari Arsalan[3], Kanda Januar Miraswan [4]

Informatics, Faculty of Computer Science, Sriwijaya University, Indonesia

[1] altundriiwahyu@gmail.com; [2]dprini@unsri.ac.id; [3] osvari.arsalan@ilkom.unsri.ac.id; [4] boy_kanda_jm@yahoo.com
* corresponding author

ARTICLE INFO

Article history
Received 16 January 2023
Revised 31 May 2023
Accepted 12 February 2024

Keywords
Case Based Reasoning (CBR)
Jaccard Similarity
K-Nearest Negihbor (KNN)
Expert System

ABSTRACT

Health ranks highest in supporting the continuity of every human activity, especially children. The availability of a doctor is still relatively lacking, especially in remote areas. This makes people have difficulty in diagnosing certain diseases so that medical treatment becomes too late and can even be fatal for the patient. So it is necessary to create a system that has the ability to be able to diagnose diseases in children like an expert. The method used in this study is Case Based Reasoning (CBR) with the Jaccard Similarity Algorithm and K-Nearest Neighbor. Jaccard Similarity is one way to calculate the similarity of two objects (items) which are binary. Similarity calculations are used to generate values whether or not there is a similarity between new cases and existing cases in the case base. While the K-Nearest Neighbor (KNN) Algorithm belongs to the instance-based learning group. The KNN algorithm allows the program to find old cases that are most similar to the current case. Based on the test results using 50 sample data, the expert system can provide diagnostic results in accordance with expert diagnoses. The accuracy results for the K-Nearest Neighbor Algorithm are 72% while the accuracy results for the Jaccard Similarity Algorithm are 70%.

1. Introduction

Children are very susceptible to germs, so as parents need to quickly obtain information about the disease suffered by the child even though there is no pediatrician so that parents must have sufficient knowledge to carry out initial treatment [1]. The availability of a doctor and medical personnel is still relatively lacking, especially in remote areas. This makes people have difficulty in diagnosing certain diseases so that medical treatment becomes too late and can even be fatal for the patient. So it is necessary to create a system that has the ability to diagnose disease symptoms in children as well as an expert.

In an expert system, you will be dealing with data that is ambiguous, vague and uncertain. Given the importance of a diagnostic result to be stored so that it can be reused in the future, it is necessary to create a system with case-based reasoning. Therefore we need an expert system based on the knowledge base of the previous case, namely using the Case Based Reasoning method[2].

Therefore, in this study, researchers will compare the results between the Jaccard Similarity and K-Nearest Neighbor (KNN) algorithms using the Case Based Reasoning (CBR) method. So it is hoped that it can help in diagnosing children's diseases without the need to come to a health expert by simply accessing the website so that the community, especially parents, can provide first aid to their children easily and quickly.

2. Literature Study

a. Expert System

Expert systems are commonly referred to as knowledge bases obtained from experience or knowledge of experts or experts aimed at assisting decision making in certain fields [3].The purpose of developing an expert system itself is not to replace the role of experts but to implement expert knowledge into software so that it can be used by many people at relatively little cost.

b. Child Disease

Immunity in young children is not as good and perfect as the immunity of adults. His knowledge and awareness of cleanliness and hygiene is also still very lacking. This is what makes small children susceptible to viral diseases. With the variety of health information and medical treatment that must be known, it is necessary to know in advance what symptoms a patient is suffering from, so that the type of disease and the first treatment for the disease can be known before being treated medically by a doctor. While not all parents have the medical knowledge to perform first aid on their children [4]. The following are 5 diseases that will be used in this study:

1. Dengue hemorrhagic fever (DHF) is a disease caused by infection with the dengue virus. Dengue fever is caused by one of four viral serotypes of the genus Flavivirus, family Flaviviridae.
2. Throat disease is a type of inflammatory disease that attacks the throat caused by viruses and bacteria, due to a weak immune system.
3. Morbili or rubeola is one of the causes of death in children. This disease is caused by the measles virus of the paramyxovirus group which is in the nasopharyngeal secretions and in the blood.
4. Chickenpox or varicella is a disease caused by the Varicella zoster virus (VZV). This disease is highly contagious, pandemic and seasonal
5. Typhoid fever is a systemic infection caused by Salmonella enterica serovar typhi (S. typhi). The incidence of this disease is often found in Asian countries and can be transmitted through contaminated food or water

c. Case Based Reasoning (CBR)

Case-Based Reasoning (CBR) is a problem solving approach by emphasizing the role of previous experience. New problems can be solved by reusing and perhaps making adjustments to problems that have similarities that have been solved previously . Case-Based Reasoning (CBR) has become a major paradigm in automated reasoning and machine learning. In CBR, someone who does reasoning can solve a new problem by paying attention to its similarities with one or more solutions to the previous problem [5].

In CBR (Case-Based Reasoning) there are 4 stages which include [6]:

1. Retrieve

   Retrieve or retrieve the case that most closely resembles or is relevant to the new case. This retrieve stage begins by describing or describing part of the problem, and ends if a match is found to the previous problem with the highest level of match. This section refers to the aspects of identification, initial match, search and selection and execution.

2. Reuse

   Modeling or reusing knowledge and information the old case based on the weight of the most relevant similarity to the new case, resulting in a proposed solution where an adaptation to the new problem may be needed.

3. Revise

   Reviewing the proposed solution and then testing it on real cases (simulations). If necessary, the solution will be corrected to match the new case.

4. Retain

Integrate or save new cases that have succeeded in getting a solution so that they can be used by further cases similar to the case. But if the new solution fails, then explain the failure, fix the used solution, and test it again.

d. K-Nearest Neighbor (K-NN)

The K-Nearest Neighbor (K-NN) algorithm belongs to the instance-based group learning. This algorithm is also a lazy learning technique. K-NN This is done by looking for groups of k objects in the most appropriate training data close (similar) to the object in the new data or testing data. K-NN Algorithm allows the program to search for the old case that is most similar to the case at hand now [7].

The formula for calculating the similarity weight (Similarity) with K-NN [8], namely:

$$Similirity(P,Q) = \frac{S1 \times W1 + S2 \times W2 + \cdots + S_n \times W_m}{W1 + W2 + \cdots + W_n} \quad (1)$$

Where:

P = New Case

Q = Cases in storage

W = weight (the weight given to the i attribute)

S = similarity (similarity)

N = amount of data

e. Jaccard Similarity

Jaccard Similarity is one way to calculate the similarity of two objects (items) which are binary. Similarity calculations are used to generate a value whether or not there is a similarity between new cases and existing cases on the case base [9]. Jaccard similarity can be formulated as follows:

$$Jaccard\ Similarity\ (A, B) = S = |A \cap B|\ |A| + |B| - |A \cap B| \quad (2)$$

Where:

A = New Case

B = Cases in Storage

|A ∩ B | = The number of symptoms of old cases and new cases

f. Rational Unified Process (RUP)

The Rational Unified Process (RUP) is a software engineering method developed by collecting various best practices found in the software development industry. The main feature of this method is that it uses a use case driven and iterative approach to the software development cycle. RUP uses an object oriented concept, with activities that focus on model development using the Unified Model Language (UML) [10]. RUP has four phases of system development including:

1. Inception is the initial stage in the RUP process which aims to gain understanding from all interested parties. Things that are usually done in this phase are analyzing benefit costs, analyzing initial risks, recording system requirements (requirements) and so on.
2. Elaboration is a stage for developers to carry out a complete design based on the results of the analysis at the inception stage.
3. Construction is the stage where the system development process is executed. The implementation and testing process will be carried out at this stage so that it will produce software

4. Transition is the last stage in the RUP process which focuses on the maintenance of the software that has been created.

3. Methodology

a. Data Collection

Data collection techniques are divided into two, namely primary and secondary data. The method for collecting primary data is to conduct direct interviews with doctors or an expert to evaluate the knowledge acquisition process in building a knowledge base. While secondary data is obtained by collecting data from previous research journals.

The list of symptoms and diseases is shown in table 1 where data is obtained through journals which have been verified by experts through interviews.

Table 1. Table of Symptoms of Early Diagnosis of Childhood Disease

No	Symptoms List	Symptoms Code
1	Fever	G01
2	Cough	G02
3	Red eyes and sensitive to light	G03
4	Skin Lesions Appear in The Form Of Erythematous Papules	G04
5	Headache	G05
6	Sore Throat	G06
7	Nausea and vomiting	G07
8	Itchy	G08
9	Decreased Appetite	G09
10	Loss of Consciousness	G10
11	Body Weak and Lethargic	G11
12	Easily Restless	G12
13	Diarrhea	G13
14	Swollen Glands in The Neck	G14
15	Platelet Levels Down	G15
16	Nosebleed	G16
17	Dehydration	G17
18	Muscle Ache	G18
19	Heartburn	G19

The data for 5 pediatric diseases are shown in Table 2.

Table 2. Table of Child Disease

No	Disease List	Disease Code
1	Dengue Hemorrhagic Fever	P01
2	Laryngitis	P02
3	Morbili	P03
4	Varicella	P04
5	Typhoid Fever	P05

b. Framework

In this study, a comparison will be made between the two algorithms, namely Jaccard Similarity and KNN. The process will start with the user entering the appropriate symptom. After that, new cases will be checked against old cases on knowledge data in the system. Followed by calculations using the Case Based Reasoning method with the Jaccard Similarity Algorithm and will also be carried out calculations using the Case Based Reasoning method with the KNN Algorithm. After doing the calculations with the two algorithms, it will display the results of calculations from each algorithm and will display the results of the diagnosis of childhood diseases where the results of calculations that have a percentage above 50% can be given a solution while the results of calculations below 50% cannot be given a solution.

The framework in this research can be seen in Figure 1 below.

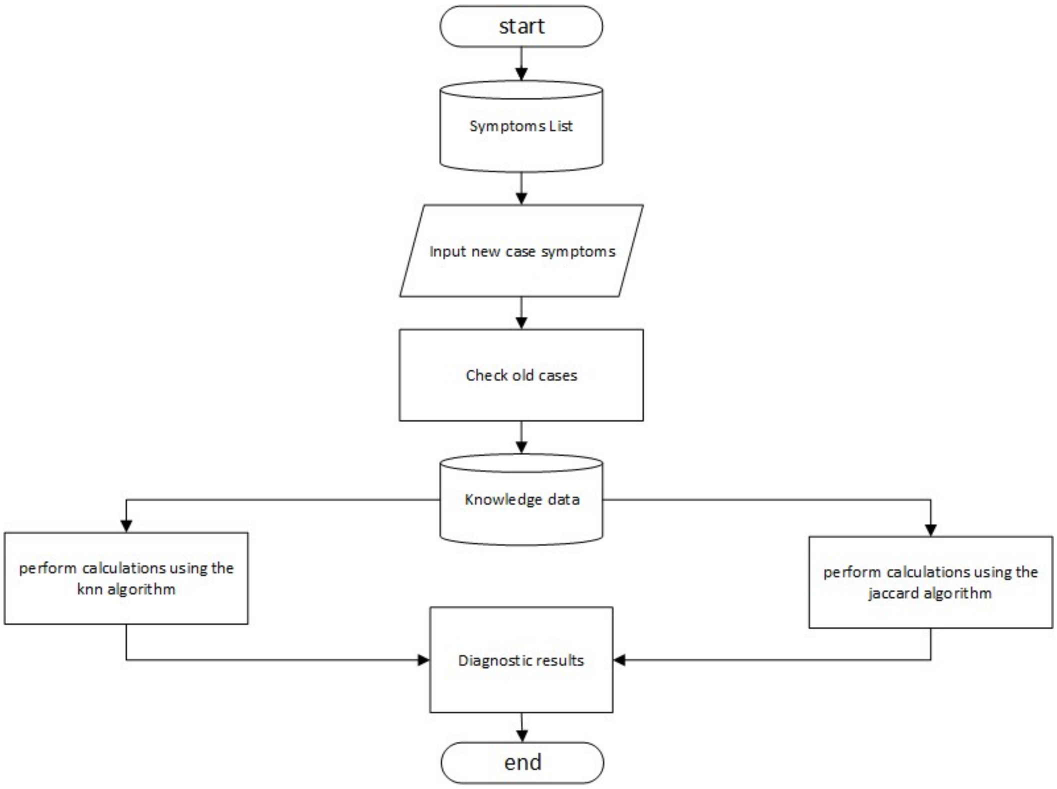
start
Symptoms List
Input new case symptoms
Check old cases
Knowledge data
perform calculations using the knn algorithm
perform calculations using the jaccard algorithm
Diagnostic results
end

Fig 1. Research Framework

4. Result and Discussion

This test is carried out using data that has been taken by researchers through a questionnaire totaling 50 data that have been consulted with experts to get diagnostic results from experts. the experiment results can be seen in the following table.

Table 2. Experiment Result

	Expert Diagnostic Results	System Diagnostic Results		Conclusion (Same/NotSame)	
		K-NN Algorithm	Jaccard Algorithm	K-NN Algorithm	Jaccard Algorithm
1	Laryngitis	Laryngitis	Laryngitis	Same	Same
2	Dengue HemorrhagicFever	Dengue HemorrhagicFever	Dengue HemorrhagicFever	Same	Same
3	Varicella	Varicella	Typhoid Fever	Same	Not Same
…	……..	……..	……..	……..	……..
50	Varicella	Varicella	Typhoid Fever	Same	Not Same

From the results obtained between the results of system diagnostics and expert diagnoses, the results obtained in the form of accuracy values can be calculated by the following formula:

$$\text{Accuracy Value} = \text{appropriate amount of data} / \text{amount of testing data} \times 100\% \quad (3)$$

Then the accuracy value for the K-Nearest Neighbor Algorithm:

$$\text{Accuracy Value} = 36 / 50 \times 100\% = 72\%$$

and the accuracy value for the Jaccard Similarity Algorithm:

$$\text{Accuracy Value} = 35 / 50 \times 100\% = 70\ \%$$

Based on the above calculation results, the percentage value of the system accuracy level using the K-Nearest Neighbor Algorithm is 72% and for the Jaccard Similarity, the accuracy value is 70%. In the KNN Algorithm there are 36 results that have a percentage above 50% so that a solution can be given and the Jaccard Algorithm there are 18 results that have a percentage above 50% so that a solution can be given. In this study the K-Nearest Neighbor has a greater accuracy of 2% compared to the Jaccard Similarity Algorithm.

5. Conclusion

Based on the explanation in the previous chapter, namely implementation, it can be concluded that:

1. The Jaccard Similarity Algorithm and KNN (K-Nearest Neighbor) Algorithm can be applied to the Pediatric Disease Diagnosis Expert System. By looking for the similarity value of the similarity of the old case and the new case. The KNN algorithm will perform calculations with the Nearest Neighbor Retrieval formula to get the results of the diagnosis and continue using the Jaccard Algorithm so as to produce disease diagnoses from the two algorithms.

2. Based on the data that was successfully tested in this study, it shows that the expert value produces an accuracy for the K-Nearest Neighbor Algorithm of 72% and for the Jaccard

Similarity Algorithm using the subjective value of the expert produces an accuracy value of 70% so that there is a difference between the two methods of 2%. where the KNN Algorithm has greater accuracy than the Jaccard Similarity Algorithm.

References

[1] A. F. Indriani, E. Y. Rachmawati, and J. D. Fitriana, “Pemanfaatan Metode Certainty Factor dalam Sistem Pakar Diagnosa Penyakit pada Anak,” *Techno.Com*, vol. 17, no. 1, pp. 12–22, 2017, doi: 10.33633/tc.v17i1.1576.

[2] C. R. Pasalli, V. Poekoel, and X. Najoan, “Sistem Pakar Diagnosa Penyakit Anak Menggunakan Metode Forward Chaining Berbasis Mobile,” *J. Tek. Inform.*, vol. 8, no. 1, 2016, doi: 10.35793/jti.7.1.2016.12828.

[3] T. R. Maulidia, “Membuat Sistem Pakar Jauh Lebih Besar Dari Pada Pembuatan Sistem Biasa . Pakar Digunakan Untuk Memecahkan Masalah Yang Memang Sulit Untuk Dipecahkan Dengan Pemrograman Biasa , Mengingat Biaya Yang Diperlukan Untuk,” *Coding J. Komput. dan Apl. Untan*, vol. 05, no. 03, 2017.

[4] M. Sari, S. Defit, and G. W. Nurcahyo, “Sistem Pakar Deteksi Penyakit pada Anak Menggunakan Metode Forward Chaining,” *J. Sistim Inf. dan Teknol.*, pp. 130–135, 2020, doi: 10.37034/jsisfotek.v2i4.34.

[5] A. F. Prayuda, S. Wibisono, and W. Hadikurniawati, “Implementasi Sistem Pakar untuk Rekomendasi Masakan Tradisional Jawa dengan Metode Case Based Reasoning Menggunakan Algoritma Similaritas Czekanowski,” *Pros. SENDI_U*, pp. 978–979, 2018.

[6] F. Fatmayati, - Kusrini, and E. T. Lutfi, “Implementasi Case Base Reasoning Untuk Mendiagnosa Penyakit Gigi dan Mulut,” *Techno.Com*, vol. 16, no. 1, pp. 70–79, 2017, doi: 10.33633/tc.v16i1.1331.

[7] F. D. Wahyudi, D. Remawati, and P. Harsadi, “Sistem Pakar Deteksi Kerusakan Mesin Bubut Dengan Metode Knn,” *J. Teknol. Inf. dan Komun.*, vol. 6, no. 2, pp. 7–13, 2019, doi: 10.30646/tikomsin.v6i2.370.

[8] D. A. Pranggono, Sabar, “Sistem Pakar Diagnosa Penyakit Kucing Menggunakan Metode Forward Chaining (Fc) Berbasis Web,” pp. 1–27, 2017.

[9] A. Gunawan, C. Suhery, and T. Rismawan, “Implementasi Metode Case-Based Reasoning Dan Similarity Jaccard Coefficient Dalam Identifikasi Kerusakan Laptop,” *Coding J. Komput. dan Apl.*, vol. 09, no. 02, pp. 292–305, 2021.

[10] A. R. Fabiyanto, Y. T. Mursityo, and D. Pramono, “Pengembangan Sistem Informasi Penilaian Kinerja Guru Menggunakan Metode Rational Unified Process (RUP) Berbasis Web (Studi Pada SD Negeri Prigen 1),” *J. Pengemb. Teknol. Inf. dan Ilmu Komput.*, vol. 3, no. 4, pp. 3888–3895, 2019.