Application of the Minkowski Distance Similarity Method in Case-Based Reasoning for Stroke Diagnosis

Angelina Rumuy[1], Rosa Delima[2*], Kuncora Probo Saputra[3], Joko Purwadi[4]

[1,2,3,4]*Informatics Department, Universitas Kristen Duta Wacana, Indonesia*

*corr_author: rosadelima@staff.ukdw.ac.id

Abstract - A Stroke is a cerebrovascular disease characterized by impaired brain function due to damage or death of brain tissue caused by reduced or blocked blood and oxygen flow to the brain. Expert systems can be used as learning aids for medical students to diagnose stroke. Medical records of stroke cases can be reused as a reference for diagnosing stroke when there are new cases, known as the case-based reasoning (CBR) method. This study implements the Minkowski distance similarity method in CBR to calculate the similarity value between cases, where each similar case has the same solution. This study uses the Minkowski distance similarity method in CBR to obtain the most optimal value of r and the most appropriate threshold value in the expert system for stroke diagnosis. The diagnosis process is carried out by inputting the patient's condition, symptoms, and risk factors. Then, the system will calculate the similarity value and take the case with the highest similarity value as the solution, providing that the similarity value must be greater than or equal to the threshold value. Based on system testing, the best accuracy value was achieved by applying a threshold value of 75 with an *r*-value of 3 or 4, with an accuracy rate of 88.89%, a recall value of 88%, and a precision of 100%.

Keywords: Case-based reasoning, stroke diagnosis, Minkowski distance similarity

I. INTRODUCTION

A stroke is a cerebrovascular disorder that happens when blood and oxygen flow to the brain is obstructed, causing brain tissue to die or be damaged [1]. Technology has advanced quickly as a means of increasing production and efficiency across a variety of industries. Expert systems are one of the technologies that are developing. Expert systems are intelligent systems and a subset of Artificial Intelligence (AI) that use an expert's expertise to teach a computer how to solve problems accurately, much like an expert would [2]-[3]. A knowledge base, inference engine, working memory, and user interface are needed to develop an expert system [2], [4].

A knowledge base is a component that contains facts and rules, where facts represent information about objects, and rules are used to derive new facts from known ones. An inference engine is a component that searches for connections between the rules in the knowledge base and the input facts, working memory, database, and user interface. Working memory contains data received from the user during the expert system session, and the user interface provides facilities for interaction between the user and the system [2].

In the field of medicine, numerous systems have been created, including [5]- [9]. Expert systems can be used in the classroom to diagnose strokes as a learning tool. The Case-Based Reasoning (CBR) method allows for the reuse of medical records from previous stroke cases as references when diagnosing new stroke cases. Expert systems use Case-Based Reasoning to solve problems by remembering and using prior information and experiences [10]-[11]. If a problem is successfully resolved during the CBR process, the solution will be saved to address similar problems in the future. If it cannot resolve a problem, the case will be noticed and stored to help prevent the same mistake in the future [12].

The four stages of the CBR process are retrieve, reuse, revise, and retain. Retrieve is the process of determining the issue and comparing it to prior cases. When a new case and the cases in the case base are compared, a similarity value is calculated, and the old case with the highest similarity value to the new case is chosen. Identifying the degree of case similarity is the most important part of this stage [13]. The system employs the reuse process to search the database for comparable earlier circumstances to discover a solution for the current problem. The system then utilizes knowledge from earlier, comparable cases to address the new issue [14]. An expert's revision process involves making the offered solution better. If successful, the new case will be saved in the database alongside the new solution, and retaining involves merging or saving new

cases that have achieved solutions successfully for reference in subsequent instances similar to these [13].

Calculating document similarity at the retrieval stage becomes a crucial component of the CBR system. The level of document similarity is calculated using this formula. The Minkowski Distance Similarity approach, a generalization of the Euclidean Distance and Manhattan Distance methods, is one of the numerous methods to determine the degree of similarity between a new case and cases in the case base in CBR [15]. The only distinction between these techniques is the magnitude of r, the Minkowski power constant. The accuracy of the system being created is significantly impacted by the values chosen for r and the threshold.

There have been numerous studies on CBR for disease diagnosis, such as Nelson et al.'s 2018 study on CBR for stroke diagnosis using the Jaccard Coefficient technique, applying the Siriraj Score to distinguish between ischemic and hemorrhagic stroke, and then employing dense indexing [16]. A threshold value of 0.7 resulted in greater sensitivity (89.88%) and accuracy (81.67% with indexing and 84.44% without indexing) compared to threshold values of 0.8, 0.9, and 1. The system was tested with 45 cases as test data and 135 cases as the case base.

Using the K-Nearest Neighbour algorithm with an 80% threshold value, Zainuddin et al. conducted research on CBR for diagnosing stroke in 2016 and reported that out of 15 evaluated cases, the system properly diagnosed 93.3% of cases, according to expert diagnosis [17]. Using CBR and the K-Nearest Neighbour approach for calculating distance, Warman et al. [18]. They investigated an expert system for spotting illnesses in rice plants. Fifty-two test data sets with a threshold value of 70% were used to assess the system's sensitivity and accuracy; the results showed that the system's sensitivity was 100%, and its accuracy rate was 82.69%.

Minkowski Distance method has been used in several studies, including [19]- [23]. This study aims to develop an expert system that uses the Minkowski Distance Similarity approach in CBR to diagnose strokes with the maximum degree of accuracy and the most suitable threshold value.

This essay is divided into four pieces, starting with an introduction detailing the research background and a literature review on the method's use. The system development and testing process is covered in the second section. The results and discussion are covered in the third part. The paper's conclusion concludes the research findings.

II. METHOD

Four stages comprise the system development process: requirement analysis, system design, implementation of the program code, and system testing. The research by Nelson et al. in 2018, titled Case-Based Reasoning for Stroke Diseases Diagnosis, provided the data for this study. It is made up of data from the medical records of stroke patients treated at Yogyakarta's DKT Dr. Soetarto Hospital during 2015 and 2016 [16]. Based on its etiology and anatomical pathology, stroke is divided into four categories: embolic, thrombotic, subarachnoid hemorrhage, and intracerebral hemorrhage.

A. System Design

The American Stroke Association [24] states that stroke symptoms generally include face dropping, arm weakness, and slurred speech. Face drooping is when one side of the face droops or is numb, and arm weakness is when one arm is weak or numb. Apart from that, several other symptoms can accompany a stroke, including difficulty walking, vision problems, confusion, weakness in one part of the body, and headaches without knowing the cause. In this study, 42 symptoms were defined that accompany the diagnosis of the disease. Apart from the main symptoms, twelve risk factors stroke patients have, including a history of various diseases such as hypertension and diabetes mellitus, a family history of stroke, obesity, and smoking habits. Case representation can be seen in Table I.

The CBR approach was used to create this expert system, which will take user input, including the patient's personal information, symptoms, and known risk factors. The system will calculate the local and global similarity values between the new case data and the case base. The solution to the latest case, where the similarity value must be above the threshold, will be taken from the example with the highest similarity value. The case will be stored in the case base and updated by an expert if the similarity value does not surpass the threshold. The system will output the patient's type of stroke disease if the value exceeds the threshold. The CBR system's process for identifying stroke disorders is shown in Fig. 1.

The expert system construction uses the CBR approach and the Minkowski Distance Similarity method to determine how similar new and old examples are. The user enters information into the system, including the patient's personal information, symptoms, and known risk factors. The system then uses the Minkowski Distance Similarity method to determine the local and global similarity values between the new case data (user

input) and the old cases in the case base. The case with the highest similarity value—which must be more than the threshold—will be utilized as the answer to the new case. The case will be stored in the case base and updated by an expert if the similarity value does not surpass the threshold. The system will then output the name of the patient's specific type of stroke disease.

TABLE I
EXAMPLE OF CASE REPRESENTATION [16]

Case-based		
Patient Code:		K00007
General Condition:		
1	Age	60
2	Gender	Male
3	Awareness	Compos Mentis
Symptom:		
G1	Confusion	No
G3	Trouble balancing	No
Gn	n-th symptom	…
Risk Factor:		
FR1	History of heart disease	No
FR2	History of hypertension	Yes
FRn	n-th risk factor	…
Diagnosis:	Embolism Stroke	

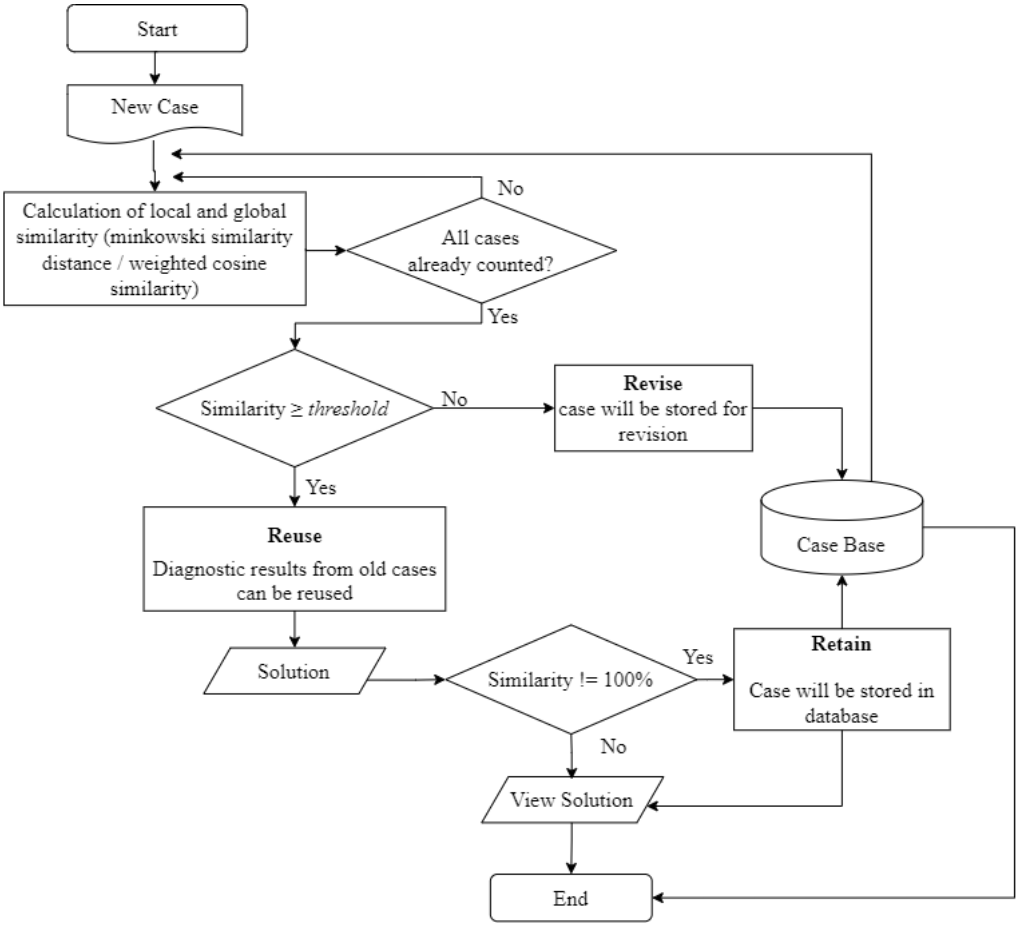


Fig. 1 Flowchart of the CBR system for stroke diagnosis

The goal of similarity measurement is to calculate the degree of similarity between two objects. Local and global similarity values are the variables that determine similarity values. Local similarity measurement aims to calculate the degree of similarity between two things. The global similarity calculates the similarity between a

problem and cases in the case base. Local similarity computation assesses the degree of similarity between problem attributes and identical attributes from a case. Based on the data type of the features, the local similarity is calculated [25]. The two data types in local similarity are boolean and numeric. Eq. (1) and (2) show, respectively, the formula for local similarity with numeric and boolean data types, where *s* and *t* are the values of the features being compared, *R* is the range of values for that feature for numeric data, and *s*, *t* is {*true*, *false*} for boolean data.

$$f(s,t) = 1 - \frac{|s-t|}{R} \quad (1)$$

$$f(s,t) = \begin{cases} 1, jika\ s = t \\ 0, jika\ s \neq t \end{cases} \quad (2)$$

This study will analyze the system's accuracy using Minkowski Distance Similarity with both local dan global similarity. Fig. 2 illustrates calculating local and global similarity with Minkowski Distance Similarity. The formula for calculating local and global similarity using Minkowski Distance Similarity [22] is shown in (3).

$$\left[\frac{\sum_{k=1}^{n} w_k^r * |d_k(C_{ik}, C_{jk})|^r}{\sum_{k=1}^{n} w_k^r}\right]^{1/r} * T(C_j) * \frac{n(C_i, C_j)}{n(C_i)} \quad (3)$$

Eq. (3) is the formula for calculating global similarity where $E(C_i, C_j)$ is a global similarity between the target case (C_i) and the source case (C_j), w_k Attribute k weight value, and $d_k(C_{ik}, C_{jk})$ is a local similarity value between the k-th attribute of the target case and the k-th attribute of the source case . variabel r is a Minkowski factor (positive integer), $T(C_j)$: Confidence level of the case in the case base, $n(C_i,\ C_j)$ is the total attributes of the target case (C_i)that appear in the source case (C_j), and $n(C_i)$: Total number of attributes in the target case (C_i).

B. System Coding

The system development was carried out as web-based software using HTML/CSS and PHP, with Apache as the web server and MySQL as the database.

C. System Testing

The confusion matrix approach is used to test the system, and the results include numbers for accuracy, recall or sensitivity, and precision. The confusion matrix is a table that lists the outcomes of test data that were correctly and wrongly labeled [26]. The matrix will compute accuracy, precision, and recall by comparing the actual and anticipated values, as shown in Table II. Eq. (4)–(6) illustrate the relationship between accuracy, precision, and recall.

$$accuracy = \frac{TP+TN}{Total} \quad (4)$$

$$precision = \frac{TP}{TP+FP} \quad (5)$$

$$recall = \frac{TP}{TP+FN} \quad (6)$$

The testing uses information from [16] study, which includes 180 cases worth of medical records from stroke patients treated at DKT Dr. Soetarto Hospital in Yogyakarta between 2015 and 2016. Of those, 54 cases, or 30% of the total, are used as test data. Different threshold values, including (0.6), (0.65), (0.7), (0.75), (0.8), (0.85), (0.9), and (0.95), are used to test the system. The applied Minkowski distance similarity exponent (r) will rise from 1 until there are no longer noticeable changes in the system's accuracy.

III. RESULT AND DISCUSSION

A. Results of the System

The patient's overall health, current symptoms, and risk factors are input into the system. The system evaluates the new case's highest level of similarity to the case base's imposed threshold. A solution is offered if the similarity value is equal to or more than the threshold. An expert will revise the case if the similarity value falls below the cutoff. Otherwise, it cannot be solved. Fig. 3 shows the diagnosis interface implemented, along with the diagnosis outcomes.

TABLE II
CONFUSION MATRIX VARIABEL

		Actual Value	
		Positive	**Negative**
Prediction Value	Positive	True Positive (TP)	False Negative (FP)
	Negative	False Negative (FN)	True Negative (TN)

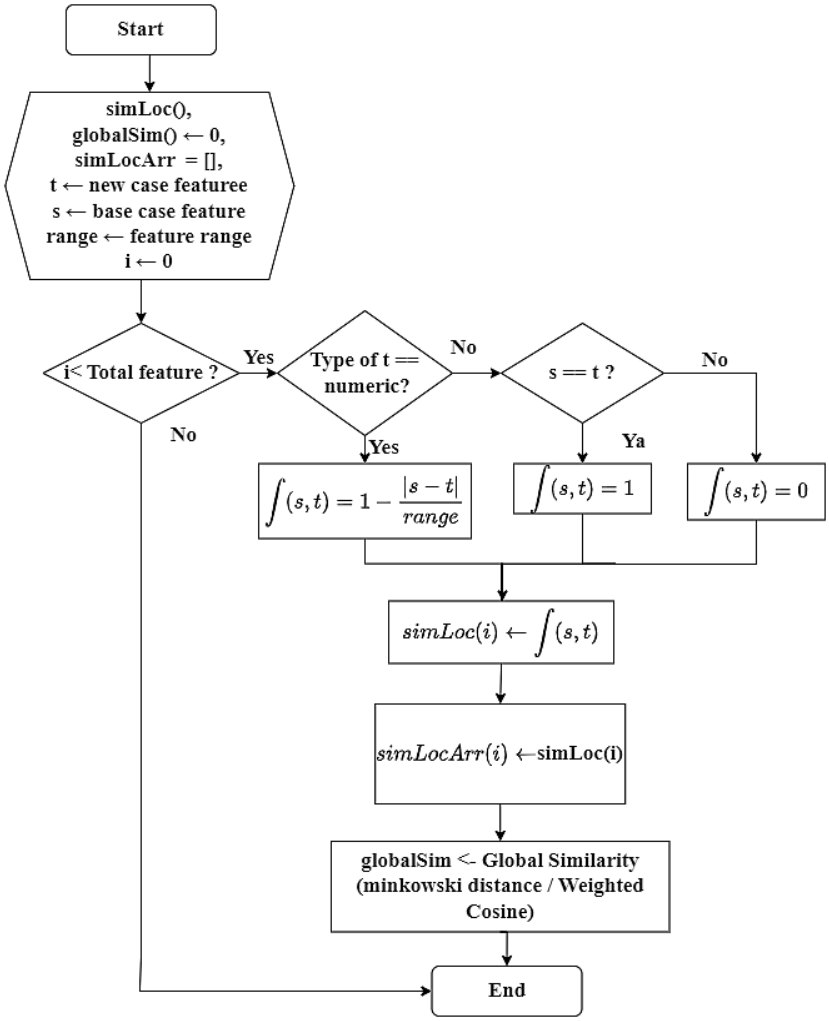
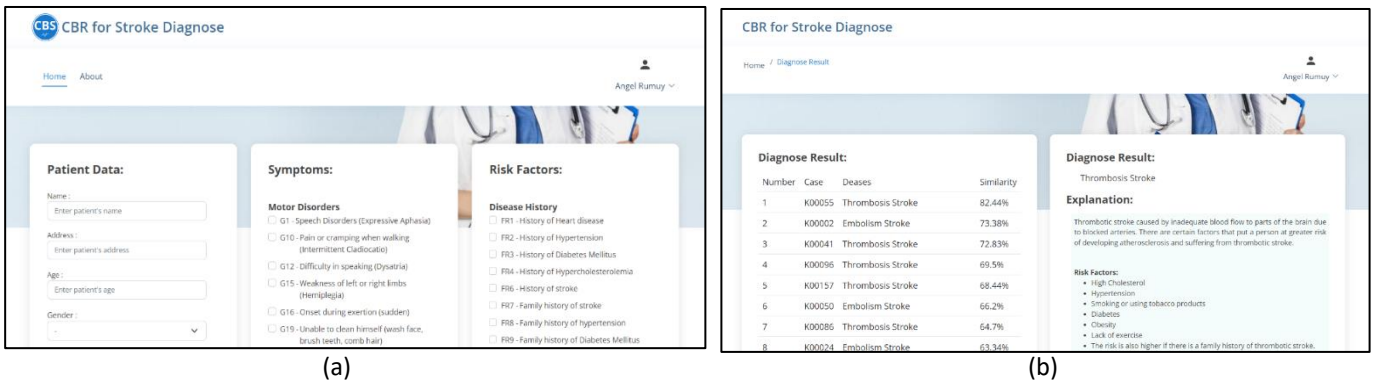


Fig. 2 Minkowski similarity calculation steps on the system



(a) (b)

Fig. 3 Diagnosis (a) and result page (b)

B. CBR Process on the System

By determining the similarity between the new case and the case base, the retrieval process comprises looking for examples comparable to the new case. The Minkowski distance similarity method is used in the similarity computation.

Based on the data type of the characteristics, the similarity level between the features of the new case and the case base is calculated to determine local similarity. Equation (1) is used to calculate the numerical data types for features like age, gender, systolic and diastolic blood pressure, headache, temperature, pulse, condition, and level of consciousness, while Equation (2) is used to calculate the boolean data types for features like symptoms and risk factors. The global similarity is calculated by giving each characteristic a weight, determining the expert's confidence level in the case, and utilizing Minkowski distance similarity to calculate the global similarity value.

The reuse procedure is implemented by employing the previous examples with the most significant similarity to the current case. The case with the highest degree of similarity is then obtained, and its similarity value is contrasted with the threshold. The old case can be used as a solution for the new instance if the similarity value is greater than or equal to the threshold. The case will instead move onto the revision phase if the similarity value is below the cutoff.

The expert revises the case solution that has been provided as part of the procedure. The system will update the case's diagnosis results when the expert presses the revise button. Figure 4 depicts the expert's application of the revised procedure.

New cases are kept in the case base as part of the retain procedure. The case will be saved into the case base when the user clicks the save button to be utilized as a solution for upcoming new cases. The case will be added to the case base with the diagnosis result revised if it moves into the revision phase so that the expert can revise it.

C. System Testing Results

Eq. (4), (5), and (6) are used to calculate accuracy, precision, and recall during system testing. A total of 54 cases, or 30% of the case data, are used in the testing. For efficiency, the testing is carried out using automation scripts, which log into the system and automatically fill in patient, symptom, and risk factor information depending on test data. The results (accuracy, precision, and recall) are then recorded in Excel and PDF files.

Fig. 5 exhibits the system testing outcomes using the confusion matrix and displays various accuracy, sensitivity, and recall levels for each threshold and value r. The maximum level of accuracy is attained when a threshold value of 75 is combined with values for r = 3 and r = 4, as illustrated in Figure 6. The accuracy rating represents how well the system can diagnose. Hence, a more excellent accuracy value means the system will deliver more accurate diagnosis results or solutions. Figure 7 compares the accuracy levels of the two systems at a threshold of 70, demonstrating that the Minkowski Distance method system has a better accuracy level than the Jaccard Coefficient approach without indexing [16].

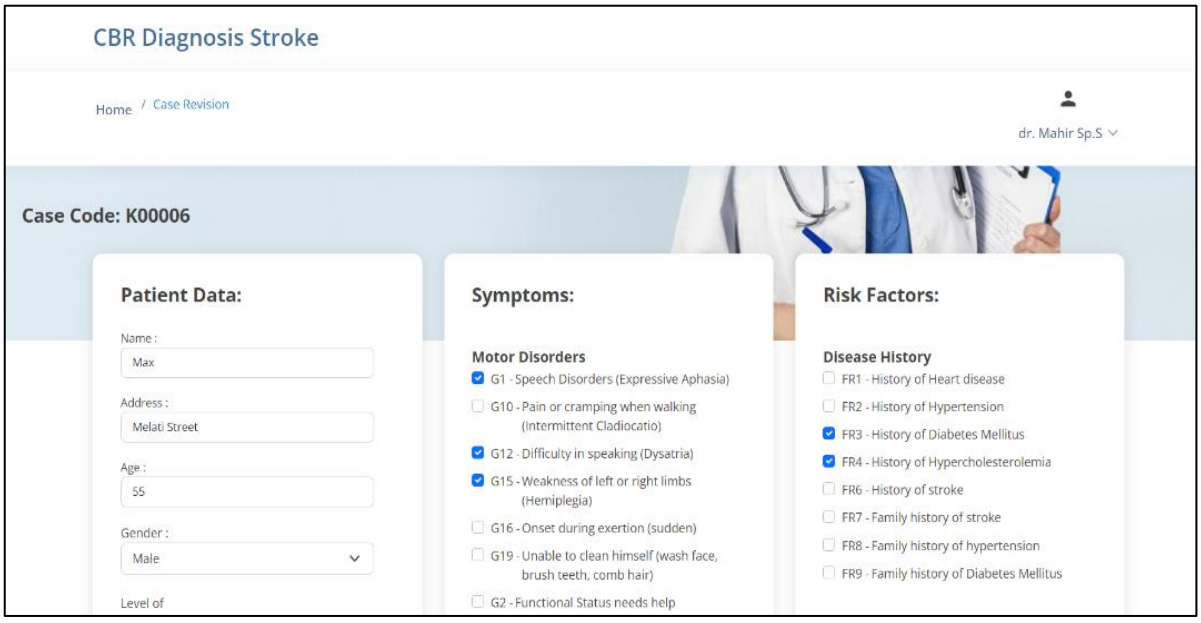


Fig. 4 Implementation of the revise process

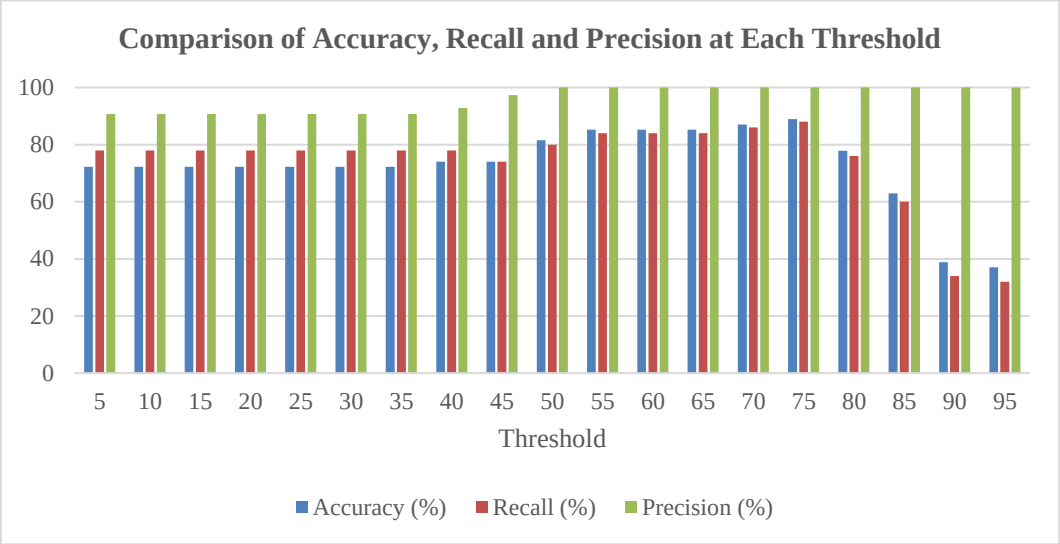


Fig. 5 Testing results graph

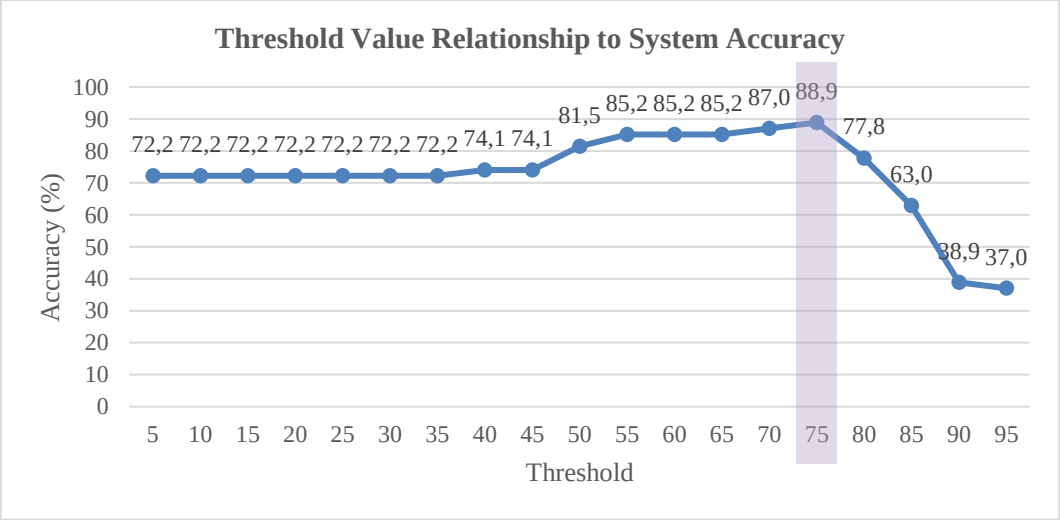


Fig. 6 System accuracy level graph using Minkowski distance similarity

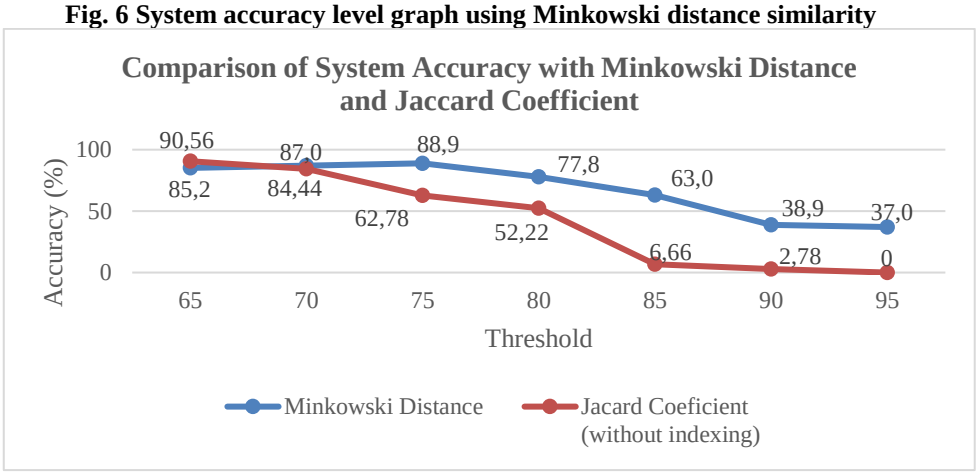


Fig. 7 Graph comparing the accuracy level of the stroke diagnosis system between Minkowski distance and Jaccard coefficient [16]

Accuracy does not have a preference for any one label but calculates all true prediction values. As a result, increased accuracy does not always imply that the system predicts labels accurately; hence, recall and precision values must be calculated. A recall measures the system's ability to retrieve information successfully, and the higher the recall value, the more effectively the system can recognize affirmative cases.

Applying a threshold value of 75 with values for r = 3 and r = 4 results in an 88% recall value for the system utilizing the Minkowski distance approach, as illustrated in Figure 8. As shown in Figure 9, the system's recall value when utilizing the Minkowski distance method is only 86.95%. However, the system's recall value when using the Jaccard Coefficient approach is lower. Positive prediction accuracy is measured by a statistic called precision; the higher the precision value, the more accurate the positive forecasts. As demonstrated in Figure 10, the Minkowski distance method system's best precision value is attained at a threshold of 50, achieving 100%.

According to the system's test results, a similarity level of 75 and a value of r = 3 or r = 4 are appropriate threshold values for the CBR system to identify stroke disease. Compared to various similarity threshold values, this selection yields the best accuracy and recall and reaches a sensitivity of 100%. Recall is a more significant evaluation parameter than precision in a system for identifying high-risk conditions (such as stroke). A poor recall value would indicate that several patients with stroke disease were misdiagnosed as healthy people, which might be extremely risky for the patients' lives. Therefore, a threshold value of 75 with a value of r = 3 or r = 4 results in a system accuracy of 88.89%, a recall of 88%, and a precision of 100% for the CBR system's diagnosis of stroke disease.

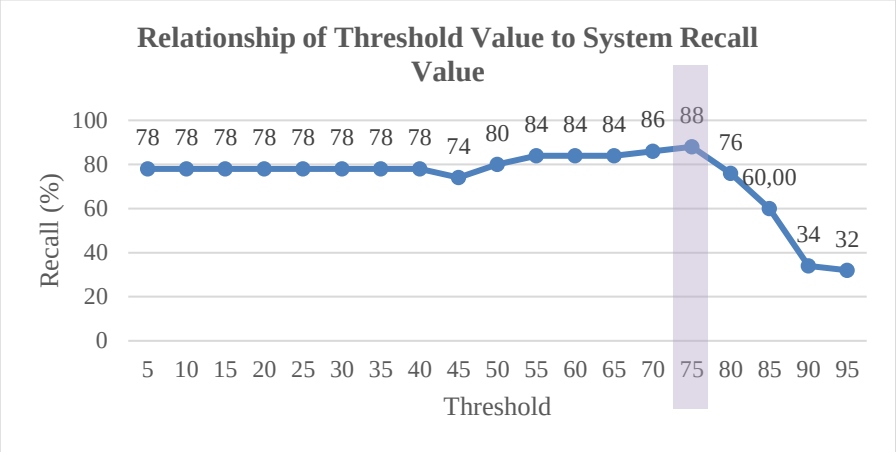


Fig. 8 System recall level graph using Minkowski distance

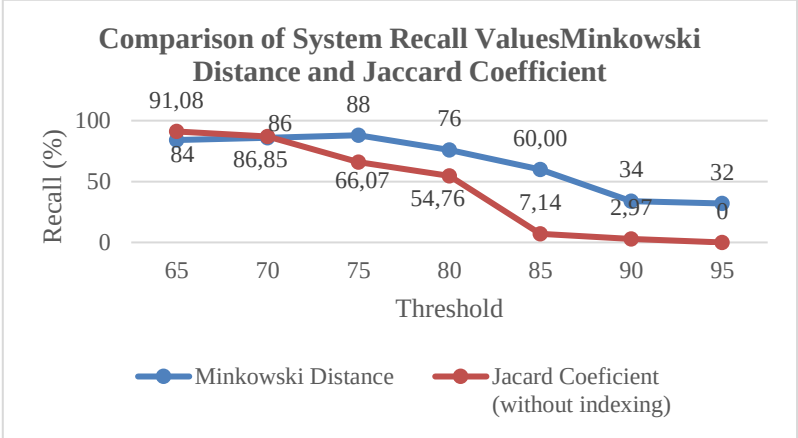


Fig. 9 Graph comparing the recall level of the system between Minkowski Distance and Jaccard Coefficient [16]

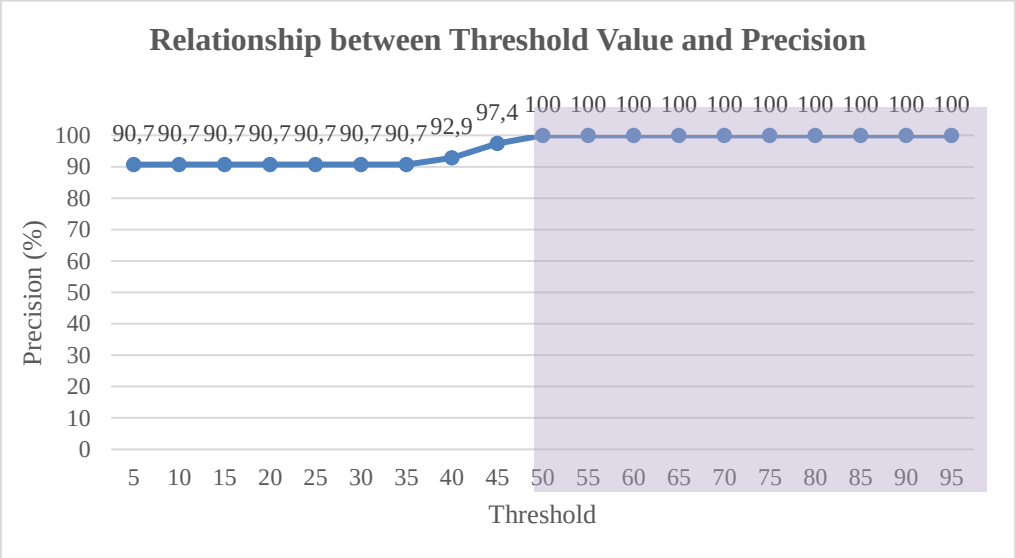


Fig. 10 System Precision Level Graph

IV. CONCLUSION

The Minkowski distance similarity method is utilized in this study to create a case-based reasoning (CBR) reasoning system for identifying stroke disease that may be used as a teaching aid for medical students. When using a threshold value of 75 with a value r = 3 or r = 4, the stroke disease diagnosis system developed in CBR that employs the Minkowski distance similarity method performs best in terms of accuracy rate (88.89%), recall rate (88%), and high precision (100%) rates. Comparing the Minkowski Distance Similarity approach to the Jaccard Coefficient method, the CBR system for identifying stroke disease offers superior accuracy and sensitivity/recall scores.

ACKNOWLEDGEMENT

Thank you to LPPM and the Faculty of Information Technology Universitas Kristen Duta Wacana, who have provided support and funds for the research and publication of this article.

REFERENCES

[1] Kementerian Kesehatan, "Apa itu Stroke," *http://p2ptm.kemkes.go.id/infographic-p2ptm/stroke/apa-itu-stroke*, 2018.

[2] P. J. F. Lucas and L. C. Van Der Gaag, *Principles of expert systems*. Singapore: Addison-Wesley Publishing, 1991. [Online]. Available: https://www.researchgate.net/publication/220694050

[3] R. Rizky, A. H. Wibowo, Z. Hakim, and L. Sujai, "Sistem Pakar Diagnosis Kerusakan Jaringan Local Area Network (LAN) Menggunakan Metode Forward Chaining," *J. Tek. Inform. UNIS*, vol. 7, no. 2, pp. 145–152, 2020, doi: 10.33592/jutis.v7i2.396.

[4] A. Abraham, "Rule-based Expert Systems," in *Handbook of Measuring System Design*, P. H. Sydenham and R. Thorn, Eds. John Wiley & Sons, Ltd., 2005, pp. 909–919.

[5] A. Gunawan, C. Suhery, and T. Rismawan, "Implementasi Metode Case-Based Reasoning Dan Similarity Jaccard Coefficient Dalam Identfikasi Kerusakan Laptop," *J. Komput. dan Apl.*, vol. 09, pp. 292–305, 2021.

[6] H. Henderi, F. Al Khudhorie, G. Maulani, S. Millah, and V. T. Devana, "A Proposed Model Expert System for Disease Diagnosis in Children to Make Decisions in First Aid," *INTENSIF J. Ilm. Penelit. dan Penerapan Teknol. Sist. Inf.*, vol. 6, no. 2, pp. 139–149, 2022, doi: 10.29407/intensif.v6i2.16912.

[7] B. Basiroh and S. W. Kareem, "Analysis of Expert System for Early Diagnosis of Disorders During Pregnancy Using the Forward Chaining Method," *Int. J. Artif. Intell. Res.*, vol. 5, no. 1, pp. 44–52, 2021, doi: 10.29099/ijair.v5i1.203.

[8] X. Huang *et al.*, "A Generic Knowledge Based Medical Diagnosis Expert System," *ACM Int. Conf. Proceeding Ser.*, vol. 1, no. 1, pp. 462–466, 2021, doi: 10.1145/3487664.3487728.

[9] I. Setiawan and M. Batara, "Expert System Design to Diagnose Pests and Diseases on Local Red Onion Palu Using Bayesian Method," *BAREKENG J. Math. Its Appl.*, vol. 17, no. 1, pp. 371–382, 2023.

[10] M. M. Richter and R. O. Weber, *Case-Based Reasoning*. New York: Springer Berlin Heidelberg, 2013. doi: 10.1007/978-3-642-40167-1.

[11] I. Nugraha and M. Siddik, "Penerapan Metode Case Based Reasoning (CBR) Dalam Sistem Pakar Untuk Menentukan Diagnosa Penyakit Pada Tanaman Hidroponik," *J. Mhs. Apl. Teknol. Komput. dan Inf.*, vol.

2, no. 2, pp. 91–96, 2020, [Online]. Available: https://www.ejournal.pelitaindonesia.ac.id/JMApTeKsi/index.php/JOM/article/view/575/387

[12] I. Y. Subbotin and M. G. Voskoglou, "Applications of fuzzy logic to Case-Based Reasoning," vol. 11, pp. 7–18, 2012, [Online]. Available: https://www.researchgate.net/publication/223129950

[13] A. S. Soroto, A. Fuad, S. Lutfi, J. J. Metro, and K. T. Selatan, "Penerapan Metode Case Based Reasoning (CBR) untuk Sistem Penentuan Status Gunung Gamalama," *J. Inform. dan Komput.*, vol. 02, no. 2, pp. 70–75, 2018.

[14] A. Yuli Vandika and A. Cucus, "Sistem Deteksi Awal Penyakit TBC dengan Metode CBR," *Pros. Semin. Nas. Darmajaya*, 2017.

[15] M. Nishom, "Perbandingan Akurasi Euclidean Distance, Minkowski Distance, dan Manhattan Distance pada Algoritma K-Means Clustering berbasis Chi-Square," *J. Inform. J. Pengemb. IT*, vol. 4, no. 1, pp. 20–24, 2019, doi: 10.30591/jpit.v4i1.1253.

[16] R. Nelson, A. Harjoko, and A. Musdholifah, "Case-Based Reasoning for Stroke Disease Diagnosis," *IJCCS (Indonesian J. Comput. Cybern. Syst.*, vol. 12, no. 1, p. 33, 2018, doi: 10.22146/ijccs.26331.

[17] M. Zainuddin, K. Hidjah, and W. Tunjung, "Penarapan Case-Based Reasoining (CBR) untuk Mendiagnosis Penyakit Stroke Menggunakan Algoritma K-Nearest Neighbor," *CITISEE*, 2016.

[18] I. Warman, "Sistem Pakar Identifikasi Penyakit Tanaman Padi Menggunakan Case-Based Reasoning," 2017.

[19] R. Adawiyah, "Implementasi Metode Minkowsky Distance untuk Deteksi Kelahiran Bayi Prematur Berbasis Case-Based Reasoning," *J. Inform. dan Komputer) Akreditasi KEMENRISTEKDIKTI*, vol. 3, no. 1, 2020, doi: 10.33387/jiko.

[20] A. Mubarak, M. Salmin, A. Fuad, and S. Do Abdullah, "Penalaran Berbasis Kasus Untuk Diagnosis Penyakit Malaria Dengan Menggunakan Metode Minkowsky Distance," *J. Ilm. Ilk. - Ilmu Komput. Inform.*, vol. 5, no. 1, 2022, doi: 10.47324/ilkominfo.v4i3.136.

[21] A. Labellapansa, A. Efendi, A. Yulianti, and A. K. Evizal, "Lambda value analysis on Weighted Minkowski distance model in CBR of Schizophrenia type diagnosis," in *2016 4th International Conference on Information and Communication Technology, ICoICT 2016*, 2016, vol. 4, pp. 1–4. doi: 10.1109/ICoICT.2016.7571898.

[22] E. Faizal and H. Hamdani, "Weighted Minkowski Similarity Method with CBR for Diagnosing Cardiovascular Disease," *Int. J. Adv. Comput. Sci. Appl.*, vol. 9, no. 12, 2018, doi: 10.14569/IJACSA.2018.091244.

[23] E. Wahyudi and N. I. Pradasari, "Case Based Reasoning untuk Diagnosis Penyakit Jantung Menggunakan Metode Minkowski Distance," *INTECOMS J. Inf. Technol. Comput. Sci.*, vol. 1, no. 1, pp. 119–123, 2018, doi: 10.31539/intecoms.v1i1.170.

[24] "Stroke Symptoms," *American Stroke Association*, 2021. https://www.stroke.org/en/about-stroke/stroke-symptoms

[25] M. K. Jha, D. Pakhira, and B. Chakraborty, "Diabetes Detection and Care Applying CBR Techniques ," *IJSCE*, vol. 2, no. 6, 2013.

[26] D. Normawati and S. A. Prayogi, "Implementasi Naïve Bayes Classifier Dan Confusion Matrix Pada Analisis Sentimen Berbasis Teks Pada Twitter," 2021.