Indonesian Journal of Artificial Intelligence and Data Mining (IJAIDM)
Vol 7, No.1, March 2024, pp. 80 – 85
p-ISSN: 2614-3372 | e-ISSN: 2614-6150

Early Prediction of Stroke Disease Diagnosis Patients Using Data Mining Algorithm Comparison

[1]Pungkas Subarkah, [2]Wenti Risma Damayanti, [3]Arbangi Puput Sabaniyah
[1,2]Departement of Informatics, Universitas Amikom Purwokerto
[3]Departement of Information System, Universitas Amikom Purwokerto
Email: [1]subarkah@amikompurwokerto.ac.id, [2]wrismadamayanti@gmail.com, [3]arbangipuputsabaniyah@gmail.com

Article Info
Article history:
Received Oct 9th, 2023
Revised Dec 27th, 2023
Accepted Jan 28th, 2024

Keyword:
Comparison Algorithm
Data Mining
Diagnosis
Early Prediction
Stroke

ABSTRACT
Stroke constitutes a medical emergency of paramount significance, characterized by a notably elevated mortality rate, and stands as the foremost cause of mortality within hospital settings. The dataset employed for this analysis is sourced from Kaggle, denoted as the Brain Stroke Dataset, encompassing a total of 4981 records. This research aims to carry out early prediction of stroke sufferers using several algorithms including the ANN algorithm, CART, KNN, and the NBC algorithm. The results obtained in the ANN algorithm obtained an accuracy of 93.53%, in the CART algorithm of 95.02%, in the KNN algorithm of 91.09% and in the NBC algorithm of 88.44%. With the outcomes of this research, the use of the cart set of rules may be used for early evaluation of stroke sufferers because it has a good degree of accuracy and is listed inside the excellent type kind..

Corresponding Author:
Pungkas Subarkah,
Departement of Informatics,
Amikom Purwokerto University,
Jl. Letjend Pol. Soemarto No.127, Watumas, Purwanegara, Kec. Purwokerto Utara,
Banyumas Regency, Central Java.
Email: subarkah@amikompurwokerto.ac.id
DOI: http://dx.doi.org/10.24014/ijaidm.v7i1.25955

1. INTRODUCTION

It is crucial to promptly identify the symptoms of a disease as the initial measure in anticipating the onset of a condition that can jeopardize one's health and, in severe cases, lead to fatality. Stroke is among the life-threatening ailments associated with this concern. Patients affected by stroke tend to experience cognitive impairment ranging from decreased awareness, visuospatial disorders, non-verbal learning disorders communication disorders, and decreased patient attention levels[1].

Stroke is the leading cause of death and disability worldwide[2]. Assessment is based on clinical features and brain imaging to distinguish between ischemic stroke and intracerebral hemorrhage [3], [4] Healthcare providers in Indonesia diagnosed 43.1% of stroke cases involving individuals over 75 years old and 0.2% of cases involving individuals between 15 and 24 years old [5]. Stroke is a sudden change in the brain, lasting more than 24 hours or causing death[6]. It affects the workings of the brain locally and globally stroke is a sudden change in the brain, lasting more than 24 hours or causing death [7]. It affects how the brain works locally and globally numbers [8].

Stroke is categorized as a less prevalent ailment, distinct from widespread diseases such as cardiovascular disease, cancer, diabetes, and chronic respiratory conditions, all of which contribute to mortality [9]. Stroke is a medical emergency because it has a greater mortality rate [10]. As per data released by the World Health Organization in 2016, cardiovascular disease accounted for 31% of global mortality. In 2017, stroke subsequently emerged as the third leading cause of death worldwide. Findings from the Lower Health Study, conducted by the Indonesian government in 2013, 2017, and 2018, revealed an escalation in the prevalence of this less common ailment [11], [12]. In Indonesia, stroke is the number one cause of death in hospitals, given the number of stroke sufferers and the significance of vital organs affected by stroke, predicting

stroke early is a priority for doctors[13]. Prediction of stroke-related problems in apps usually fails to achieve high accuracy [6], [8].

Certain studies have been conducted in connection with this research. Initially, a study was conducted focusing on breast cancer, employing the K-Nearest Neighbors (KNN) algorithm. The outcomes of this research reveal that an accuracy rate of 93% was achieved through the utilization of the KNN algorithm.[14]. Furthermore, research was conducted using the CART algorithm and the C 4.5 algorithm for stroke disease classification. The purpose of this research is to carry out prevention and early treatment to reduce mortality in stroke patients. The best accuracy result in the CART algorithm obtained is 95.11% [15]. Moreover, researchers conducted a study wherein they compared the Naive Bayes algorithm with the Random Forest algorithm and the Neural Network Algorithm. The objective of this study was to forecast heart disease using classification algorithms. In this investigation, the Naive Bayes algorithm yielded the highest accuracy, with a recorded value of 83%.[16]. Research was conducted by [17] regarding the comparison of PSO and GA-based C4.5 algorithms for diagnosing stroke disease. The results obtained are GA-based C4.5 optimization gets an accuracy value of 99.38% greater than PSO-based C4.5 0.10%.

This study aims to assess the early prediction of stroke disease by conducting a comparative analysis of the Artificial Neural Network (ANN) algorithm, Classification and Regression Trees (CART), K-Nearest Neighbor (KNN), and Naive Bayes Classifier (NBC) algorithm. The anticipated outcomes of this research endeavor are expected to furnish a valuable reference concerning the accuracy of stroke diagnosis when employing the the ANN algorithm, CART, KNN, and NBC algorithm. By predicting early stroke disease using a data mining approach, namely using the ANN, CART, KNN and NBC algorithms, it is expected to find the best pattern with the accuracy results obtained from each algorithm used.

Based on the problems described above, the authors propose to analyze the early prediction of stroke disease by comparing the methods of the ANN algorithm, CART, KNN, and NBC algorithm. The data used in this study are primary and secondary data obtained from Kaggle public data consisting of taken from the Kaggle Dataset consisting of 11 attributes which consist of 10 Feature attributes and 1 destination class attribute. Data obtained from the Kaggle Dataset site as many as 4981 records[18]

2. RESEARCH METHOD

The research methods used in this research are ANN algorithm, CART, KNN, and NBC algorithm.The following are the stages of the research, on figure 1.

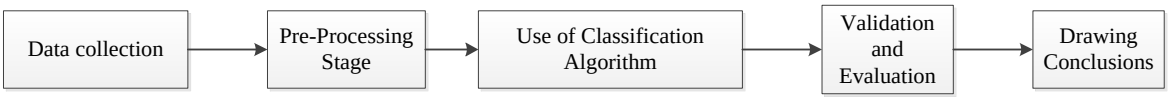

Figure 1. Research Flow

Figure 1 shows the research flow. The following is the result of the explanation:

2.1. Data Collection

This research, the secondary data used is taken from the Stroke Dataset from Kaggle, consisting of 11 attributes which consist of 10 Feature attributes and 1 destination class attribute. The data obtained from the Kaggle Dataset site is 4981 records. The dataset consists of 4981 records, which have 11 attributes (10 attributes and 1 target attribute). The target attribute consists of two outputs, namely the patient died due to stroke and the patient is being followed up by medical At this stage the member is in charge of downloading the dataset from the Kaggle site[18].

2.2. Pre-Processing stage

This stage involves a meticulous data selection process aimed at obtaining clean and research-ready data. The data pre-processing stage encompasses attribute identification and selection, addressing missing attribute values or incomplete attributes (handling missing values), and discretizing the value process.

2.3. Use of Classification Algorithm

The use of classification methods, the methods that researchers use are the ANN algorithm, CART, KNN, and NBC algorithm from the confusion matrix results can be calculated precision, recall and F-Measure values.

2.4. Validation and evaluation

This stage aims to measure the accuracy of the results achieved using the technical model used, namely confusion matrix and 10-Fold cross-validation.

2.5. Drawing conclusions

The final stage involves the culmination of research findings derived from the utilization of the ANN algorithm, CART, KNN, and NBC algorithm to determine the most accurate method for diagnosing diabetes retinopathy. This assessment relies on the precision, recall, and F-Measure metrics of each algorithm. The performance levels are categorized as follows: excellent classification = 0.90 - 1.00, good classification = 0.80 - 0.90, fair classification = 0.70 - 0.80, poor classification = 0.60 - 0.70, and failure = 0.50 - 0.60[19].

3. RESULTS AND ANALYSIS

3.1. Data Collection

Stroke Dataset from Kaggle, consisting of 11 attributes which consist of 10 Feature attributes and 1 destination class attribute. The data obtained from the Kaggle Dataset site is 4981 records. The dataset consists of 4981 records, which have 11 attributes (10 attributes and 1 target attribute).

Tabel 1. Data Stroke Dataset

No	Atrributes Name	Description
1	Gender	Female or Male
2	Age	Age of the patient
3	Hypertension	“0” if the patient doesn't have hypertension, “1” if the patient has
4	Heart Disease	“0” if the patient doesn't have any heart diseases, “1” if the patient has a heart disease
5	Ever Married	“No” or “Yes”
6	Work Type	"children", "Govtjov", "Never worked", "Private" or "Self-employed”
7	Residensce Type	“Rural” Or “Urban”
8	AVG Glucose Level	Average glucose level in blood
9	BMI	Body Mass Index
10	Smooking Status	“formerly smocked’, “never smocked”, “smokes” or “Unknown”
11	Class_Target	Yes or No

3.2. Pre-Processing Stage

The pre-processing phase comprises several steps, notably data transformation, which aims to enable the computation of values among class attributes during the classification stage. The outcomes of converting non-numeric data into numeric data are illustrated in Table 2.

Tabel 2. Data transformation of the target class stroke dataset

No	No-Numerical Data	Numerical Data
1	Patient had a stroke	“1”
2	Patient had not a stroke	“0”

Moreover, during this pre-processing stage, the activities of attribute identification, modification, and the selection of stroke datasets are undertaken to ensure that the obtained data is suitably prepared for utilization in the subsequent stage. The outcomes of the adjustments made to the stroke dataset's attributes for the Weka software are depicted in Table 3 below.

Table 3. Data Pre-processing

Original Data	Pre-Processing Data	Information
Male	Male	Gender
80	80	Age
0	0	Hypertension
1	1	Heart Disease
Yes	Yes	Ever Married
Private	Private	Work Type
Urban	Urban	Residensce Type
68.53	68.53	AVG Glucose Level
24.2	24.2	BMI
Smokes	Smokes	Smooking Status
1	1	Class_Target

3.3. Use of Classification Algorithm

Upon the conclusion of the pre-processing stage, the study advances to the classification phase. In this phase, our primary objective is to determine the accuracy value of the confusion matrix. This involves employing a testing methodology that incorporates the utilization of the ANN algorithm, CART, KNN, and NBC algorithm. The resulting accuracy metrics generated by these algorithms when applied to the stroke dataset are presented in Table 4.

Table 4. Accuracy Results of Stroke Dataset

No	Algorithm	Accuracy Result
1	ANN	93.53%
2	CART	95.02%
3	KNN	91.09%
4	NBC	88.44%

For enhanced clarity in observing the outcomes of the accuracy comparison among the ANN algorithm, CART, KNN, and NBC algorithm in relation to the stroke dataset, the visual representation of accuracy test results is presented in Figure 2.

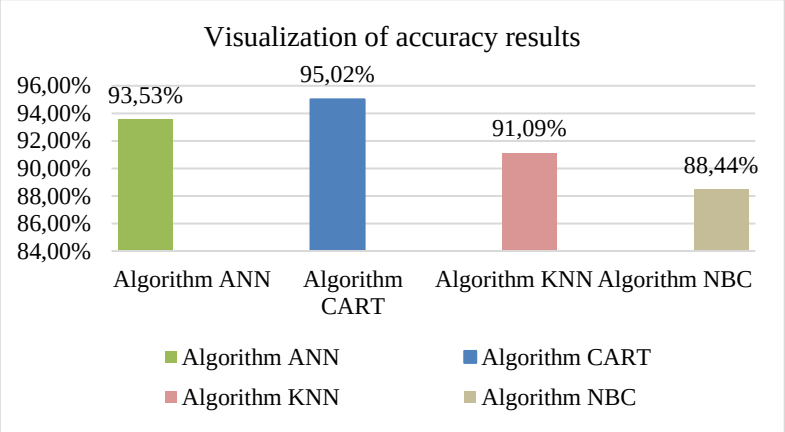

Figure 2. Visualization of accuracy results

Based on the Figure 5 provided above, the accuracy results for the stroke dataset using the ANN algorithm are 93.53%, while the CART algorithm attains an accuracy of 95.02%. Additionally, the accuracy outcomes for the KNN algorithm are 91.09%, and lastly, the NBC algorithm yields an accuracy of 88.44%. It is noteworthy that the highest accuracy result achieved during the algorithm testing is attributed to the Classification And Regression Trees (CART) algorithm, which attains a rate of 95.02%. This observation aligns with the fact that a substantial portion of the dataset employs numeric data types predominantly, contributing to the superior accuracy of this algorithm [20].

3.4. Validation and evaluation

Following this, in the validation and evaluation stage, which relates to the accuracy data presented in Table 6, we have the opportunity to scrutinize the accuracy results obtained from the confusion matrix. These results stem from the examination of the stroke dataset using ANN algorithm, CART, KNN, and NBC algorithm algorithms. These outcomes are depicted in Table 5, 6, 7, and 8.

Table 5. Confusion matrix ANN Algorithm

	0 (Patient had not a Stroke)	1 (Patient had a stroke)
0 (Patient had not a Stroke)	229	19
1 (Patient had a stroke)	93	4640
4981	322	4659

Table 6 Presented below are the outcomes of the confusion matrix of the CART algorithm.

Table 6. Confusion matrix CART Algorithm

	0 (Patient had not a Stroke)	1 (Patient had a stroke)
0 (Patient had not a Stroke)	248	0
1 (Patient had a stroke)	4733	0
4981	4981	0

Furthermore, Figure 8 Presented below are the outcomes of the confusion matrix of the KNN algorithm.

Table 7. Confusion matrix K-Nearest Neighbor (KNN) Algorithm

	0 (Patient had not a Stroke)	1 (Patient had a stroke)
0 (Patient had not a Stroke)	224	24
1 (Patient had a stroke)	4513	220
4981	4737	264

Finally, Table 8 Presented below are the outcomes of the confusion matrix of the Naive Bayes Classifier (NBC) algorithm.

Table 8. Confusion matrix Naive Bayes Classifier (NBC) Algorithm

	0 (Patient had not a Stroke)	1 (Patient had a stroke)
0 (Patient had not a Stroke)	164	84
1 (Patient had a stroke)	4321	412
4981	4485	496

4. CONCLUSION

The research conducted, comparing the accuracy of the ANN algorithm, CART, KNN, and NBC algorithms on early stroke disease datasets sourced from Kaggle, specifically the Brain Stroke Dataset, revealed that the CART algorithm achieved the highest accuracy among all algorithms. The CART algorithm yielded an accuracy rate of 95.02%, surpassing the performance of the ANN algorithm, KNN, and NBC. These findings underscore the utility of the CART algorithm for early stroke patient diagnosis, as it demonstrates a commendable level of accuracy and falls within the "Good Classification" category.

ACKNOWLEDGEMENTS

We would like to thank Amikom Purwokerto University through LPPM Amikom Purwokerto University with moral and material assistance so that this research runs smoothly and is completed.

REFERENCES

[1] M. Firdaus Banjar, Irawati, F. Umar, and L. N. Hayati, "Analysis of stroke classification using Random Forest method," vol. 14, no. 3, pp. 186–193, 2022.
[2] B. Imran, E. Wahyudi, A. Subki, S. Salman, and A. Yani, "Classification of stroke patients using data mining with adaboost, decision tree and random forest models," *Ilk. J. Ilm.*, vol. 14, no. 3, pp. 218–228, 2022.
[3] R. Vijayakumar *et al.*, "Prediction of protein aggregation on key proteins involved in ischemic stroke," *J. King Saud Univ. - Sci.*, vol. 35, no. 2, p. 102474, 2023.
[4] A. Byna and M. Basit, "Penerapan Metode Adaboost Untuk Mengoptimasi Prediksi Penyakit Stroke Dengan Algoritma Naïve Bayes," *J. Sisfokom (Sistem Inf. dan Komputer)*, vol. 9, no. 3, pp. 407–411, 2020.
[5] Y. Azhar, A. K. Firdausy, and P. J. Amelia, "Perbandingan Algoritma Klasifikasi Data Mining Untuk Prediksi Penyakit Stroke," *SINTECH (Science Inf. Technol. J.*, vol. 5, no. 2, pp. 191–197, 2022.
[6] M. Bahrudin, P. Yudha Pratama Putra, and D. Amalia Eka Putri, "Comparison of accuracy, sensitivity and specifity of Bahrudin score vs Siriraj score vs Gajah Mada algorithm in diagnosing type of stroke," *Brain Hemorrhages*, vol. 3, no. 4, pp. 184–188, 2022.
[7] Y. S. Huang *et al.*, "Exploring the pivotal variables of tongue diagnosis between patients with acute ischemic stroke and health participants," *J. Tradit. Complement. Med.*, vol. 12, no. 5, pp. 505–510, 2022.
[8] P. Kunwar and P. Choudhary, "A stacked ensemble model for automatic stroke prediction using only raw electrocardiogram," *Intell. Syst. with Appl.*, vol. 17, no. December 2022, p. 200165, 2023.
[9] WHO, *Noncommunicable Diseases Progress Monitor*. Switzerland, 2020.
[10] S. Tan *et al.*, "Delays in the diagnosis of ischaemic stroke presenting with persistent reduced level of consciousness : A systematic review," *J. Clin. Neurosci.*, vol. 115, no. March, pp. 14–19, 2023.
[11] Kementerian Kesehatan Republik Indonesia, *Profil Kesehatan Indonesia*. Jakarta: Kementerian Kesehatan RI, 2018.
[12] R.T. Pinzon, *Awas Stroke*. Yogyakarta: Betha Grafika, 2016.
[13] A. Ahmed *et al.*, "Stroke risk in older British men : Comparing performance of stroke-specific and composite-CVD risk prediction tools," *Prev. Med. Reports*, vol. 31, no. September 2022, p. 102098, 2023.
[14] D. Cahyanti, A. Rahmayani, and S. A. Husniar, "Analisis performa metode KNN pada Dataset pasien pengidap Kanker Payudara," *Indones. J. Data Sci.*, vol. 1, no. 2, pp. 39–43, 2020.
[15] Suryani *et al.*, "Analisis Perbandingan Algoritma C4. 5 dan CART Untuk Klasifikasi Penyakit Stroke: Comparative Analysis of C4. 5 and CART Algorithms for Classification of Stroke," *SENTIMAS Semin. Nas. Penelit. dan Pengabdi. Masy.*, vol. 1, no. 1, pp. 197–206, 2022.
[16] D. Derisma, "Perbandingan Kinerja Algoritma untuk Prediksi Penyakit Jantung dengan Teknik Data Mining," *J. Appl. Informatics Comput.*, vol. 4, no. 1, pp. 84–88, 2020.
[17] R. S. Rohman, R. A. Saputra, and D. A. Firmansaha, "Komparasi algoritma c4.5 berbasis pso dan ga untuk diagnosa penyakit stroke," vol. 5, no. 1, pp. 155–161, 2020.
[18] J. S. Tech, "Brain Stroke Dataset," *Agustus 2022*, 2022. [Online]. Available:

https://www.kaggle.com/datasets/jillanisofttech/brain-stroke-dataset. [Accessed: 01-Jan-2023].
[19] F. Gorunescu, *Data mining Concepts, Models and Techniques*. Verlen Berlin: Springer, 2011.
[20] P. Subarkah, M. M. Abdallah, and S. O. N. Hidayah, "Komparasi Akurasi Algoritme CART Dan Neural Network Untuk Diagnosis Penyakit Diabetes Retinopathy," *CogITo Smart J.*, vol. 7, no. 1, p. 121, 2021.

BIBLIOGRAPHY OF AUTHORS

Pungkas Subarkah is a lecturer at the Informatics Study Program, Amikom Purwokerto University. The author's education continued his S-1 Study at Amikom Purwokerto University and S-2 Master of Informatics Engineering at Amikom University Yogyakarta. Research interests in Data Mining, Machine Learning, and Information System.

Wenti Risma Damayanti is a 7th semester student majoring in Informatics, Faculty of Computer Science at Amikom Purwokerto University. Home address in Pandansari Village. Ajibarang District, Regency. Banyumas. Hobbies cooking

Arbangi Puput Sabaniyah, 3rd semester student majoring in Information Systems, Faculty of Computer Science, at Amikom Purwokerto University. Home address in Pekuncen Village, Jatilawang District, Banyumas Regency. Hobbies reading.