Expert System for Stroke Classification Using Naive Bayes Classifier and Certainty Factor as Diagnosis Supporting Device

Khusnul Ain[1], Hanik B. Hidayati[2] and Olivia Aulia Nastiti[1]

[1]Biomedical Engineering, Airlangga University, Indonesia
[2]Neurosurgery Department, Dr. Soetomo Hospital, Indonesia
corresponding author : k_ain@fst.unair.ac.id

Abstract. Stroke is a disease that has high mortality rate in Indonesia and worldwide. This disease is dangerous if not immediately treated because the brain circulatory disorder can cause a permanent disability or death. Diagnosis process of stroke disease can be assisted by an expert system using Naive Bayes Classifier and Certainty Factor. Decision making by Naive Bayes Classifier uses total probability value of all the diagnostic criteria on the existing database. Certainty Factor uses a weight value combination from measure of believe provided by expert. This expert system research aims to assist in the early diagnosis of a neurologist in diagnosing potential stroke patient from several diagnostic criteria as well as determine the accuracy of the expert system program created. The sample used in this study are the hospital medical records of 130 patients from Dr. Soetomo Hospital of Surabaya consists of 80 stroke patients data and 50 non-stroke patients data. Expert system program uses 12 category for diagnosis as an input and two statement outputs between stroke or non-stroke. Based on the analysis from the 25 testing data of 105 training data we obtained the accuracy of the method Naive Bayes classifier by 96% and of the method of Certainty Factor of 84%. Expert system program using Naive Bayes Classifier and Certainty Factor can be a device for stroke diagnosis.

keywords : expert system, diagnosis, naive bayes classifier, certainty factor

1. Introduction

Stroke is one of the third largest causes of death in the United States with an incidence of 130,000 cases / year [1]. Whereas in Indonesia stroke is a cardiovascular disease that most often occurs after coronary heart disease. Estimates of stroke patients in Indonesia are 1,236,825 people while symptoms are 2,137,941 people in 2013. This is believed to continue to increase until 2030.

World Health Organization defined stroke as a functional disorder of the brain that occurs suddenly with focal and global clinical signs and symptoms that last 24 hours or more. Stroke is a circulatory disorder in the brain as a cerebrovascular accident in the form of ischemia and bleeding. This circulatory disorder will cause brain function to be disrupted and if the case is severe it will result in infarction or death of some brain cells. The survivors of the stroke experience physical activity disorders. This is caused by the loss of some sensory and motor functions of the brain so that patients become less able to communicate, walk and numb [2].

Data obtained from the American Heart Association [3] suggests that stroke patients as much as 87% are ischemic stroke patients and the rest are hemorrhagic strokes. Data on stroke patients in the period October 2014- October 2015 at the Nerve Poly Outpatient Installation Dr. Soetomo stated that there were 4,238 people (82%) ischemic strokes, 895 people (17.4%) hemorrhagic strokes and the remaining 0.6% were strokes which could not be identified. The first two types of stroke cause reduced blood flow to the brain so that the body's oxygen needs are disrupted and can cause brain lesions that cannot recover [4].

Stroke is considered dangerous because it occurs suddenly and the symptoms appear suddenly. In diagnosing stroke, a neurologist processes 2 types of data, namely symptom analysis and neurological examination [2]. The diagnosis of stroke requires high accuracy because the identification process is quite difficult. The mechanism of diagnosis is only an estimate of the probability [2]. Therefore knowledge is needed to diagnose stroke. This knowledge is only possessed by a neurologist. The number of neuroscientists is not much, only about 900 neurologists in Indonesia. Stroke requires rapid diagnosis and management to reduce the risk of permanent disability.

The need for rapid diagnosis can be supported by information technology. Computer science as the root of information technology that underlies artificial intelligence that is able to represent data and manipulate it. Artificial Intelligence is a method that has the ability to express data, process data and solve problems. Searching for existing data then extracting it to build new information by doing four types of work, namely prediction models, cluster analysis, association analysis and detection anomalies. expert system based on certainty factors for early detection of red chili disease [5] and internal diseases [6] and naive bayes for detection of corn disease [7].

Several studies on artificial intelligence methods for the diagnosis of stroke have been carried out. one study was an Expert Diagnosis System for the Potential Stroke Attack using the Fuzzy Method. In this study it was mentioned that this expert system was able to process a number of input 9 symptom data using Fuzzy logic and Forward Chaining inference which then gave the output of stroke types and 3 categories of potential stroke attacks namely low, normal and high. Symptoms entered in the system are laboratory examination data such as blood pressure, blood sugar levels, total cholesterol levels, LDL cholesterol levels, uric acid levels, blood urea nitrogen and creatinine levels.

Another study to diagnose stroke is the Early Symptom Recognition Expert System of Stroke Using the Fuzzy Mamdani Method. This expert system uses the Fuzzy Mamdani method with a minimum rule for processing symptom data from 8 diseases that trigger stroke and then gives an output of 3 possible strokes, namely low, moderate and detected a stroke.

Another method of expert stroke diagnosis system is a Stroke Classification study based on Pathological Abnormalities with Learning Vector Quantization. This system can process input data as many as 32 data in the form of physical examination results, symptoms felt by patients, medical history and laboratory tests of patients' blood. The Learning Vector Quantization neural network method has a higher training speed compared to the Back propagation Neural Network so that this system can provide output in the form of classification of stroke suffered, namely ischemic stroke and hemorrhagic stroke with 96% accuracy.

Another Artificial Intelligence Method is an expert system with research topics Diagnosis of Patients Affected by Stroke by Using the Naive Bayes Method and Artificial Neural Network Methods. This system uses 2 methods, namely the single-layer neural network and the Naive Bayes method. Data processing is done on 13 symptoms to provide 2 kinds of system output, namely suspect or not suspect against stroke. The results given in this method are the accuracy of the Artificial Neural Network method at 99% and the Naive Bayes method at 97%.

Based on some of these studies, researchers will compare two reliable methods for the detection and classification of strokes, namely Naive Bayes Classifier and Certainty Factor. The results of this study are expected to provide a classification statement suffered by patients in the form of a statement of the presence or absence of a potential stroke from a number of diagnostic criteria data that have been accommodated previously in a database containing medical records of stroke and non-stroke patients.

2. Methods

The object of the study was 130 medical records of stroke and non-stroke patients from Dr. Soetomo Hospital Surabaya. Data collection was carried out with 2 methods, namely interviews with neurologist specialists and medical record analysis of stroke and non-stroke patients. The data obtained are selected based on diagnostic criteria which include weak facial muscles, speech disorders, weakness on one side of the body, sensory disorders, headache, sudden onset, history of hypertension, history of TIA (Transient Ischemic Attack), history of diabetes mellitus, consumption of cigarettes, and atrium fibrillation or a history of heart disease. The output is a diagnosis of stroke and non-stroke classes.

Table 1. Certainty Factor weight

	Symptom Features	CF Stroke	CF Non-Stroke
weak of facial muscle	No	0.2	0.8
	Yes	0.8	0.2
Speech disorder weak	No	0.2	0.8
	Yes	0.8	0.2
one side of the body	No	0.2	0.8
	Yes	0.8	0.2
sensory disorder	No	0.2	0.8
	Yes	0.8	0.2
Headache	No	0.2	0.8
	Yes	0.8	0.2
symptom	No	0.2	0.8
occur suddenly	Yes	0.8	0.2
age	< 55 years old	0.4	0.6
	≥ 55 years old	0.6	0.4
Waist size	Normal	0.3	0.7
	Obesity	0.7	0.3
history of	No	0.3	0.7
hypertension	Yes	0.7	0.3
history of TIA	No	0.3	0.7
	Yes	0.7	0.3
history of DM	No	0.3	0.7
	Yes	0.7	0.3
total cholesterol	No	0.4	0.6
	Yes	0.6	o.4
Uric Acid	No	0.4	0.6
	Yes	0.6	o.4
Smoke	No	0.4	0.6
	Yes	0.6	o.4
alcohol consumption	No	0.4	0.6
	Yes	0.6	o.4
history of heart disease	No	0.3	0.7
	Yes	0.7	0.3

Table 2. Interpretation of Certainty Factor

Certainty Factor	Uncertain Term
0.2	Small possibility
0.4	Maybe
0.5	Most likely
0.8	Almost certain
1.0	Certainly

There are two classification methods in the study, namely the Naive Bayes Classifier Method and the Certainty Factor Method, they are :

2.1. Naive Bayes Classifier Methods

a. Defines variables and classes

Bi is the result of the diagnosis class and Ai is a symptom that will get the probability of diagnosing a stroke or non-stroke.

b. Prior probability calculation

Calculation of prior probabilities for the possibility of a Bi class based on the equation:

$$P(B_i) = \frac{B_i}{B} \quad (1)$$

c. Calculation of posterior probability

Calculation of posterior probability uses the following equation:

$$\text{Probabilitas posterior} = \frac{\text{likekyhood clas prior}}{\text{predictor prior}} \quad (2)$$

The calculation of P (Ai | Bi) or likelihood is the probability of occurrence of each feature symptom that affects the emergence of the incidence of diagnosis with the equation (3),

$$P(A_i|B_i) = \frac{A_i}{B_i} \quad (3)$$

d. Calculation of diagnosis class events

$$P(A|B_i) = \prod_{k=1}^{n} P(A_k|B_i) \quad (4)$$

The description of each possible calculation of P (A | Bi) is as follows:

1) P(A | B_1) as a result of the possibility of a stroke diagnosis class

P(B1|A1,A2,A3,A4,A5,A6,A7,A8,A9,A10,A11,A12,A13,A14,A15,A16) =

P(A1)xP(A2)xP(A3)xP(A4)xP(A5)xP(A6)xP(A7)xP(A8)xP(A9)xP(A10)xP(A11)x P(A12)xP(A13)x P(A14)x P(A15)xP(A16)xP(B1)

2) P(A|B_2) as a result of the possibility of a non-stroke diagnosis class

P(B2|A1,A2,A3,A4,A5,A6,A7,A8,A9,A10,A11,A12,A13,A14,A15,A16) =

P(A1)xP(A2)xP(A3)xP(A4)xP(A5)xP(A6)xP(A7)xP(A8)xP(A9)xP(A10)xP(A11)x P(A12)xP(A13)x P(A14)x P(A15)xP(A16)xP(B2)

Furthermore, by following the rules of HMAP (Hypothesis Maximum Applicability Probability) to infer the final results of the diagnosis class decision. Therefore the calculation of P (A | Bi). P (Bi) used as a diagnosis decision is the calculation of the maximum value.

2.2. Certainty Factor Methods

Each input is given a quality weight according to the reference from the expert and then the calculation is done using the sequential combination equation of CF for each criterion according to the equation. (5),

$$CF_{combine} = CF_{old} + CF_{new} - (CF_{old} x CF_{new}) \quad (5)$$

Testing expert systems as a tool for diagnosing diseases is done matching the analysis of an expert. Correct data matching with classification inference results so that the level of accuracy of expert systems is obtained. The level of accuracy can be calculated by equation (6),

$$Accuration = \frac{\sum right\ data}{n} x100\% \quad (6)$$

where :

Σ right data = total data according to expert decisions
n = number of experiments conducted

3. Result and Discussion

Interview is information on criteria for stroke diagnosis. Medical records obtained 130 data, namely 80 data on stroke and 50 non-stroke patients. Data obtained from diagnostic criteria are identified according to the presence or absence of events based on the diagnosis. The database is 105 training data with details of 66 stroke patients and 39 non-stroke patients. This database is then connected with an expert system. The Certainty Factor weight given by experts is a value with a range of 0 to 1. By using equation (1) obtained prior probability of 0.628 for stroke diagnosis class and 0.371 for non-stroke. By using the equation (3) obtained by likelihood shown in table 3.

Table 3. Likelihood training data

	Feature	Class B1 (Stroke)		Class B2 (Non stroke)	
		Number	P (Ai | B1)	Number	P (Ai | B2)
Age (A1)	≥55 years old	48	0.727	29	0.743
	< 54 years old	18	0.272	10	0.256
Oblique face (A2)	Yes	36	0.545	2	0.051
	No	30	0.454	37	0.948
Speech disorder (A3)	Yes	45	0.681	1	0.025
	No	21	0.318	38	0.974
Hemiparese (A4)	Yes	52	0.787	9	0.23
	No	14	0.212	30	0.769
Hemi anestese (A5)	Yes	13	0.196	16	0.41
	No	53	0.803	23	0.589
Headache (A6)	Yes	22	0.333	26	0.666
	No	44	0.666	13	0,333
Sudden time (A7)	Yes	62	0.939	5	0,128
	No	4	0.06	34	0,871
Hypertension (A8)	Yes	62	0,939	28	0,717
	No	4	0,06	11	0,282
TIA (A9)	Yes	39	0,59	12	0,307
	No	27	0,409	27	0,692
DM (A10)	Yes	35	0,53	10	0,256
	No	31	0,469	29	0,743
Heart (A11)	Yes	23	0,348	8	0,205
	No	43	0,651	31	0,794
Smoke (A12)	Yes	9	0,136	4	0,102
	No	57	0,863	35	0,897

Both likelihood values are classified into the presence or absence of potential strokes using the Naive Bayes Classifier formulation. With the HMAP equation (Maximum Priori Probability Hypothesis), the expert system will make decisions based on the largest value of P (A|Bi). The results of the diagnosis decisions of the Naive Bayes Classifier method in 25 data are given in table 4.

Table 4. Expert System Decision Table Naive Bayes Classifier Method

Data	HMAP Stroke	HMAP Non-Stroke	Diagnosis Decision
1	0.00246943	$7.347196x10^{-7}$	Stroke
2	0.00026951	$7.335382x10^{-8}$	Stroke
3	0.00017251	0.0012478	Non-stroke
4	0.0001172	0.0029948	Non-stroke
5	0.006994	$2.2273x10^{-6}$	Stroke
6	0.0006858	$1.74393x10^{-6}$	Stroke
7	0.0001535	$1.92883x10^{-6}$	Stroke
8	0.0019878	$2.77045x10^{-5}$	Stroke
9	0.0006574	$4.58318x10^{-7}$	Stroke
10	0.0021824	$2.09656x10^{-7}$	Stroke
11	0.0001681	$2.27603x10^{-9}$	Stroke
12	0.0006968	$5.1868x10^{-8}$	Stroke
13	0.0038499	$1.32981x10^{-6}$	Stroke
14	$1.17566x10^{-5}$	0.000676	Non-stroke
15	$3.7498x10^{-7}$	0.005702	Non-stroke
16	$1.8561x10^{-6}$	0.004001	Non-stroke
17	$5.6987x10^{-7}$	0.003634	Non-stroke
18	$2.1851x10^{-5}$	0.02101	Non-stroke
19	$6.9237x10^{-7}$	0.00099	Non-stroke
20	$7.318x10^{-6}$	0.000123	Non-stroke
21	$5.5497x10^{-6}$	0.012474	Non-stroke
22	0.000136	0.002994	Non-stroke
23	0.000594	$4.2424x10^{-7}$	Stroke
24	0.002476	$1.1254x10^{-7}$	Stroke
25	$1.17566x10^{-5}$	0.000676	Non-stroke

The Certainty Factor method uses the equation (5) to obtain a combination of Certainty Factor. The CF weight of the diagnostic criteria for each test data was carried out in a CF combination in each diagnosis class to obtain CF combinations in the stroke and non stroke diagnosis classes. The expert system will make a diagnosis decision based on the greatest value among the two CF combinations. The results of the diagnosis decision of the Certainty Factor method in 25 data are given in table 5.

Table 5. Decision on Expert System Diagnosis The Certainty Factor method

Data	CF-stroke	CF-non-stroke	Diagnosis decision
1	0.999992012	0.999763965	Stroke
2	0.9999969	0.999393405	Stroke
3	0.999925673	0.999974727	Non-stroke
4	0.99982661	0.999989151	Non-stroke
5	0.999956667	0.999956667	Non-stroke
6	0.999956667	0.999956667	Non-stroke
7	0.999983787	0.999884366	Stroke
8	0.999989151	0.99982661	Stroke
9	0.999993026	0.999730229	Stroke
10	0.999537587	0.999995946	Stroke
11	0.999999225	0.997572302	Stroke

12	0.999999582	0.995684027	Stroke
13	0.999993681	0.999702751	Stroke
14	0.999393045	0.9999969	Non-stroke
15	0.999055922	0.999998033	Non-stroke
16	0.998381495	0.999998807	Non-stroke
17	0.994335293	0.999999642	Non-stroke
18	0.998381495	0.999998807	Non-stroke
19	0.9992661	0.99999151	Non-stroke
20	0.9992661	0.99999151	Non-stroke
21	0.999595403	0.99999535	Non-stroke
22	0.99982661	0.999989151	Non-stroke
23	0.999956667	0.999956667	Non-stroke
24	0.999997258	0.9990638	Stroke
25	0.999393045	0.9999969	Non-stroke

The accuracy of the Naive Bayes Classifier inference method gives a diagnosis of 96% while the Certainty Factor method is 84%. The Naive Bayes Classifier method is a fact-based decision-making system from statistical data collected in the field so that it is more stable. While the Certainty Factor method is a decision-making system based on the beliefs of an expert so that it is subjective. The weight of Certainty Factor given by experts depends on the quality of experts. The weight set used in this study is a fixed variable derived from the subjectivity of an expert based on the knowledge and experience of the expert. The accuracy of the Certainty Factor method with the weight set given in this study is lower than the Naive Bayes Classifier method.

4. Conclutions

1. The expert system of diagnosis of stroke potential classification by the Naive Bayes Classifier method has a prior probability of 0.628 for the stroke diagnosis class and 0.371 for the non-stroke diagnosis class. While the Certainty Factor method is made with a set of weight degrees of confidence whose value is given by experts and is subjective in nature.
2. The diagnosis of stroke potential with the Naive Bayes Classifier inference method produces an accuracy of 96% while the Certainty Factor inference method produces an accuracy of 84%.

References

[1] National Stroke Association. 2014. Stroke 101 : Fast Facts on Stroke. Colorado: National Stroke Association.
[2] Caplan, Louis R. 1993. *Stroke : A Clinical Approach 2nd Edition*. Massachussetts : Butterworth – Heinemann.
[3] American Heart Association. 2008. *Heart Disease and Stroke Statistics*. USA : American Stroke Association
[4] Bilotta, Kimberly A.J. 2009. Kapita Selekta Penyakit: Dengan Implikasi Keperawatan Edisi 2. Jakarta : EGC.
[5] Fahrul Agus, Hernandha Eka Wulandari, Indah Fitri Astuti, Journal of Applied Intelligent System, Vol.2, No. 2 (2017), pp. 52 – 66
[6] Munandar, Suherman and Sumiati, International Journal of Application or Innovation in Engineering & Management, Vol. 1, No. 1 (2012), pp.58-64
[7] Mohammad Syarief, Novi Prastiti, Wahyudi Setiawan , International Journal of Engineering Research and Application, Vol. 7, No. 11(2017), pp.30-34