LONTAR KOMPUTER VOL. 13, NO. 1 APRIL 2022 **p-ISSN 2088-1541**
DOI : 10.24843/LKJITI.2022.v13.i01.p03 **e-ISSN 2541-5832**
Accredited Sinta 2 by RISTEKDIKTI Decree No. 158/E/KPT/2021

Classification of Stroke Using K-Means and Deep Learning Methods

I Putu Kerta Yasa[a1], Ni Kadek Dwi Rusjayanthi[a2], Wan Siti Maisarah Binti Mohd Luthfi[b3]

[a]Information Technology, Faculty of Engineering, Udayana University
Bukit Jimbaran, Bali, Indonesia
[1]iputukerta@gmail.com
[2]dwi.rusjayanthi@unud.ac.id

[b]Data Science, School of Computing, Universiti Utara Malaysia
Bukit Kayu Hitam, Kedah, Malaysia
[3]wsmaisrh@gmail.com

Abstract

Stroke is a disease caused by blockage or rupture of blood vessels in the brain due to disruption of blood flow, where the blood supply to an area of the brain is suddenly interrupted. This study discusses stroke classification using the K-Means and Deep Learning methods. This study aims to segment patient data to produce patient class labels and classify the results of grouping the data to test the performance of the classification algorithm used. The 4,906 patient data used in this study were grouped using the K-Means method into multiple clusters, including 2 clusters, 3 clusters, 4 clusters, and 5 clusters, and the data grouping findings will be classified. The cluster validation method is the Davies Bouldin Index and the Silhouette Index, while the algorithm used in the classification process is the Deep Learning Algorithm. The classification results produce the most excellent accuracy value in the number of clusters tested, namely 2 clusters of 99.71%.

Keywords: *Stroke, K-Means, Davis-Bouldin Index, Silhouette Index, Deep Learning*

1. Introduction

Stroke is a disease that occurs due to disruption of blood flow due to blockage or rupture of blood vessels in the brain so that the blood supply to parts of the brain suddenly becomes disrupted. Brain nerve cells can be damaged due to a series of biochemical reactions caused by reduced blood flow in brain tissue. Functions controlled by brain tissue can decrease or even disappear due to the death of the tissue in that part of the brain. Part of the brain cannot function properly if the supply of oxygen and nutrients to the brain is stopped, causing blood flow to stop. Stroke is a disease with the second-largest contributor to death globally based on data from the World Health Organization (WHO) [1], reaching 6.7 million in 2012. As many as 69% of strokes occur in low-income, middle-income, and third-world countries. In 2018, the prevalence of stroke in Indonesia rose from 7% to 10.9%. The number of deaths caused by stroke ranks second for those aged over 60 years and fifth for those aged 15-59 years. A total of 57.9% of strokes have been diagnosed by health workers in Indonesia, where the prevalence of stroke based on symptoms is 12.1 per mile and based on diagnosis is 7.0 per mile. As the respondent's age increases, stroke also appears to increase, where the prevalence of stroke in men and women is the same [2].

Technology is used in various circles. Some activities have used technology to support various activities ranging from light to heavy activities. One of them is a technology used in the health sector. The technology needed in the form of computers and internet networks is the primary means of delivering information. Information is circulating more and more through computers. The impact caused by the rapid flow of information is the amount of data stored in the network. To take advantage of data that is spread a lot in the digital world, we require a tool to process data so that the information can be absorbed and presented as desired. One form of data processing is data mining. Finding patterns and knowledge from large amounts of data can be called data mining. Data sources are the basic things that must exist to carry out the data mining process [3].

In the health sector, data mining can be implemented into medical record data to predict disease. With the classification method in data mining, data such as age, gender, blood pressure, and other attributes can be used as a supporting factor in predicting the possibility of a patient getting a disease [4].

Irene Lishania and friends have researched the classification of stroke inpatient data by applying the Decision Tree Method and the Naive Bayes Method using the WEKA tools with a case study at the Abdul Wahab Sjahranie Hospital Samarinda. The results showed that the classification of stroke using the decision tree algorithm resulted in better classification performance. The classification accuracy results using the Naive Bayes method is 81.25%, and the decision tree algorithm obtained an accuracy rate of 87.5% [5]. Knowing the importance of data mining to analyze data to provide added value and new knowledge from large amounts of data faster than manually, a study was conducted with the title Classification of Stroke Using K-Means and Deep Learning Methods using Rapidminer tools version 9.9. The Deep Learning Algorithm method is used in this case study because Deep Learning performs well when the amount of data used continues to increase and how to solve a problem. Deep Learning Algorithm is used to create an artificial neural network that cannot process small amounts of data optimally. This is because the Deep Learning algorithm requires large amounts of data and can solve the problem as a whole from beginning to end without the need to separate it into several parts. This study aimed to determine the suitability of the algorithm used in dealing with grouping and classification problems using stroke patient data. In the future, the application of stroke classification data mining using a combination of clustering algorithms and deep learning classification algorithms can be implemented directly in the health sector in the form of information systems or stroke classification applications so that the implementation can support the decisions of health workers in diagnosing stroke and can provide treatment for indicated stroke patients and can reduce the risk of death from stroke by taking treatment early.

2. Research Methods

Patient data in this study used 4,906 records from the site kaggle.com [6]. The following is the data used in the process of classifying patient data.

Ro...	id	gender	age	hype...	heart...	ever...	work_type	Reside...	avg_gluc...	bmi	smoking_sta...
1	9046	Male	67	0	1	Yes	Private	Urban	228.690	36.6	formerly smok...
2	51676	Female	61	0	0	Yes	Self-employ...	Rural	202.210	N/A	never smoked
3	31112	Male	80	0	1	Yes	Private	Rural	105.920	32.5	never smoked
4	60182	Female	49	0	0	Yes	Private	Urban	171.230	34.4	smokes
5	1665	Female	79	1	0	Yes	Self-employ...	Rural	174.120	24	never smoked

Figure 1. Dataset

Figure 1 shows a stroke dataset containing 11 attributes; not all attributes are used to classify stroke. Therefore some attributes are not used. The attributes entered in the classification process consist of age, hypertension, heart disease, average glucose levels, and BMI. These attributes are used based on calculating the most influential information gain value. In this study, the author will group the datasets and identify them into several labels such as "Minor Ischemic Stroke," "Serious Ischemic Stroke," and "No Stroke." It is intended that the identified stroke is by the treatment that will be carried out on the seriousness of the stroke experienced.

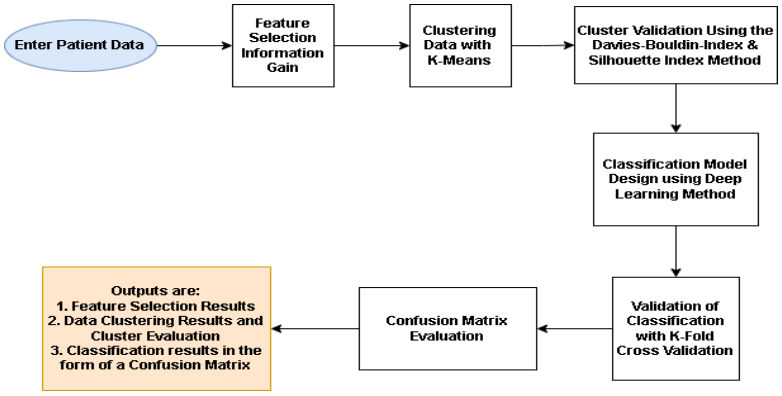

Figure 2. Stages of Research

Figure 2 is an overview of the system for classifying patient data. The stages of designing a patient data classification system begin with the data input process into a patient data classification system. The data used is stroke patient data downloaded through the kaggle.com site. The Feature Selection Process is carried out to select attributes that are considered relevant in the data mining process. The method used in this research is Information Gain. Information Gain is used to determine an important attribute's boundaries, where researchers widely use this feature selection method. The data clustering process is the process of grouping data based on the specified number of clusters. The purpose of this data clustering process is to determine the segmentation of patient data from each number of clusters tested. The data clustering process in this study uses the K-Means method. The results of data clustering will be evaluated using cluster validation. The cluster validation method used in this study is the Davies-Bouldin-Index (DBI) and the Silhouette Index (SI). The design of the classification model is carried out on the RapidMiner Studio Application using the Deep Learning Method. The validation process is carried out using k-fold cross-validation, where the number of folds (k-fold) tested is 10-fold. The classification evaluation process is carried out by comparing the results of the confusion matrix from each number of clusters that have been tested. The processing of patient data will produce output in the form of feature selection results, data clustering, cluster evaluation results, classification results in the confusion matrix, and results of deep learning algorithm models.

2.1. Information Gain

One of the important tasks of data pre-processing in data analytics is Feature Selection [7]. The advantage of the feature selection method is that it can better understand data in machine learning or a pattern recognition application, reduce computation time, and improve prediction performance [8]. The most widely used method by researchers in determining the boundaries of the importance of an attribute is the Information Gain feature selection method. The Information Gain feature selection method is mostly applied in Intrusion Detection System (IDS) research [7]. The entropy value before the separation is reduced by the entropy value after the separation will produce the Information Gain Value. Determination of attributes that will be used or discarded is done by measuring the Information Gain value as the initial stage. Attributes that will be used in the classification process of an algorithm are attributes that have met the weighting criteria. The following are three stages of feature selection using Information Gain.

a. Each attribute in the original dataset will be calculated for its Information Gain value.
b. After calculating the Information Gain value for each attribute, the next step is determining the desired threshold. This aims to see whether the weight of each attribute is equal to the limit or greater so that it can be maintained or discarded if the attribute Gain Information value is below a predetermined threshold.
c. By reducing unnecessary attributes, the dataset can be improved.

This attribute measurement was first pioneered by Claude Shannon in Information Theory and is written as:

$$info\ (D) = -\sum_{j=1}^{v} \frac{|Dj|}{|D|} x\ I(D_j) \quad (1)$$

Information:
D: Set of cases
A: Attributes
v: Number of partition attribute A
$|D_j|$: Number of cases on j partition
|D|: Number of cases in D
$I(D_j)$: Total entropy in partition

Meanwhile, to find the Information Gain attribute A, the following formula can be used:

$$Gain\ (A) = I(D) - I(A) \quad (2)$$

Information:
Gain (A): Information attribute A
I (D): Total entropy
I (A): entropy A

The result of the expected entropy reduction caused by the introduction of the attribute value A results in the value of Gain (A). The largest Information Gain value is selected as a test of the attribute set S for each attribute. After that, a node will be created and labeled accordingly to that attribute, then a branch is created for each of the other attribute values.

2.2. Normalization

According to Madhiarasan, Data transformation is carried out to convert data into values that are easy to understand. One of the data transformation techniques is normalization or data normalization. The normalization method based on the data's average value (mean) and standard deviation can be called Z-score normalization. The Z-score normalization method improves the model's accuracy[8]. This method is beneficial when the actual minimum and maximum data value is unknown. The equation used in Z-score normalization is as follows.

$$New\ value = \frac{Old\ value - Mean}{Stdev} \quad (3)$$

Information:
New Value: Normalized data value
Old value: Data value before normalization
Mean: The average value of the data values per column
Stdev: Value of standard deviation

2.3. *Clustering*

This paper uses the K-Means algorithm to group data. The K-Means belong to unsupervised learning that functions to the group in pattern recognition and machine learning. The algorithm is always influenced by initialization with several clusters required a priori [9]. The initialization of the cluster center of the K-Means algorithm is very sensitive because it is done randomly. The cluster center of the K-Means algorithm is determined from its mean value. The stages of the K-Means algorithm can be seen as follows.

a. The initial cluster center is determined by randomly selecting k values.
b. The value of k will divide each data into several parts as much as k, and Euclidean Distance is used to determine the center of the cluster.

$$d_{euc} = \sum_{i=0}^{n} \sqrt{(z_i - X_i)^2} \quad (4)$$

c. Each cluster centroid is calculated as the average of the clusters received.

d. Choice of steps 2 and 3 if there is a change in the subgroup. If no changes are made to the cluster, the process is aborted.

2.4. Cluster Validation with Davies-Bouldin Index (DBI) and Silhouette Index (SI)

The validity of the cluster that has been formed will be tested using the Davies-Bouldin Index (DBI) method. Validation calculations are performed with the help of operators in RapidMiner. The following are the steps for calculating cluster validation using the Davies-Bouldin Index Method.

a. Calculate the variance of each cluster using the following equation.

$$S = \frac{\sum n-1 \, (\mathrm{X}i-x)^2}{n-1} \tag{5}$$

n is the number of data, X is the square of the centroid, and x is the centroid.

b. Calculate the distance between the centers using the formula Euclidean Distance.
c. Find the DBI value with the following equation.

$$DB = \frac{Rij1+Rij2+\cdots Rij}{n} \tag{6}$$

Rij is the average value of the cluster using the following equation:

$$Rij = \frac{var\,(Ci)+var\,(Cj)}{|\,ci-cj\,|} \tag{7}$$

var (C) is the cluster variance value, and C is the cluster average value.

Clusters that have been formed will also be tested for validity using the Silhouette Index Method to see a comparison with the Davies-Bouldin-Index Method. Silhouette Index validation calculation is done manually. The following are the steps for calculating cluster validation using the Silhouette Index Method.

a. Calculate the average of object i so that the average value of a(i) is obtained.
b. Calculate the average of object i in other clusters by taking the smallest value of the average value to obtain the average value of b(i).
c. Calculation of all variables using the following equation.

$$S(i) = \frac{\mathrm{b}(i)-\mathrm{a}(i)}{\max\{\mathrm{a}(i),\mathrm{b}(i)\}} \tag{8}$$

S(i) is the overall average calculation value, a(i) is the average core point distance with all points in the same cluster, and b(i) is the average core point distance value with all points in different clusters.

d. Calculate the average silhouette value with the following equation.

$$GSu = \frac{1}{n}\sum\frac{n}{j=1} \tag{9}$$

GSu is the Silhouette mean value, and n is the number of data.

e. The cluster with the highest GSu value is the optimal cluster.

2.5. Data Modelling

The cluster that has been formed will then go through a data modeling process. The data modeling stage aims to determine each cluster's class by finding each cluster's average value, which is then compared with the range of values in domain value [11]. Data modeling uses attributes that have been selected or previously selected using the Information Gain method, namely age, hypertension, heart disease, avg. glucose levels, and BMI. Below is Table 1, which shows the Linguistic variables and domain for each mean value.

Table 1. Linguistic Variables and Domain Values of Attributes

Attribute	Linguistic Variable	Value Random
Age	Mature	*Age* ≤ 45
	Seniors	46 ≤ a*ge*
Hypertension	Low Risk of Suffering	0 ≤ *Hypertension* < 0,1
	High Risk of Suffering	0,1 ≤ *Hypertension*
Heart Disease	Low Risk of Suffering	0 ≤ *Heart Disease* < 0,1
	High Risk of Suffering	0,1 ≤ *Heart Disease*
Avg. Glucose Level	Normal	0 ≤ *Avg. Glucose Level* ≤ 140
	Diabetes	141 ≤ *Avg. Glucose Level*
BMI	Mild Excess Weight	25,1 ≤ *BMI* < 30,0
	Overweight Look at Weight	30,1 ≤ *BMI*

The class for each cluster can be determined by comparing the mean value of each attribute of the cluster with the predefined domain values. Each model class has a patient label that states the characteristics of each patient class. Below is Table 2, which shows the class descriptions for each cluster.

Table 2. Description of Linguistic Variables from Patient Label

Description of Linguistic Variable					Patient Label
Age	*Hyper tension*	*Heart Disease*	*Avg. Glucose Level*	*BMI*	
Mature	Low Risk of Suffering	Low Risk of Suffering	Normal	Mild Excess Weight	No Stroke
Mature	High Risk of Suffering	Low Risk of Suffering	Normal	Overweight Look at Weight	Minor Ischemic Stroke
Mature	High Risk of Suffering	High Risk of Suffering	Diabetes	Overweight Look at Weight	Serious Ischemic Stroke
Seniors	Low Risk of Suffering	Low Risk of Suffering	Normal	Mild Excess Weight	No Stroke
Seniors	High Risk of Suffering	Low Risk of Suffering	Normal	Overweight Look at Weight	Minor Ischemic Stroke
Seniors	High Risk of Suffering	High Risk of Suffering	Diabetes	Overweight Look at Weight	Serious Ischemic Stroke

Table 2 is a patient label linguistic variable that shows the category of stroke patients. This linguistic variable labeling stroke patient is based on the rules of stroke risk factors, where the older the patient, the greater the risk of stroke, and the more symptoms or disease the patient suffers, the worse the condition of stroke patients.

2.6. Classification

This paper uses a deep learning method to classify the result of clustering data grouping. Deep learning is a method that utilizes a multi-layered Artificial Neural Network. Artificial Neural Networks are made similar to the human brain, where a complex network of neurons is formed from the connection of neurons. The deep learning method is a learning method that utilizes multiple non-linear transformations. Addressing significant problems in statistical machine learning is required in Deep Learning Method [12].

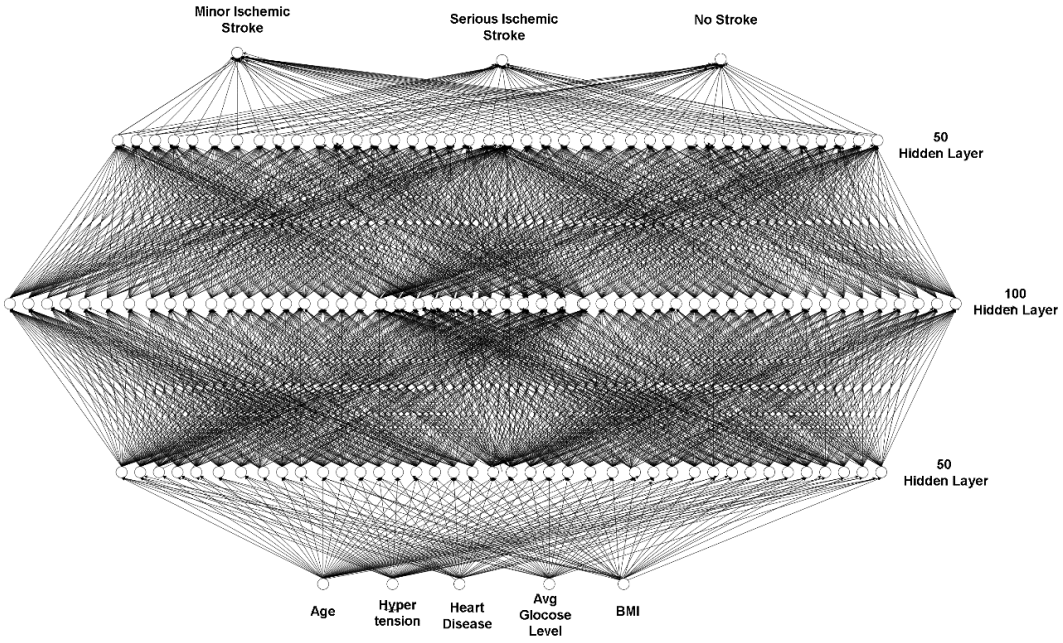

Figure 3. Deep Learning Architecture

Figure 3 shows the architecture of Deep Learning, which has three different layer functions: input, The depth of the model is indicated by the length of the chain as a whole, where the process can be referred to as Deep Learning [13]. The basis of the Deep Learning Model is determined from cross multiplication Feedforward neural networks or Multilayer Perceptrons (MLPs), where data is entered, the machine will pass through two or more layers to perform processing [14]. The accuracy level obtained will increase when more layers are used [15]. The attributes entered in the classification process consist of age, hypertension, heart disease, average glucose level, and BMI. These attributes are used based on calculating the most influential information gain value. The determination of the parameters used in this algorithm experiment are activation used is Rectifier, three hidden layers (with each size of 50 units, 100 units, 50 units), local random seed (active) of 1992, epochs is 10, the train samples per iteration used is -2, the epsilon is 1.0E-8, rho of 0.99, L1 of 1.0E-5, L2 of 1.0E-5, and max w2 of 10.

3. Result and Discussion

Clustering was tested with the K-Means method to form 2 until 5 clusters. The results of the experiments can be seen as follows.

Table 3. The Results of Segmentation (2 Clusters)

Clusters	Patient Value		Linguistic Variable	Patient Label
1	85,81%	*Age*	40 Mature	No Stroke
		Hypertension	0,068 Low Risk of Suffering	
		Heart Disease	0,035 Low Risk of Suffering	
		Avg. Glucose Level	89,497 Normal	
		BMI	28 Mild Excess Weight	
2	14,19%	*Age*	57 Seniors	Serious Ischemic Stroke
		Hypertension	0,236 High Risk of Suffering	
		Heart Disease	0,136 High Risk of Suffering	

Clusters	Patient Value		Linguistic Variable	Patient Label
		Avg. Glucose Level	200,944 Diabetes	
		BMI	32 Overweight Look at Weight	

Table 3 results in grouping 2 clusters using the K-Means Cluster Method. The segmentation results of 2 clusters produce two patient labels, No Stroke and Serious Ischemic Stroke.

Table 4. The Results of Segmentation (3 Clusters)

Clusters	Patient Value		Linguistic Variable	Patient Label
1	46,72%	*Age*	57 Seniors	Minor Ischemic Stroke
		Hypertension	0,116 High Risk of Suffering	
		Heart Disease	0,066 Low Risk of Suffering	
		Avg. Glucose Level	89,759 Normal	
		BMI	30,5 Overweight Look at Weight	
2	13,68%	*Age*	58 Seniors	Serious Ischemic Stroke
		Hypertension	0,244 High Risk of Suffering	
		Heart Disease	0,133 High Risk of Suffering	
		Avg. Glucose Level	202,872 Diabetes	
		BMI	33 Overweight Look at Weight	
3	39,60%	*Age*	20 Mature	No Stroke
		Hypertension	0,010 Low Risk of Suffering	
		Heart Disease	0,001 Low Risk of Suffering	
		Avg. Glucose Level	91,136 Normal	
		BMI	26 Mild Excess Weight	

Table 4 is the result of grouping 3 clusters using the K-Means Cluster Method. The segmentation results of forming 3 clusters produce three patient labels: Minor Ischemic Stroke, Serious Ischemic Stroke, and No Stroke.

Table 5. The Results of Segmentation (4 Clusters)

Clusters	Patient Value		Linguistic Variable	Patient Label
1	11,58%	*Age*	61 Seniors	Serious Ischemic Stroke
		Hypertension	0,266 High Risk of Suffering	
		Heart Disease	0,153 High Risk of Suffering	
		Avg. Glucose Level	211,151 Diabetes	
		BMI	33,4 Overweight Look at Weight	
2	32,18%	*Age*	20 Mature	No Stroke
		Hypertension	0,010 Low Risk of Suffering	

Clusters	Patient Value		Linguistic Variable	Patient Label
		Heart Disease	0,001 Low Risk of Suffering	
		Avg. Glucose Level	81,993 Normal	
		BMI	26 Mild Excess Weight	
3	19,16%	*Age*	38 Mature	No Stroke
		Hypertension	0,060 Low Risk of Suffering	
		Heart Disease	0,030 Low Risk of Suffering	
		Avg. Glucose Level	124,485 Normal	
		BMI	28 Mild Excess Weight	
4	37,08%	*Age*	59 Seniors	Minor Ischemic Stroke
		Hypertension	0,126 High Risk of Suffering	
		Heart Disease	0,070 Low Risk of Suffering	
		Avg. Glucose Level	82,585 Normal	
		BMI	30,3 Overweight Look at Weight	

Table 5 is the result of grouping 4 clusters using the K-Means Cluster Method. The segmentation results of 4 clusters produce three patient labels: Minor Ischemic Stroke, Serious Ischemic Stroke, and No Stroke.

Table 6. The Results of Segmentation (5 Clusters)

Clusters	Patient Value		Linguistic Variable	Patient Label
1	26,80%	*Age*	58 Seniors	Minor Ischemic Stroke
		Hypertension	0,121 High Risk of Suffering	
		Heart Disease	0,063 Low Risk of Suffering	
		Avg. Glucose Level	75,532 Normal	
		BMI	30,4 Overweight Look at Weight	
2	11,75%	*Age*	60 Seniors	Serious Ischemic Stroke
		Hypertension	0,264 High Risk of Suffering	
		Heart Disease	0,151 High Risk of Suffering	
		Avg. Glucose Level	210,590 Diabetes	
		BMI	33,3 Overweight Look at Weight	
3	20,03%	*Age*	57 Seniors	Minor Ischemic Stroke
		Hypertension	0,114 High Risk of Suffering	
		Heart Disease	0,073 Low Risk of Suffering	
		Avg. Glucose Level	108,295 Normal	
		BMI	30,3	

Clusters	Patient Value		Linguistic Variable	Patient Label
			Overweight Look at Weight	
4	12,27%	*Age*	22 Mature	No Stroke
		Hypertension	0,020 Low Risk of Suffering	
		Heart Disease	0,002 Low Risk of Suffering	
		Avg. Glucose Level	125,590 Normal	
		BMI	26 Mild Excess Weight	
5	29,15%	*Age*	20 Mature	No Stroke
		Hypertension	0,010 Low Risk of Suffering	
		Heart Disease	0,001 Low Risk of Suffering	
		Avg. Glucose Level	80,214 Normal	
		BMI	26 Mild Excess Weight	

Table 6 results from grouping 5 clusters using the K-Means Cluster Method. The segmentation results of forming 5 clusters produce three patient labels: Minor Ischemic Stroke, Serious Ischemic Stroke, and No Stroke.

In the DBI validity index, the number of clusters with the lowest DBI value indicates that the number of clusters is optimal [16], while in the SI validity index, the number of clusters that have the highest Silhouette index value is the optimal number of clusters [17]. The following table 7 shows the result of the DBI value and the SI for each number of clusters that have been tested.

Table 7. Results of Cluster Validation

Number of Clusters	DBI Value	Silhouette Index Value
2 Clusters	-0,529	0,617
3 Clusters	-0,933	0,408
4 Clusters	-0,831	0,338
5 Clusters	-0,891	0,347

Based on testing using the two methods of validity index above for each number of clusters, it is obtained that the optimal number of clusters is 2 clusters with the lowest DBI value and the highest SI value index.

Table 8 shows the results of testing data grouping with a deep learning algorithm using 10-Fold Cross Validation [18].

Table 8. Result of Classification Using Deep Learning Method

	2 *Cluster*	**3 *Cluster***	**4 *cluster***	**5 *cluster***
Accuracy Value	99,71%	99,02%	98,57%	98,17%
MSE	9,433001E-4	0,005762797	0,0083509665	0,009102712
RMSE	0,030713191	0,07591309	0,09138362	0,095408134
R^2	0,9922516	0,9876279	0,99225473	0,9951603
logloss	0,003705171	0,019558994	0,029368008	0,03301215
mean_per_class_error	0,0014367816	0,006768372	0,011219117	0,009040273

Based on the tests to determine the performance of the deep learning algorithm in classifying stroke diagnosis with each number of clusters using the Rapidminer 9.9 application, 2 clusters have the highest accuracy value of 99.71%. Comparison of the results of deep learning algorithm models based on MSE (Mean Square Error), RMSE (Root Mean Square Error), R^2, logloss, and

mean_per_class_error in 2 clusters, 3 clusters, 4 clusters, and 5 clusters. MSE (Mean Square Error) is the average squared error class between the actual value and the predicted or forecasted value. A low MSE value or a mean squared error value close to zero indicates that the prediction results are by the actual data and can be used for prediction calculations in the future period. The mean squared error method is usually used to evaluate measurement methods with predictive models. RMSE (Root Mean Square Error) is a measurement method by measuring the difference in the value of a model's prediction as an estimate of the observed value. RMSE is the product of the square root of MSE. The accuracy of the measurement error estimation method is indicated by the presence of a small RMSE value. The estimation method that has a smaller RMSE (Root Mean Square Error) is said to be more actual than the estimation method that has a larger RMSE. R Squared is a number that ranges from 0 to 1, indicating the magnitude of the combination of independent variables that jointly affect the value of the independent variable, where the closer to number one, the model issued by the regression will be better. Cross-entropy loss (log loss) is used to calculate the error in the model. Smaller log loss values represent a better model than larger ones. A model that predicts probability perfectly has a cross-entropy or log loss of 0.0.

4. Conclusion

In this study, the implementation of the K-Means and Deep Learning can be applied in the application of stroke classification through the stages of data selection, feature selection using the Information Gain, the clustering process using the K-Means, cluster validation using the Davies Bouldin Index and the Silhouette Index, the process classification using Deep Learning Method, and validation and evaluation process using k-fold cross validation and confusion matrix. The feature selection results using the Information Gain Method produce the five most influential attributes in the classification process. Based on the results of cluster validation measurements using DBI validation, 2 clusters has the lowest value of 0.529 among the number of other clusters, so it can be said to be the optimum cluster. Based on the measurement of cluster validity using the Silhouette Index Method, the number of 2 clusters tested had cluster quality with a good structure with a SI value of 0.617. The classification using the Deep Learning Method yielded the most excellent accuracy value in the number of clusters tested, namely 2 clusters of 99.71%.

References

[1] G. Sailasya and G. L. A. Kumari, "Analyzing the Performance of Stroke Prediction using ML Classification Algorithms," *International Journal of Advanced Computer Science and Applications*, vol. 12, no. 6, 2021, DOI: 10.14569/IJACSA.2021.0120662.

[2] Riskesdas, “Rikesdas 2013,” *Riset Kesehatan Dasar*, 2013.

[3] J. Han and M. Kamber, "Data Mining: Concepts and Techniques (Solution Manual)," *San Francisco, CA, itd: Morgan Kaufmann*, 2006.

[4] M. Mirqotussa’adah, M. A. Muslim, E. Sugiharti, B. Prasetiyo, and S. Alimah, “Penerapan Dizcretization dan Teknik Bagging Untuk Meningkatkan Akurasi Klasifikasi Berbasis Ensemble pada Algoritma C4.5 dalam Mendiagnosa Diabetes,” *Lontar Komputer*, vol. 8, no. 2, p. 135, Aug. 2017, DOI: 10.24843/LKJITI.2017.v08.i02.p07.

[5] I. Lishania, R. Goejantoro, and N. Nasution, “Perbandingan Klasifikasi Metode Naive Bayes dan Metode Decision Tree Algoritma (J48) pada Pasien Penderita Penyakit Stroke di RSUD Abdul Wahab Sjahranie Samarinda,” *Jurnal EKSPONENSIAL*, vol. 10, no. 2, pp. 135–142, 2019.

[6] FEDESORIANO, "Stroke Prediction Dataset," 2020. [Online]. Available: https://www.kaggle.com/datasets/fedesoriano/stroke-prediction-dataset.

[7] Kurniabudi, D. Stiawan, Darmawijoyo, M. Y. bin bin Idris, A. M. Bamhdi, and R. Budiarto, "CICIDS-2017 Dataset Feature Analysis with Information Gain for Anomaly Detection," *IEEE Access*, vol. 8, pp. 132911–132921, 2020, DOI: 10.1109/ACCESS.2020.3009843.

[8] G. Chandrashekar and F. Sahin, "A survey on feature selection methods," *Computers and Electrical Engineering*, vol. 40, no. 1, pp. 16–28, Jan. 2014, DOI: 10.1016/j.compeleceng.2013.11.024.

[9] D. A. Anggoro and W. Supriyanti, "Improving accuracy Bb applying Z-score normalization in linear regression and polynomial regression model for real estate data," *International*

Journal of Emerging Trends in Engineering Research, vol. 7, no. 11, pp. 549–555, Nov. 2019, DOI: 10.30534/ijeter/2019/247112019.

[10] K. P. Sinaga and M. S. Yang, "Unsupervised K-means clustering algorithm," *IEEE Access*, vol. 8, pp. 80716–80727, 2020, DOI: 10.1109/ACCESS.2020.2988796.

[11] R. W. Sembiring Brahmana, F. A. Mohammed, and K. Chairuang, "Customer Segmentation Based on RFM Model Using K-Means, K-Medoids, and DBSCAN Methods," *Lontar Komputer*, vol. 11, no. 1, p. 32, Apr. 2020, DOI: 10.24843/LKJITI.2020.v11.i01.p04.

[12] Z. Q. Zhao, P. Zheng, S. T. Xu, and X. Wu, "Object Detection with Deep Learning: A Review," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 30, no. 11, pp. 3212–3232, 2019, DOI: 10.1109/TNNLS.2018.2876865.

[13] J. Heaton, "Ian Goodfellow, Yoshua Bengio, and Aaron Courville: Deep learning," *Genetic Programming and Evolvable Machines*, vol. 19, no. 1–2, 2018, DOI: 10.1007/s10710-017-9314-z.

[14] C. L. Zhang and J. Wu, "Improving CNN linear layers with power mean non-linearity," *Pattern Recognition*, vol. 89, pp. 12–21, 2019, DOI: 10.1016/j.patcog.2018.12.029.

[15] D. A. Prasetya and I. Mujahidin, "2.4 GHz Double Loop Antenna with Hybrid Branch-Line 90-Degree Coupler for Widespread Wireless Sensor," 2020. DOI: 10.1109/EECCIS49483.2020.9263477.

[16] B. Jumadi Dehotman Sitompul, O. Salim Sitompul, and P. Sihombing, "Enhancement Clustering Evaluation Result of Davies-Bouldin Index with Determining Initial Centroid of K-Means Algorithm," *Journal of Physics: Conference Series*, vol. 1235, no. 1, pp. 1–6, 2019, DOI: 10.1088/1742-6596/1235/1/012015.

[17] A. A. R. Fernandes, Solimun, Nurjannah, U. A. I. Billah, and N. M. A. A. Badung, "Comparison of Cluster Validity Index and Distance Measures Using Integrated Cluster Analysis and Structural Equation Modeling the Warp-PLS Approach," *Journal of Southwest Jiaotong University*, vol. 56, no. 3, pp. 157–168, Jun. 2021, DOI: 10.35741/issn.0258-2724.56.3.13.

[18] H. Yasar and M. Ceylan, "A novel comparative study for detection of Covid-19 on CT lung images using texture analysis, machine learning, and deep learning methods," *Multimedia Tools and Applications*, vol. 80, no. 4, 2021, DOI: 10.1007/s11042-020-09894-3.