Jurnal Ilmiah Teknik Elektro Komputer dan Informatika (JITEKI)
Vol. 7, No. 3, December 2021, pp. 423-432
ISSN: 2338-3070, DOI: 10.26555/jiteki.v7i3.22237

Linkage Detection of Features that Cause Stroke using Feyn Qlattice Machine Learning Model

Purwono [1], Alfian Ma'arif [2], Iis Setiawan Mangku Negara [3], Wahyu Rahmaniar [4], Jihad Rahmawan [5]
[1,3] Universitas Harapan Bangsa, Jl. Raden Patah No. 100 Kedunglongsir Ledug Kembaran, Banyumas 53182, Indonesia
[2] Universitas Ahmad Dahlan, Jln. Ring Road Selatan, Bantul, Yogyakarta, Indonesia
[4] National Taipei University of Technology, Taiwan
[5] Iwate Prefecture University, Japan

ARTICLE INFO

Article history:

Received November 13, 2021
Revised December 14, 2021
Accepted December 21, 2021

Keywords:

Stroke;
Machine Learning;
Qlattice;
Predictor;
Ehr

ABSTRACT

Stroke is a disease caused by brain tissue damage because of blockage in the cerebrovascular system that disrupts body sensory and motoric systems Stroke disease is one of the highest death cause in the world. Data collection from Electronic Health Records (EHR) is increasing and has been included in the health service big data. It can be processed and analyzed using machine learning to determine the risk group of stroke disease. Machine learning can be used as a predictor of stroke causes, while the predictor clarifies the influence of each cause factor of the disease. Our contribution in this research is to evaluate Feyn Qlattice machine learning models to detect the influence of stroke disease's main cause features. We attempt to obtain a correlation between features of the stroke disease, especially on the gender as a feature, whether any other features can influence the gender feature. This research utilizes 4908 data of the disease predictor using the Feyn Qlattice model. The result implies that gender highly impacts age and hypertension on stroke disease causes. Autorun in Feyn Qlattice model was run with ten epochs, resulting in 17596 test models at 57s. Query string parameter that was focused on age and hypertension features resulted in 1245 models at 4s. An increase of accuracy was found in training metrics from 0.723 to 0.732 and in testing metrics from 0.695 to 0.708. Evaluation results showed that the model is reasonably good as a predictor of stroke disease, indicated with blue lines of AUC in training and testing metrics close to ROC's left side peak curve.

Corresponding Author:

Purwono, Universitas Harapan Bangsa, Jl. Raden Patah No. 100 Kedunglongsir Ledug Kembaran, Banyumas 53182, Indonesia
Email: purwono@uhb.ac.id

1. INTRODUCTION

Stroke is a disease caused by brain tissue damage because of blockage in the cerebrovascular system [1] that disrupts body sensory and motoric systems [2]. This condition causes all body functions controlled by brain tissue to be disrupted. Stroke is a very dangerous disease and must be treated immediately because brain cells can die in minutes. Proper treatment must be done to prevent complications. Stroke has become one of the highest death cause diseases in the world [3]. Many low-income countries are unable to cope with the burden posed by this disease. Moreover, Indonesia placed first in the highest death cases caused by stroke disease with 193,3/100.000 cases per year [4]. Some cause factors of stroke disease are hypertension, obesity, smoke, cholesterol increase, physical activity, low-density lipoprotein increase, excessive alcohol consumption, and diabetes [5].

The utilization of Electronic Health Records (EHR) by many countries worldwide is rapidly increasing [6]. Many medical data resulting from EHR has been collected and included in big data of health and medical service [7][8]. The analysis of medical data is required to determine risk group factors of many diseases [9].

The collected data can be reprocessed using machine learning models to find various new patterns that can benefit as actionable knowledge and information [10]. One of the benefits of using machine learning is that it can be used to predict several factors that may cause stroke [11][12]. This predictor clarifies the influence of each factor causing this disease. This predictor clarifies the influence of each factor causing this disease. For example, we can investigate whether there is an effect of age and hypertension on someone's susceptibility to stroke.

Nowadays, many classifiers from machine learning models have been used in some researches, especially on stroke disease. Research conducted by Liu [1] used a machine learning model called random forest in classifying cause factors of stroke disease, resulting in 85.03% accuracy. Another research was conducted by Zhu [13] identified stroke ischemic onset time based on DWI and FLAIR imaging with Convolutional Neural Network (CNN) model, yielding an accuracy of 80.50%. Meanwhile, Jamthikar [14] used machine learning, a random forest model, to prevent stroke by integrating carotid ultrasound image-based phenotypes and their harmonics with conventional risk factors, yielding an accuracy of 93.15%.

Many machine learning models are used to predict stroke diseases, such as SVM, XGBoost, Logistic Regression, KNN, Random Forest, Decision Tree, and others. However, currently, there are not many studies that apply machine models to investigate correlation or linkage between primary cause features of stroke diseases. Therefore, we propose an alternative machine learning model, called the Feyn Qlattice model, to assess the influence of each cause feature of stroke disease. This model was developed by a startup named Abzu, which was inspired by Richard Feynman's path integral formula [15]. Compared with neural networks and decision trees, Feyn Qlattice has some superiorities. Feyn Qlattice eliminates the black box concept that can be found in neural networks, though it provides explanations similar to the decision tree model. Feyn Qlattice works by searching thousands of potential models and seeking the best feature to become the ideal machine learning model to solve a computation problem [16].

Our contribution in this study is to analyze and evaluate the Feyn Qlattice machine learning model to detect the influence of the main causative features of stroke. The analysis was carried out to obtain the correlation between the features of stroke, especially on gender as a feature, whether there are other features that can affect gender features. By applying Feyn Qlattice, thousands of training models can be obtained so that the best machine learning model can be selected and used to predict the main causes of stroke. The results of the analysis can be in the form of data which is the result of the model evaluation of each predictor feature used.

2. METHOD

This section describes the proposed framework for using the Feyn Qlattice model to predict the association of features that influences the causation of stroke disease. Several important steps are described in each subset. Overall, the methodology used in the research can be seen in Fig. 1.

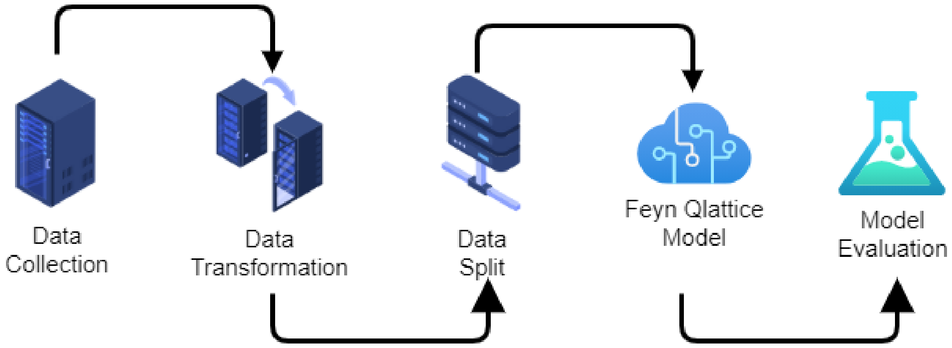


Fig. 1. Methodology

Based on Fig 1, the first is data collection, i.e., datasets related to the causal factors of stroke disease. Then, data transformation is carried out to balance data so that the machine learning model can work effectively. In the next step, the data is separated into training data and testing data. This technique is called splitting. Then, the Feyn Qlattice model was applied to produce thousands of best machine learning models that could predict the main features that cause a stroke. The Features that caused stroke were then selected to be tested with the Feyn Qlattice model as well. The model with the best performance results will be selected and evaluated. A more detailed explanation will be explained in the following subsections.

2.1. Dataset

Dataset used in this research was taken from public datasets made by Fedesoriano, which was uploaded in Kaggle [17]. This dataset was formatted in Coma Separated Value (CSV) with 5110 rows of data. It still has many noise or false-formatted data. For example, there was empty-valued or non-uniform data. The dataset has 12 main features that can be used to predict the cause of stroke disease. Available features of the dataset are id, gender, age, hypertension, heart_disease, ever_married, work_type, residence_type, avg_glucose_level, bmi, smoking_status, and stroke. The stroke feature becomes the classifier used from the dataset. Table 1 shows the dataset sample and its format used in this research.

Table 1. Healthcare Stroke Data

id	Feature Data										
	gender	**age**	**hypertension**	**heart disease**	**ever married**	**work type**	**residence type**	**Avg glucose level**	**BMI**	**smoking status**	**stroke**
9046	Male	67	No	Yes	Yes	Private	Urban	228.69	36.6	formerly	1
51676	Female	61	No	No	Yes	Self	Rural	202.21	N/A	never	1
31112	Male	80	No	Yes	Yes	Self	Rural	105.92	32.5	never	1
60182	Female	49	No	No	Yes	Self	Urban	171.23	34.4	smokes	1
1665	Female	79	Yes	No	Yes	Self	Rural	174.12	24	never	1
…..	…..	…..	….	…..	…..	…..	…..	…..	…..	…..	…..
44679	Female	44	0	0	Yes	Govt	Urban	85.28	26.2	Unknown	0

2.2. Preprocessing Dataset

The dataset that has been collected cannot be immediately used because it is imbalanced, so preprocessing is needed. This step balances the dataset by adding the sample from a smaller dataset or subtracting samples from a bigger dataset [18]. Preprocessing is essential to improve data quality so that machine learning can function properly [19]. An unprocessed dataset is usually ambiguous and incomplete because some of its attributes are missing, either in its inputs or outputs, which may negatively affect the machine learning modeling [20]. Moreover, Qlattice models immediately detect data types; incorrect detection of data types leads to poor machine learning models. Qlattice supports many variants of data transformations, such as linear, multiply, sine, tan, and gaussian transformation [16].

Data features that are ambiguous, such as columns with similar features, will be collided or selected so that only one column will remain [21]. Empty-valued features will also be deleted in preprocessing. Data consistency is carefully maintained. This can be seen in bmi feature; N/A values were found decimals where decimals are a majority in this feature. Therefore, the feature will be uniformly adjusted.

Properties with important characteristics and categorical behavior will be changed to number categories; this technique is called categorical variable encoding [22]. Values from each categorical feature will be changed into a number. For example, gender feature has “male” dan “female” as its category. The value of “male” will be identified as “1,” and “female” will be identified as “0”. A detailed change of categorical encoding can be seen in Table 2.

Table 2. Categorical Variable Encoding

Feature	Feature Data	
	Data	**Category**
Gender	Male	1
	Female	0
Ever married	Yes	1
	No	0
Work Type	Private	0
	Self	1
	Govt Job	2
Residence Type	Urban	1
	Rural	0
Smoking Status	Formerly	0
	Smokes	1
	Never	2

2.3. Data Splitting and Data Balancing

Data that has been through the preprocessing step will have better quality and become ready to be used in machine learning. Data will be divided into 75% composition of training data and 25% of testing data. The splitting of the data must be done effectively to improve the model's accuracy [23][24]. An illustration of the splitting can be seen in Fig. 2.

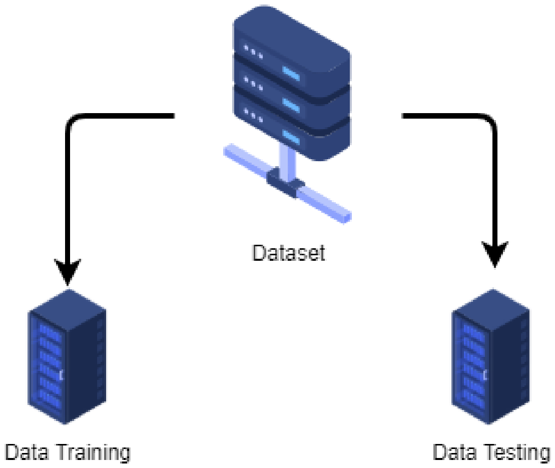


Fig 2. Data Splitting

2.4. Feyn Qlattice Model

Feyn Qlattice model was used in this research. Data that has been split into 75% training data and 25% testing data will be processed with this model. Some steps used in the Feyn Qlattice model can be seen in Fig. 3.

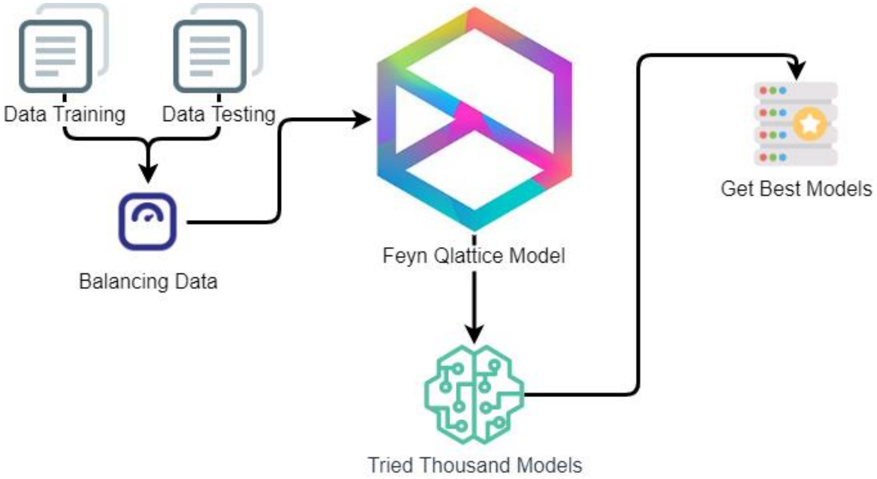


Fig 3. Feyn Qlattice Model

Based on Fig. 3, a dataset that has been split will be reprocessed with a technique called sample weight computation to balance the data. Imbalanced data usually create problems in machine learning [25]. Only balanced data will be connected with the Feyn Qlattice. The Qlattice model uses training data that has been separated in data splitting to fill its train parameter. Another parameter of the model is the output name; since the purpose of the model is to predict stroke disease, the parameter is given “stroke” as its value. The kind parameter of the model is filled with “classification” since the dataset type is classification data. Meanwhile, the stypes parameter is filled with “gender” since the influence of the gender feature on other features that cause stroke disease will be assessed using this model.

Autorun process in Qlattice takes all parameters that have been set. This process will result in thousands of models that will be tested in 10 epochs in a certain time duration. Epoch is a hyperparameter that determines how many times the machine learning model will process the training data [26]. This research used a 10/10 scale epoch; 10 experiments were done to obtain the best model and feature prediction. After processing stages were done to the dataset, machine learning will result in the best models and feature predictions. Methodological steps to get the best model and predictor feature can be seen in Fig. 4.

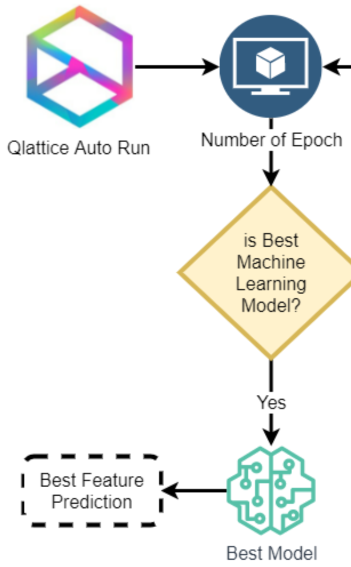
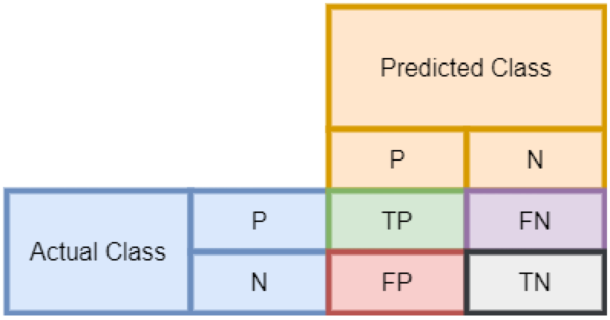


Fig 4. Qlattice AutoRun to Get Best Model and Best Feature Prediction

2.5. Evaluation Model

The best model obtained is evaluated by an evaluate machine model evaluation learning model called confusion matrix. This method can be used to measure the model's performance to various classification problems in machine learning [27]. The confusion matrix creates a representation of results such as true positive (TP), true negative (TN), false positive (FP) dan false negative (FN) [27]. TP means the positive results that are predicted by machine learning are correct. TN means the negative results predicted by machine learning are correct. Meanwhile, FP means the positive results predicted by machine learning are wrong, and FN means the negative results predicted by machine learning are wrong. Fig. 5 illustrates the confusion matrix table.



		Predicted Class	
		P	N
Actual Class	P	TP	FN
	N	FP	TN

Fig 5. Confusion Matrix

Performance evaluation with confusion matrix results in accuracy, precision, and recall [28][29]. Accuracy is the number of data points that machine learning predicted correctly among all data points. It can be calculated as follows:

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN} \tag{1}$$

Precision is a percentage of relevant elements that can tell how many times the model can predict correctly. It can be calculated as

$$Precission = \frac{TP}{TP + FP} \tag{2}$$

Meanwhile, recall is a percentage of relevant elements correctly classified by the machine learning model over the whole relevant elements. The calculation of recall can be carried out using

$$Recall = \frac{TP}{TP + FN} \tag{3}$$

Along with the confusion matrix, we used Receiver Operating Characteristic (ROC), which is a visual technique to assess and choose a suitable classifier based on its performances [30]. ROC can also be considered as a performance measurement of a classification-type machine learning model [31]. It is common to compute Area Under the ROC Curve (AUC), a recognized metric to evaluate and compare classification models [30]. AUC can be equivalent to the probability that a randomly selected positive sample will have a higher value than a negative sample [30]. As the ROC curve gets closer to the top left corner of the graph, the model can classify better [32].

3. RESULTS AND DISCUSSION

3.1. Data Transformation After Preprocessing

The data generated after the preprocessing stage will undergo data transformation, decreasing the number of data lines and each feature value with categorical data. After the data transformation, the previous sum of data, which is 5110, was reduced to 4908. The results of the data transformation can be seen in Table 3.

Table 3. Data Transformation

id	Feature Data										
	gender	age	hypertension	heart disease	ever married	work type	residence type	Avg glucose level	BMI	smoking status	stroke
9046	1	67	0	1	1	0	1	228.69	36.6	1	1
31112	1	80	0	0	1	1	0	105.92	32.5	2	1
60182	0	49	0	0	1	1	1	171.23	34.4	0	1
1665	0	79	1	0	1	1	0	174.12	24	2	1
…..	…..	…..	….	…..	…..	…..	…..	…..	…..	…..	…..
44679	0	44	0	0	1	2	1	85.28	26.2	0	0

As seen in Table 3, a change has been made to categorical features based on the categorical variable encoding. The variables previously had string as their type of data, while currently, their values are changed into encoding lists represented by integers such as 0, 1, or 2. However, special features such as avg glucose level and BMI still use decimals since their values vary or have uncategorical characteristics. The total number of transformed data is now 4908, composed of 75% of training data and 25% of testing data. Hence, the total number of training data is 3681, and the total number of testing data is 1227.

3.2. Model Feyn Qlattice

Based on stages in using the Feyn Qlattice model as in Fig. 4, as many as 17596 models were resulted and will be tested after the autorun mode was done in 10 epochs. This model with stypes input as 'gender' results in the best predictor features at 57s: hypertension and age. The autorun process of the Feyn Qlattice model can be seen in Fig. 6.

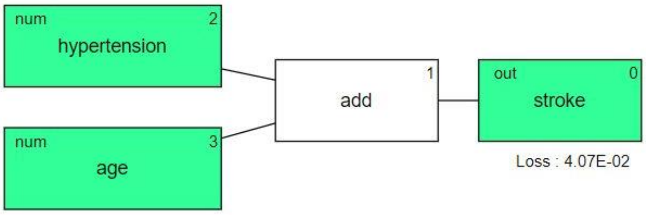


Fig 6. AutoRun Feyn Qlattice

After 17956 models were obtained, the best model, the top first or first ordered model, is chosen. The best model is visualized with a plot graph that results in the training metrics and testing metrics. Training metrics result in 0.723 of accuracy, 0.851 AUC, 0.116 precision, and 0.828 recall. Meanwhile, the testing metrics result in 0.695 of accuracy, 0.818 of AUC, 0.103 of precision, and 0.808 of recall.

According to Fig. 6, it is most likely that age and hypertension are predictors of stroke cause disease when viewed based on the age feature. The next test is to narrow the feature based on additional query string parameters, which contain age and hypertension as the parameter values. The plot graph of the model that resulted from the addition of query string parameters can be seen in Fig. 7.

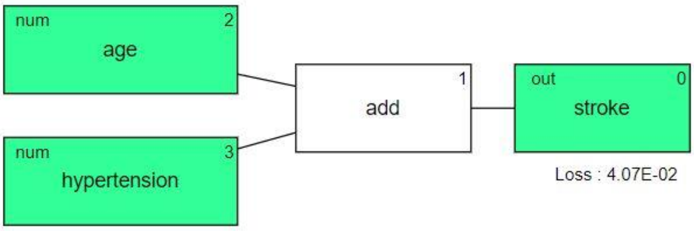


Fig 7. Addition of Query Strings: Age and Hypertension

After the query strings named age and hypertension were added, the autorun was re-run for ten epochs and resulted in 1245 machine learning models in 4s. The best model will then be visualized with a plot graph that results in training and testing metrics. Training metrics result in 0.731 of accuracy, 0.851 AUC, 0.117 precision, and 0.809 recall. Meanwhile, testing metrics result in 0.708 of accuracy, 0.818 AUC, 0.106 precision, and 0.788 of recall. An increase in accuracy can be seen based on the results. In training metrics, the accuracy was increased from 0.723 to 0.731. However, an increase of accuracy from 0.695 to 0.708 was found in testing metrics.

3.3. Evaluasi Model

The evaluation of the Feyn Qlattice model was done with a confusion matrix and ROC curve analysis. The evaluation was done to a model with the highest accuracy, which is 73.1%. The evaluation results of the training metrics can be seen in Fig. 8, while the results for testing metrics can be seen in Fig. 9. The AUC resulting in training metric evaluation was 0.85 and 0.82 for the testing metrics.

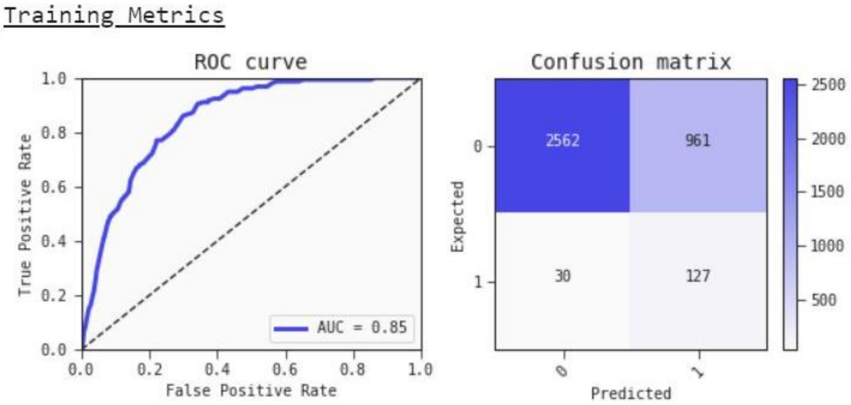


Fig 8. Evaluating Training Metrics

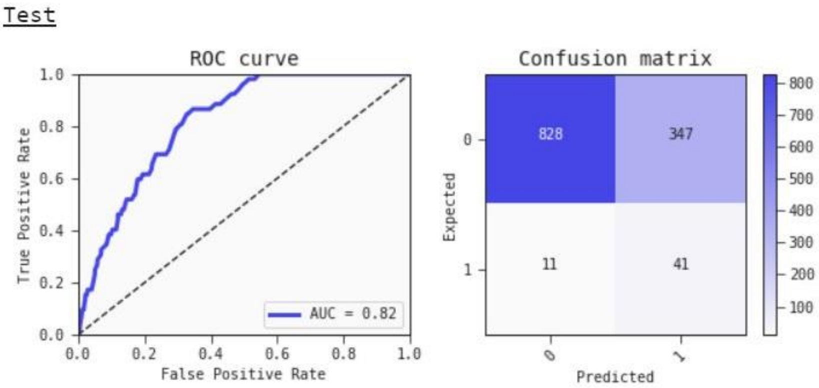


Fig 9. Evaluating Testing Metrics

According to Fig. 8 and Fig. 9, the quality of the training metrics was reasonably good since the blue line (AUC) value is close to the left corner of the graph, which is 0.85. The quality of the testing metrics was also reasonably good with a similar condition to training metrics; the AUC value of the testing metrics was 0.82. The results of this study were then compared with the results of previous researchers. A comparison of the results of using the model can be seen in Table 4.

Table 4. Comparison of Research Result

No	**Researcher**	**Results**	
		Models	**Results**
1	Zhu	CNN	80%
2	Jamthikar	CNN	93.15%
3	Liu	Random Forest	85.03%
4	Own	Qlattice	85%
5	Dobryvecher	SVM	80.2%
		Decision Tree	82.2%
		KNN	77.3%
		Naïve Bayes	80%
		Logistic Regression	84%

4. CONCLUSION

The results of the test carried out in this study indicate that the Feyn Qlattice model can be a solution to obtain features that are used to predict stroke. The Feyn Qlattice autorun method can produce the main features of stroke trigger based on a person's gender, i.e., age and hypertension. This autorun method was run for 10 epochs and produced 17596 test models in 57s. The query string parameter in the Feyn Qlattice then focused on the features of age and hypertension. Once applied, there are 1245 models in 10 epochs with a time of 4s. The experimental results showed an increase in accuracy in training metrics from 0.723 to 0.731 and in testing metrics from 0.695 to 0.708. The results of the evaluation using the confusion matrix with the ROC curve show that this model has fairly good performance where the blue curve line (AUC) has approached the top-left corner of the graph.

REFERENCES

[1] J. Liu, Y. Sun, J. Ma, J. Tu, Y. Deng, P. He, R. Li, F. Hu, H. Huang, X. Zhou, and S. Xu, "Analysis of main risk factors causing stroke in Shanxi Province based on machine learning models," *Informatics Med. Unlocked*, vol. 26, p. 100712, 2021. https://doi.org/10.1016/j.imu.2021.100712

[2] W. C. Chen, M. Y. Hsiao, and T. G. Wang, "Prognostic factors of functional outcome in post-acute stroke in the rehabilitation unit," *Journal of Formosan Medical Association*, 2021. https://doi.org/10.1016/j.jfma.2021.07.009

[3] J. D. Perkins, S. S. Wilkins, S. Kamran, and A. Shuaib, "Post-traumatic stress disorder and its association with stroke and stroke risk factors: A literature review," *Neurobiol. Stress*, vol. 14, no. 100332, pp. 1–14, 2021. https://doi.org/10.1016/j.ynstr.2021.100332

[4] A. Tjan, I. G. R. Widiana, E. D. Martadiani, I. M. D. P. Ayusta, M. W. Asih, and F. P. Sitanggang, "Carotid artery stiffness measured by strain elastography ultrasound is a stroke risk factor," *Clin. Epidemiol. Glob. Heal.*, vol. 12, no. May, pp. 1–5, 2021. https://doi.org/10.1016/j.cegh.2021.100850

[5] O. Ookeditse, T. R. Motswakadikgwa, K. K. Ookeditse, G. Masilo, Y. Bogatsu, B. C. Lekobe, M. Mosepele, H. Schirmer, and S. H. Johnsen "Healthcare professionals' knowledge of modifiable stroke risk factors: A cross-sectional questionnaire survey in greater Gaborone, Botswana," *eNeurologicalSci*, vol. 25, no. 100365, pp. 1–6, 2021. https://doi.org/10.1016/j.ensci.2021.100365

[6] F. Khennou, Y. I. Khamlichi, and N. E. H. Chaoui, "Improving the use of big data analytics within electronic health records: A case study based OpenEHR," in *Procedia Computer Science*, 2018, vol. 127, pp. 60–68. https://doi.org/10.1016/j.procs.2018.01.098

[7] M. Tavana, "Transforming healthcare one byte at a time in the world of big data," *Healthc. Anal.*, vol. 1, p. 100003, 2021. https://doi.org/10.1016/j.health.2021.100003

[8] Y. Yang, X. Zheng, W. Guo, X. Liu, and V. Chang, "Privacy-preserving fusion of IoT and big data for e-health," *Futur. Gener. Comput. Syst.*, vol. 86, pp. 1437–1455, 2018. https://doi.org/10.1016/j.future.2018.01.003

[9] Beata Butryn, I. Chomiak-Orsa, K. Hauke, M. Pondel, Agnieszka, and Siennicka, "Application of Machine Learning in medical data analysis illustrated with an example of association rules," in *Procedia Computer Science*, 2021, vol. 192, pp. 3134–3143. https://doi.org/10.1016/j.procs.2021.09.086

[10] J. Waring, C. Lindvall, and R. Umeton, "Automated machine learning: Review of the state-of-the-art and opportunities for healthcare," *Artif. Intell. Med.*, vol. 104, no. October, p. 101822, 2020.

https://doi.org/10.1016/j.artmed.2020.101822

[11] K. Kosteva, T. Wu, Y. Wang, K. Chaudhuri, and C. Tanislav, “Predicting the risk of stroke in patients with late-onset epilepsy: A machine learning approach,” *Epilepsy Behav.*, vol. 122, p. 108211, 2021. https://doi.org/10.1016/j.yebeh.2021.108211

[12] L. Velagapudi, N. Mouchtouris, M. P. Baldassari, D. Nauheim, O. Khanna, F. A. Saiegh, N. Herial, M. R. Gooch, S. Tjoumakaris, R. H. Rosenwasser, and P. Jabbour, “Discrepancies in Stroke Distribution and Dataset Origin in Machine Learning for Stroke,” *J. Stroke Cerebrovasc. Dis.*, vol. 30, no. 7, p. 105832, 2021. https://doi.org/10.1016/j.jstrokecerebrovasdis.2021.105832

[13] H. Zhu, L. Jiang, H. Zhang, L. Luo, Y. Chen, and Y. Chen, “An automatic machine learning approach for ischemic stroke onset time identification based on DWI and FLAIR imaging,” *NeuroImage Clin.*, vol. 31, p. 102744, 2021. https://doi.org/10.1016/j.nicl.2021.102744

[14] A. Jamthikar, D. Gupta, N. N. Khanna, L. Saba, J. R. Laird, and J. S. Suri, “Cardiovascular/stroke risk prevention: A new machine learning framework integrating carotid ultrasound image-based phenotypes and its harmonics with conventional risk factors,” *Indian Heart J.*, vol. 72, no. 4, pp. 258–264, 2020. https://doi.org/10.1016/j.ihj.2020.06.004

[15] Abzu, “The QLattice is a radical new machine learning model,” 2020. https://www.abzu.ai/qlattice (accessed Oct. 06, 2021).

[16] V. A. Bharadi, “QLattice Environment and Feyn QGraph Models—A New Perspective Toward Deep Learning,” in *Emerging Technologies for Healthcare*, pp. 69–92 2021. https://doi.org/10.1002/9781119792345.ch3

[17] Fedesoriano, “Stroke Dataset,” 2020. https://www.kaggle.com/fedesoriano/stroke-prediction-dataset (accessed Oct. 06, 2021).

[18] G. Y. Wong, F. H.F.Leung, and Sai-HoLing, “A hybrid evolutionary preprocessing method for imbalanced datasets,” *Information Sciences*, vol. 454–455, pp. 161–177, 2018. https://doi.org/10.1016/j.ins.2018.04.068

[19] K. Stöger, D. Schneeberger, P. Kieseberg, and A. Holzinger, “Legal aspects of data cleansing in medical AI,” *Comput. Law Secur. Rev.*, vol. 42, pp. 1–13, 2021. https://doi.org/10.1016/j.clsr.2021.105587

[20] S. Sachan, F. Almaghrabi, J.-B. Yang, and D.-L. Xu, “Evidential reasoning for preprocessing uncertain categorical data for trustworthy decisions: An application on healthcare and finance,” *Expert Syst. Appl.*, vol. 185, 2021. https://doi.org/10.1016/j.eswa.2021.115597

[21] O. A. Olabanjo, B. S. Aribisala, M. Mazzara, and A. S. Wusu, “An ensemble machine learning model for the prediction of danger zones: Towards a global counter-terrorism,” *Soft Comput. Lett.*, vol. 3, p. 100020, 2021. https://doi.org/10.1016/j.socl.2021.100020

[22] S. Gnat, “Impact of Categorical Variables Encoding on Property Mass Valuation,” in *Procedia Computer Science*, 2021, vol. 192, pp. 3542–3550. https://doi.org/10.1016/j.procs.2021.09.127

[23] K. Pawluszek-Filipiak and A. Borkowski, “On the importance of train-test split ratio of datasets in automatic landslide detection by supervised classification,” *Remote Sens.*, vol. 12, no. 18, 2020. https://doi.org/10.3390/rs12183054

[24] A. Rácz, D. Bajusz, and K. Héberger, “Effect of dataset size and train/test split ratios in qsar/qspr multiclass classification,” *Molecules*, vol. 26, no. 4, pp. 1–16, 2021. https://doi.org/10.3390/molecules26041111

[25] G. Sambasivam and G. D. Opiyo, “A predictive machine learning application in agriculture: Cassava disease detection and classification with imbalanced dataset using convolutional neural networks,” *Egypt. Informatics J.*, vol. 22, no. 1, pp. 27–34, 2021. https://doi.org/10.1016/j.eij.2020.02.007

[26] H. Seo, S. Back, S. Lee, D. Park, T. Kim, and K. Lee, “Intra- and inter-epoch temporal context network (IITNet) using sub-epoch features for automatic sleep scoring on raw single-channel EEG,” *Biomed. Signal Process. Control*, vol. 61, p. 102037, 2020. https://doi.org/10.1016/j.bspc.2020.102037

[27] A. Luque, A. Carrasco, A. Martín, and A. de las Heras, “The impact of class imbalance in classification performance metrics based on the binary confusion matrix,” *Pattern Recognit.*, vol. 91, pp. 216–231, 2019. https://doi.org/10.1016/j.patcog.2019.02.023

[28] K. R. Singh, K. P. Neethu, K. Madhurekaa, A. Harita, and P. Mohan, “Parallel SVM model for forest fire prediction,” *Soft Comput. Lett.*, vol. 3, p. 100014, 2021. https://doi.org/10.1016/j.socl.2021.100014

[29] W. Rahmaniar, W.-J. Wang, C.-W. Chiu, and N.L. Hakim “Real-Time Bi-Directional People Counting Using an RGB-D Camera”, *Sensors Review*, vol. 41, no. 4, pp. 341-349, 2021.

[30] K. Gajowniczek and T. Ząbkowski, “ImbTreeAUC: An R package for building classification trees using the area under the ROC curve (AUC) on imbalanced datasets,” *SoftwareX*, vol. 15, p. 100755, 2021. https://doi.org/10.1016/j.softx.2021.100755

[31] S. Yang and G. Berdine, “The receiver operating characteristic (ROC) curve,” *Southwest Respir. Crit. Care Chronicles*, vol. 5, no. 19, p. 34, 2017. https://doi.org/10.12746/swrccc.v5i19.391

[32] T. C. F. Polo and H. A. Miot, “Use of roc curves in clinical and experimental studies,” *J. Vasc. Bras.*, vol. 19, pp. 1–4, 2020. https://doi.org/10.1590/1677-5449.200186

BIOGRAPHY OF AUTHORS

Purwono was Born on May 16, 1989 in Banyumas Indonesia. He is a graduate of the Information Systems College of Computer Science (STIKOM) Yos Sudarso in 2019. His postgraduate education is a master's program in Informatics Engineering at Universitas Ahmad Dahlan (UAD). Currently, he is a lecturer in the informatics study program at Harapan Bangsa University (UHB) Purwokerto. Areas of interest are Data Science, Blockchain, Internet of Things. Email: purwono@uhb.ac.id

Alfian Ma'arif (Member, IEEE, IAENG, ASCEE) received the bachelor's degree from the Department of Electrical Engineering, Universitas Islam Indonesia, Indonesia, in 2014, and the M. Eng. degree from the Department of Electrical Engineering, Universitas Gadjah Mada, Indonesia, in 2017., Since October 2018, he has been a Lecturer with the Department of Electrical Engineering, Universitas Ahmad Dahlan. He got the Assistance professor in 2020. His research interests include control systems and computer programming. Email: alfian.maarif@te.uad.ac.id

Iis Setiawan Mangku Negara was Born on February 16, 1976 in Tanjung Karang, Indonesia. A graduate of Information Management and Computer Engineering at the Institute of Science & Technology "Akprind" Yogyakarta in 2000, he continued his Masters in Information Technology at the University of Indonesia and received his degree in 2013. He is currently working as a lecturer in the Bachelor of Nursing study program at the College of Health Sciences. Nation's hope. After changing his form from high school to university, he was placed as a lecturer in the Information Technology study program at Harapan Bangsa University. Areas of interest are IT Governance, IT Strategic Planning, and Data Mining. Email: iissmn@gmail.com

Wahyu Rahmaniar is currently a postdoctoral researcher at the National Taipei University of Technology, Taiwan. She is received a B.S. degree in Electronics and Instrumentation from Universitas Gadjah Mada, Yogyakarta, Indonesia, in 2009 and a Ph.D. degree in Electrical Engineering from National Central University, Taiwan, in 2020. Her research interests are in the areas of robotics, computer vision, image processing, and artificial intelligence. Email: wahyu.rahmaniar@gmail.com

Jihad Rahmawan received a bachelor's degree from the Department of Electrical Engineering, Universitas Ahmad Dahlan, Indonesia, in 2019. He is now doing a study as a research student at the Faculty of Software and Informatic Science, Iwate Prefectural University, Japan, from 2020. And start for M.Sc in 2021. His research interests include Control Systems, robotics, and Artificial Intelligence. Email: g231s502@s.iwate-pu.ac.jp