Journal of Advanced Health Informatics Research (JAHIR)
Vol. 1, No. 2, August 2023, pp. 95-105
ISSN: 2985-6124, DOI: 10.59247/jahir.v1i2.87

Comparison of Classification and Regression Model Approaches on the Main Causes of Stroke with Symbolic Regression Feyn Qlattice

Purwono Purwono [1], Agung Budi Prasetio[2], Burhanuddin bin Mohd Aboobaider[3]
[1] *Department of Informatics, Universitas Harapan Bangsa, Indonesia*
[2] *Faculty Computer Science, Institut Teknologi Tangerang Selatan, Indonesia*
[3] *Faculty of Information & Technology, Universiti Teknikal Malaysia Melaka, Malaysia*

ARTICLE INFO

Article history:

Received July 28, 2023
Revised September 18, 2023
Published September 23, 2023

Keywords:

Stroke;
Qlattice;
Classification;
Regression;
Comparison;
Feyn

ABSTRACT

Stroke is one of the deadliest diseases in the world, caused by damage to brain tissue resulting from a blockage in the cerebrovascular system. Proper treatment is essential to avoid worsening complications in patients. Several main triggering factors for stroke include hypertension, obesity, smoking habits, lack of physical activity, excessive alcohol consumption, diabetes, and high cholesterol levels. The advancement of information technology allows for early disease prediction through the utilization of AI and Machine Learning technology. The vast amount of data available on medical and health services worldwide can be maximized to identify risk factors for various diseases, including stroke. Machine learning techniques can be employed to predict the causes of stroke. In this study, we were inspired to use the Feyn Qlattice model approach to address stroke. Both classification and regression models were tested in this study. The results indicate that the classification model performs better, achieving an accuracy rate of 0.95. In contrast, the regression model yielded less satisfactory results, with R2, MAE, and RMSE values considered inadequate. This conclusion is supported by the regression plot and residual plot, both of which indicate suboptimal performance. Hence, maximizing the use of the Feyn Qlattice regression model in datasets related to the causes of stroke is recommended.

Corresponding Author:

Purwono Purwono, Department of Informatics, Universitas Harapan Bangsa, Indonesia
Email: purwono@uhb.ac.id

1. INTRODUCTION

Stroke is the second leading cause of death worldwide, and the number of cases continues to rise, particularly in countries with low and middle-income levels [1]. According to data from the World Stroke Organization, 13 million people suffer from a stroke each year, resulting in approximately 5.5 million deaths [2]. One of the symptoms of stroke that you should be vigilant about is the rapid death of brain cells within minutes [3]. During those critical minutes, there appear to be many delays in the treatment of acute ischemic stroke, leading to the loss of 1.9 million neurons [4]. Early recognition is essential in the emergency department to reduce the long-term disability caused by stroke [5].

Stroke is caused by damage to brain tissue when there is a blockage in the cerebrovascular system [6]. This damage leads to disruptions in the sensory and motor areas of the human body, preventing proper functioning of bodily functions controlled by brain tissue [7]. Treatment for this condition must be administered correctly to avoid exacerbating complications in the affected individual. Several main triggering factors for stroke include hypertension, obesity, smoking habits, a lack of physical activity, excessive alcohol consumption, diabetes, and elevated cholesterol levels [8]. Stroke can occur in anyone, regardless of age, gender, or physical condition [9].

Stroke develops very quickly and presents with a wide range of symptoms [2]. Stroke symptoms can occur suddenly, including sudden paralysis in the legs or arms on one side of the human body, numbness, difficulty speaking, walking difficulties, reduced vision, and, in severe cases, loss of consciousness leading to a coma [2].

The advancement of science, particularly in the field of information and communication technology, especially in artificial intelligence (AI) and machine learning (ML), has become a valuable tool for predicting various types of diseases. This includes the use of classification models for diabetes [10], regression models for predicting glucose levels [11], as well as predicting hypertension [2], cholesterol levels [12], and even Covid-19 [13], among others.

The trend of using Electronic Health Records (EHR) continues to grow worldwide [3]. As a result, there has been a significant accumulation of medical and health services data [14] [15]. From this extensive dataset, we can explore various methods to identify risk factors for different diseases [3]. Machine learning technology offers a valuable solution for extracting patterns from large datasets, which can then contribute to new knowledge [16]. Machine learning has been employed to predict the primary causes of stroke [17] [18].

Various types of research related to the use of machine learning as a tool for predicting stroke have been conducted by numerous researchers worldwide. For instance, Biswas [19] conducted research comparing various machine learning algorithms on stroke data. Biswas concluded that the SVM algorithm achieved the highest accuracy, recall, and precision values at 99.99%.

Similarly, Veerle [20] conducted research using logistic regression and SVM algorithms to predict the cause of stroke. Veerle found that the model created was capable of producing an average AUC value of 0.76 for eye movement features. Chong [21] conducted research related to stroke to investigate the relationship between the structure and function of the corticomotor system early after stroke using machine learning. Chong obtained research results in the form of an SVM classification model with an 81% accuracy in detecting motor generating potential, although false positives were more common than false negatives. Additionally, Zhu [22] conducted research using a machine learning approach to identify the time of onset of ischemic stroke based on DWI and FLAIR images. This research yielded classification model testing results with an accuracy value of 0.805, a sensitivity value of 0.769, and a specificity value of 0.840.

Based on various similar studies, we were inspired to employ the Feyn Qlattice model approach. According to information from the official website, Abzu, accessible at https://docs.abzu.ai/, Qlattice is a form of supervised learning designed for symbolic regression needs, and it was developed using Feynman's path integral formulation [3]. Qlattice collaboratively constructs functions to establish a mathematical model between input and output within the dataset [23]. In general, symbolic regression approaches tend to maintain high performance and uphold generalization.

In contrast to previous research, our contribution to this study is conducting trials to compare the results of classification and regression model approaches on the stroke dataset using Feyn Qlattice. These two approaches will be utilized to identify the primary features contributing to the main causes of stroke. The classification and regression models will be assessed using established machine learning evaluation standards.

2. METHODS

2.1. Data Preparation

The dataset used in this study belongs to Fedesoriano [24]. This dataset consists of several important features to determine the causes of stroke. Details of the features contained in this dataset can be seen in Table 1. Based on Table 1, there are 11 clinical features utilized in this study. However, this dataset cannot be fully implemented with machine learning models at this stage. We have identified a substantial amount of data containing empty or NaN values. Additionally, we require data transformation to convert it into numerical format. The quality of the machine learning model to be employed will be influenced by data processing [25].

The dataset is subsequently processed to optimize the performance of the machine learning model. The initial stage of data processing involves removing all null or NaN data, a process commonly referred to as "handling missing data" [26]. The next step is converting categorical data into numerical data. One commonly

used method for this purpose is One-Hot encoding (1H), which transforms categorical data into numerical format, often considered effective in handling memory issues when dealing with high cardinalities [27]. For instance, gender data containing "male" and "female" can be transformed into numerical values, such as "male" = 0 and "female" = 1. We implement this conversion for all categorical data in the dataset. Once the dataset is prepared, it undergoes a splitting process, separating it into training data and test data [28]. In this research, we used a 75% to 25% composition for training versus testing.

Table 1. Clinical Features for Predicting Stroke Events

Attribute	Information
id	unique identifier
gender	"Male", "Female" or "Other"
age	age of the patient
hypertension	0 if the patient doesn't have hypertension, 1 if the patient has hypertension
heart_disease	0 if the patient doesn't have any heart diseases, 1 if the patient has a heart disease
ever_married	"No" or "Yes"
work_type	"children", "Govt_jov", "Never_worked", "Private" or "Self-employed"
Residence_type	"Rural" or "Urban"
avg_glucose_level	average glucose level in blood
bmi	body mass index
smoking_status	"formerly smoked", "never smoked", "smokes" or "Unknown"*
stroke	1 if the patient had a stroke or 0 if not

2.2. Feyn Qlattice

In an effort to identify the primary features that cause stroke, we conducted a trial using the Feyn Qlattice. The various stages of utilizing the Feyn Qlattice model [3] involve creating thousands of models, fine-tuning the model with the backpropagation version, assessing several criteria like the loss function and information criteria, discarding the least performing model, and continuing this iterative process until we find the best model. You can observe the iterative process of using the Feyn Qlattice in Figure 1 [29].

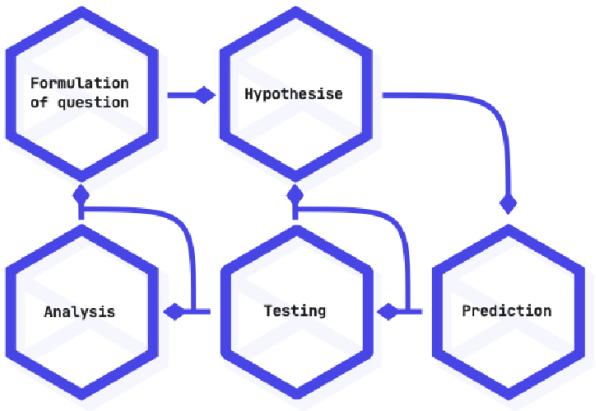

Fig. 1. The Scientific Method as an Iterative Process

Qlattice can be regarded as a probability distribution from which the model is sampled. The initial distribution is uniform and is established after each update call. This process aids qlattices in converging and shaping a distribution towards the best solution.

2.3. Classification Model Approach

Feyn Qlattice incorporates classification features that can be employed for predicting stroke data. This model boasts an autorun feature capable of generating thousands of models for selection, ultimately identifying the best model [23]. The Feyn Qlattice also includes an epoch hyperparameter that functions similarly to an artificial neural network. A single run of the autorun function generates at least 10/10 epoch Feyn Qlattices. From these 10 epochs, thousands of training data models for machine learning are produced. Parameters within autorun comprise data, which houses the training data; output_name, representing the classification target; and kind, indicating the type of learning—either classification or regression. The autorun function generates a dynamic illustration illustrating the features contributing to stroke, with the primary stroke-causing features clearly displayed and linked to the prediction target.

For model evaluation, the classification model in Feyn Qlattice employs the receiver operating characteristic (ROC) curve and Confusion Matrix. The ROC Curve is a two-dimensional plot that illustrates the performance of a classification model as the discrimination threshold value is altered across a range of predictor variables [30]. On the x-axis, we have the false positive rate for predictive tests, while the y-axis represents the true positive rate for predictive tests. ROC analysis provides the likelihood as the slope of the tangent line to the ROC curve at the point corresponding to the test results. The calculation of this slope is demonstrated in Equation 1 [31].

$$B_{i,n}(t) = \tag{1}$$

$$\binom{n}{i} t^i (1-t)^{n-i} \ \ with\ t\ ranging\ from\ 0\ to\ 1$$

For this purpose, we require a cubic Bézier curve defined by Equation 2.

$$B(t) \ = \ (1-t)^3 P_0 \ + \ 3t(1-t)^2 P_1 \ + \ 3t^2(1-t)P_2 \ + \ t^3 P_3 \tag{2}$$

In principle, a curve is defined by two endpoints (P0 and P3) with corresponding tangent lines and a control point of the form B(t) = S(u) on the curve. To interpret the ROC Curve, you can follow these guidelines: it is considered to have a poor value if the resulting curve closely follows the baseline or the horizontal line at the point (0.0), while it is considered to have a good value if the curve approaches the point (0.1).

Evaluation of classification models can also be performed using a confusion matrix. This method serves as a type of predictive analytical tool for displaying and comparing actual values with model-predicted values in the form of evaluation metrics such as accuracy, precision, recall, and F1-score [32].

2.4. Regression Model Approach

In our quest to identify the primary features contributing to stroke, we turned to a regression model. In addition to utilizing classification models, we can also employ the autorun function with regression models. The regression function in Feyn Qlattice also incorporates the epoch hyperparameter, capable of generating thousands of regression models until one of the best regression models is selected. The autorun function in the regression model will also generate illustrations highlighting relevant features associated with stroke.

To evaluate the regression model, we can employ regression plots and residual plots. The regressor serves as a comparison tool between the actual and predicted values. The regressor plot can be represented as a tuple, with the actual value of the target variable on the x-axis and the predicted value of the regressor on the y-axis. An ideal prediction would have all points lying on the dotted line y=x. Additionally, we can use the residual plot to analyze whether errors follow a normal distribution. An atypical distribution may indicate bias in the model, while a random distribution suggests an unbiased model. The regressor provides values for R2 score, Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), regression plot, and residual plot. R2, or R-squared, is calculated using a mathematical equation as shown in Equation 3 [33].

$$R^2 = 1 - \frac{\sum_{i=1}^{m}(X_i - Y_i)^2}{\sum_{i=1}^{m}(Y - Y_i)^2} \tag{3}$$

The coefficient of determination R2 has a range of values from the worst at -∞ to the best at +1. When R2 has a value of 0.50, it means that half of the observed variation can be explained by the input model. The coefficient of determination can be interpreted as the proportion of variance in the dependent variable that can

be predicted from the independent variable. MAE is calculated using mathematical equations, as seen in Equation 4 [33].

$$MAE = \frac{1}{m}\sum_{i-1}^{m}|X_i - Y_i| \quad (4)$$

MAE has a worst value of +∞, with the best value being 0. MAE is a measure of the error between paired observations that represent the same phenomenon. The mathematical equation for evaluating RMSE can be seen in Equation 5 [33].

$$RMSE = \sqrt{\frac{1}{m}\sum_{i=1}^{m}(X_i - Y_i)^2} \quad (5)$$

This equation has a best value of 0 and a worst value of +∞. RMSE measures the level of accuracy of a model's prediction results [33].

3. RESULTS AND DISCUSSION

3.1. Data Transformation

The data concerning clinical features contributing to stroke has undergone the preprocessing stage, which includes removing null or NaN values. Additionally, data transformation has been conducted using one-hot encoding, converting categorical data into numeric form. Table 2 represents the outcome of the completed data transformation.

Table 2. Data Transformation

gender	age	hypertension	heart_disease	…	smoking
1	67	0	1	…	1
1	80	0	1	…	1
0	49	0	0	…	1
0	79	1	0	…	1
1	82	0	0	…	0
…	…	…	…	…	…

Based on Table 2, there are no longer any visible data with NaN or null values as they have been selected and deleted. It is also apparent that the data has been transformed into numeric format, containing only encoded numbers.

3.2. Classification Model

The classification model is executed using the auto-run function provided by Feyn Qlattice. The training data used comprises 3680 features. Subsequently, Feyn Qlattice generates an illustration depicting the relationship between the main features that contribute to stroke, as shown in Fig. 2.

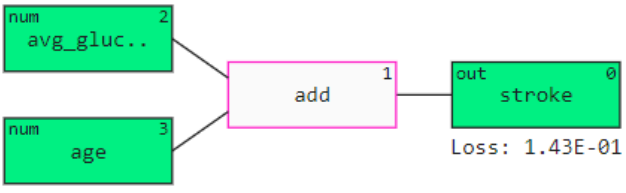

Fig. 2. The Main Features of The Stroke Causes Classification Model.

Based on Fig. 2, we have identified the main features that can cause a stroke, namely 'age' and 'average glucose level in the blood.' It appears that a person's age is correlated with their blood sugar levels, which is a significant factor that can potentially lead to a stroke. These two features hold greater importance compared to several other variables. As an individual ages, it becomes increasingly crucial to maintain balanced blood sugar levels to avoid experiencing stroke symptoms.

In this classification type, Feyn Qlattice generated a total of 10,372 models, each with 10/10 epochs, completing the process in 49 seconds. Among the 10,372 models, the best one identified 'age' and 'average glucose level in the blood' as the main features with the potential to cause a stroke.

3.3. Regression Model

The regression model was also executed using the auto-run function. This model aims to predict the primary features of stroke. Similar to the classification model, we also utilized 3680 training data. An illustration depicting the relationship between the features that contribute to stroke can be observed in Fig. 3.

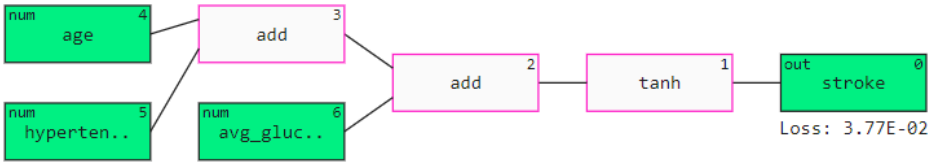

Fig. 3. The main features contributing to the causes of stroke are addressed by the regression models

Based on Fig. 3, the regression model identifies several features that are interrelated and serve as triggers for stroke. These features include 'age,' 'hypertension,' and 'average glucose level in the blood.' The Feyn Qlattice auto-run function generated 11,186 models, each with a total of 10/10 epochs, and completed the process in 45 seconds. Among these 11,186 models, the best one predicts that 'age,' 'hypertension,' and 'average glucose level in the blood' are the primary features with the potential to cause a stroke.

3.4. Evaluation of Classification Model

The evaluation of the classification model involves the use of the ROC Curve and confusion matrix. The ROC Curve is generated from the best model among the 10,372 models created, each with a total of 10/10 epochs. The ROC Curve for the training data is displayed in Fig. 4, and the ROC Curve for the testing data is presented in Fig. 5.

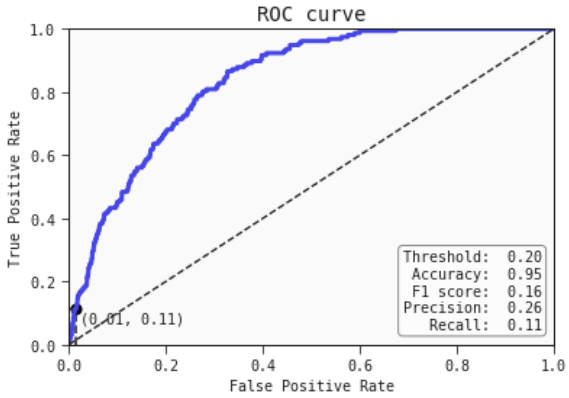

Fig. 4. ROC Curve Data Training

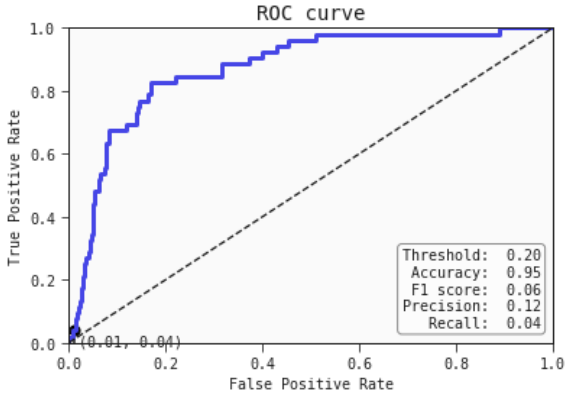

Fig. 5. ROC Curve Data Testing

Based on Fig. 4, it is evident that the point is close to 0.1, indicating that the resulting model is sufficiently effective in classifying the causes of stroke within the training data. Additionally, Fig. 4 demonstrates that the resulting accuracy of the classification model is 0.95. Similarly, based on Fig. 5, the point is also near 0.1, signifying that this model performs quite well in classifying the causes of stroke within the test data. Furthermore, Fig. 5 shows that the accuracy of the classification model is 0.95.

There is a distinction between the ROC Curve results for the training data and the test data. Although both exhibit an accuracy of 0.95, the F1-score value for the training data is 0.16, while for the test data, it is 0.06. The precision for the training data is 0.26, whereas for the test data, it is 0.12. The recall value for the training data is 0.11, while for the test data, it is 0.04. Consequently, it can be inferred that the ROC Curve generated from the training data yields higher F1-score, precision, and recall values compared to the test data. Regarding the evaluation using the confusion matrix, two matrices are generated: one for the training data (Fig. 6) and one for the test data (Fig. 7).

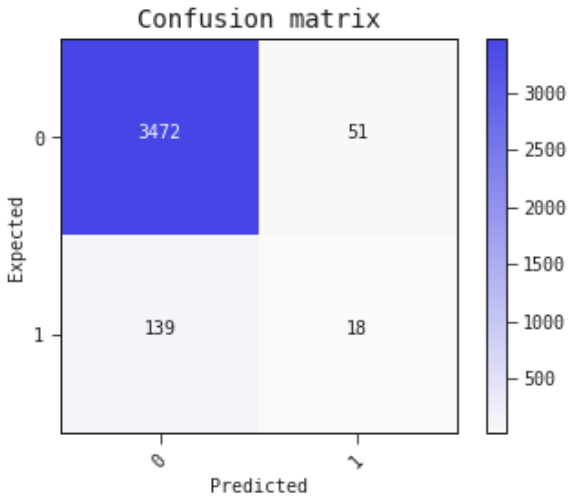

Fig. 5. Confusion Matrix Data Training

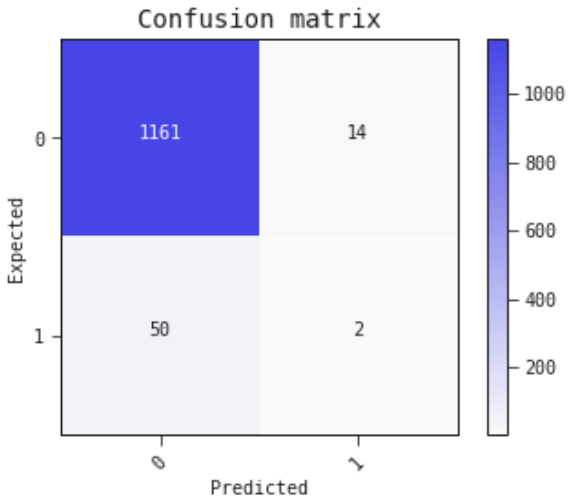

Fig. 5. Confusion Matrix Data Testing

Based on Fig. 6, we can conclude that out of the 3523 training data points classified as class 0 (not stroke), the system predicts 3472 as class 0 and 51 as class 1 (stroke). Additionally, from the 157 training data points classified as class 1, the system predicts 139 as class 1 and 18 as class 0. Furthermore, based on Fig. 7, we can

also infer that out of the 1175 test data points belonging to class 0, the system predicts 1161 as class 0 and 14 as class 1. For the 52 test data points, the system identifies 50 as class 1 and 2 as class 0.

3.5. Evaluation of Regression Model

The evaluation of the regression model yields R2, RMSE, MAE values, as well as regression plots and residual plots. The R2, RMSE, and MAE values for this model are presented in Table 3.

Table 3. Evaluation Metrics Regression Model

Training Metrics		Testing Metrics	
R2	0.0769	R2	0.117
RMSE	0.194	RMSE	0.189
MAE	0.0813	MAE	0.0797

Based on Table 3, the R2 value in the evaluation of the regression model, for both training data and test data, is still considered inadequate as the obtained values remain below 1. On the other hand, the MAE value in this evaluation is quite favorable as it is close to 0, and the RMSE value is also deemed satisfactory as it approaches

The regression plots for the training data generated by this model can be observed in Fig. 8, and for the test data in Fig. 9. The regression plot will demonstrate a perfect prediction if each point aligns with the dotted line y=x.

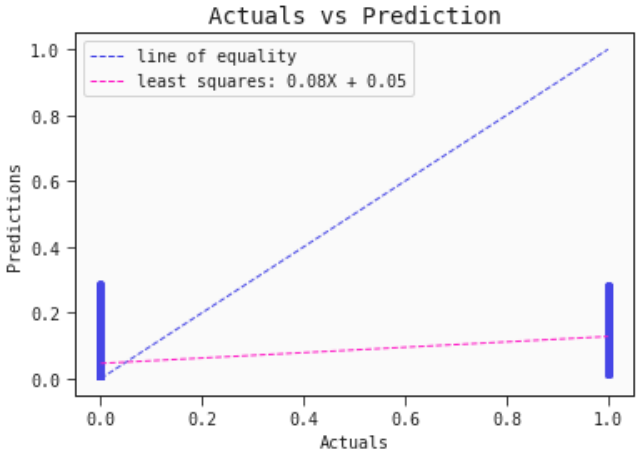

Fig. 8. Regression Plot Data Training

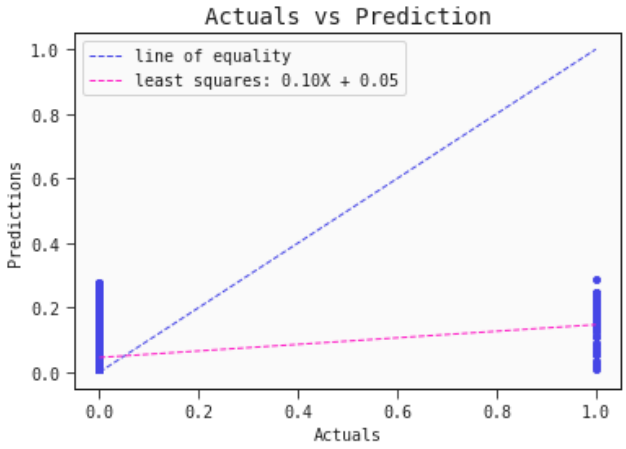

Fig. 9. Regression Plot Data Testing

Based on Fig. 8 and Fig. 9, the regression models generated are deemed insufficient as the distribution of points still falls outside the range of the dotted line y=x. Further optimization is required if a regression model is to be effectively applied to this dataset. Meanwhile, the resulting residual plots are displayed in Fig. 10 for the training data and Fig. 11 for the test data.

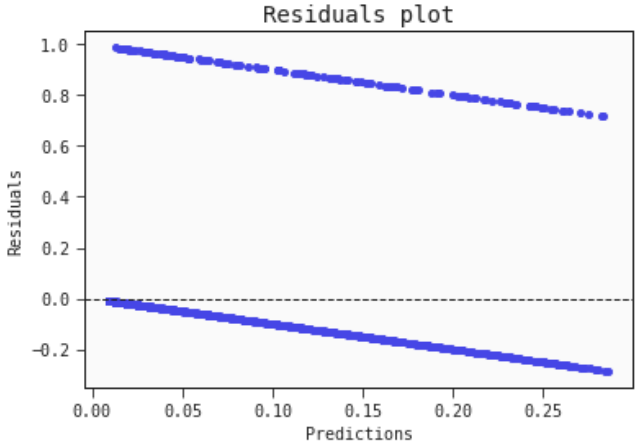

Fig. 10. Residual Plot Data Training

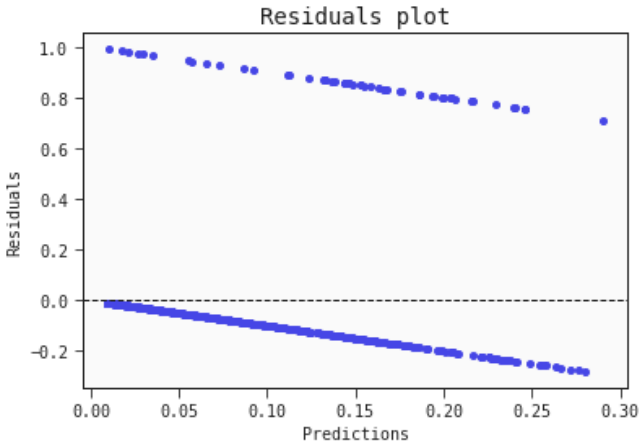

Fig. 11. Residual Plot Data Testing

Based on Fig. 10 and Fig. 11, it is evident that the regression model does not illustrate a non-random distribution of data. This is indicative of a bias in the regression model.

4. DISCUSSION

Testing of the classification and regression models has been completed on the stroke cause dataset using the Feyn Qlattice approach. This research yields an illustration suggesting that the classification model is more suitable for use with this dataset, as it achieves a reasonably high accuracy value of 0.95. On the other hand, in the context of the regression model analysis, we have observed that the R2, MAE, and RMSE values are still considered suboptimal. Additionally, the results from the regression plot and residual plots further support the conclusion that the utilization of the Feyn Qlattice regression model on the stroke cause dataset is less than optimal. We recognize the need for ongoing optimization efforts or improved data handling to obtain machine learning models with better performance.

5. CONCLUSION

This research aims to compare the performance results of thousands of Feyn Qlattice models produced by classification and regression models in handling stroke cause datasets. The study's findings reveal several primary causes of stroke. From the classification model, we identified the main features that contribute to

stroke, namely 'age' and 'average glucose level in blood.' It appears that a person's age is correlated with their blood sugar levels, which is an important factor with the potential to cause a stroke. In contrast, from the regression model, we identified several features that are interrelated as triggers for stroke. These features include 'age,' 'hypertension,' and 'average glucose level in blood.' The classification model is considered the appropriate choice for application with Feyn Qlattice to predict strokes because it can achieve a reasonably high accuracy value of 0.95. Meanwhile, for the regression model, we observed R2, MAE, and RMSE values that are still considered suboptimal, along with the results from the regression plot and residual plot. These findings collectively suggest that the use of the Feyn Qlattice regression model on the stroke cause dataset is less than optimal.

REFERENCES

[1] M. O. Owolabi *et al.*, "Primary stroke prevention worldwide: translating evidence into action," *Lancet Public Heal.*, vol. 7, no. 1, pp. e74–e85, 2022, doi: 10.1016/S2468-2667(21)00230-9.

[2] E. Dritsas and M. Trigka, "Stroke Risk Prediction with Machine Learning Techniques," *Sensors*, vol. 22, no. 13, 2022, doi: 10.3390/s22134670.

[3] P. Purwono, A. Ma'arif, I. S. Mangku Negara, W. Rahmaniar, and J. Rahmawan, "Linkage Detection of Features that Cause Stroke using Feyn Qlattice Machine Learning Model," *J. Ilm. Tek. Elektro Komput. dan Inform.*, vol. 7, no. 3, p. 423, 2021, doi: 10.26555/jiteki.v7i3.22237.

[4] J. C. M. Prick *et al.*, "Experiences with information provision and preferences for decision making of patients with acute stroke," *Patient Educ. Couns.*, vol. 105, no. 5, pp. 1123–1129, 2022, doi: 10.1016/j.pec.2021.08.015.

[5] A. Kobayashi *et al.*, "European Academy of Neurology and European Stroke Organization consensus statement and practical guidance for pre-hospital management of stroke," *Eur. J. Neurol.*, vol. 25, no. 3, pp. 425–433, 2018, doi: 10.1111/ene.13539.

[6] J. Liu *et al.*, "Analysis of main risk factors causing stroke in Shanxi Province based on machine learning models," *Informatics Med. Unlocked*, vol. 26, no. June, p. 100712, 2021, doi: 10.1016/j.imu.2021.100712.

[7] W. C. Chen, M. Y. Hsiao, and T. G. Wang, "Prognostic factors of functional outcome in post-acute stroke in the rehabilitation unit," *J. Formos. Med. Assoc.*, no. 7, 2021, doi: 10.1016/j.jfma.2021.07.009.

[8] O. Ookeditse *et al.*, "Healthcare professionals' knowledge of modifiable stroke risk factors: A cross-sectional questionnaire survey in greater Gaborone, Botswana," *eNeurologicalSci*, vol. 25, p. 100365, 2021, doi: 10.1016/j.ensci.2021.100365.

[9] T. Elloker and A. J. Rhoda, "The relationship between social support and participation in stroke: A systematic review," *African J. Disabil.*, pp. 1–9, 2018, doi: 10.4102/ajod.v7i0.357.

[10] N. Fazakis, O. Kocsis, E. Dritsas, S. Alexiou, N. Fakotakis, and K. Moustakas, "Machine Learning Tools for Long-Term Type 2 Diabetes Risk Prediction," *IEEE Access*, vol. 9, pp. 103737–103757, 2021, doi: 10.1109/ACCESS.2021.3098691.

[11] S. Alexiou, E. Dritsas, O. Kocsis, K. Moustakas, and N. Fakotakis, "An approach for Personalized Continuous Glucose Prediction with Regression Trees," 2021, doi: 10.1109/SEEDA-CECNSM53056.2021.9566278.

[12] N. Fazakis, E. Dritsas, O. Kocsis, N. Fakotakis, and K. Moustakas, "Long-term Cholesterol Risk Prediction using Machine Learning Techniques in ELSA Database," no. Ijcci, pp. 445–450, 2021, doi: 10.5220/0010727200003063.

[13] A. S. Kwekha-Rashid, H. N. Abduljabbar, and B. Alhayani, "Coronavirus disease (COVID-19) cases analysis using machine-learning applications," *Appl. Nanosci.*, no. 0123456789, 2021, doi: 10.1007/s13204-021-01868-7.

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

[15] 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, doi: https://doi.org/10.1016/j.future.2018.01.003.

[16] 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, doi: 10.1016/j.artmed.2020.101822.

[17] 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.

[18] L. Velagapudi *et al.*, "Discrepancies in Stroke Distribution and Dataset Origin in Machine Learning for Stroke," *J. Stroke Cerebrovasc. Dis.*, vol. 30, no. 7, p. 105832, 2021, doi: https://doi.org/10.1016/j.jstrokecerebrovasdis.2021.105832.

[19] N. Biswas, K. Mohammad, M. Uddin, and S. Tasmin, "Healthcare Analytics A comparative analysis of machine learning classifiers for stroke prediction : A predictive analytics approach," *Healthc. Anal.*, vol. 2, no. July, p. 100116, 2022, doi: 10.1016/j.health.2022.100116.

[20] V. H. E. W. Brouwer *et al.*, "Applying machine learning to dissociate between stroke patients and healthy controls using eye movement features obtained from a virtual reality task," *Heliyon*, vol. 8, no. 4, p. e09207, 2022, doi: 10.1016/j.heliyon.2022.e09207.

[21] B. Chong, A. Wang, V. Borges, W. D. Byblow, P. Alan Barber, and C. Stinear, "Investigating the structure-function relationship of the corticomotor system early after stroke using machine learning," *NeuroImage Clin.*, vol. 33, p. 102935, 2022, doi: 10.1016/j.nicl.2021.102935.

[22] 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, doi: https://doi.org/10.1016/j.nicl.2021.102744.

[23] P. A. Riyantoko, Sugiarto, I. G. S. M. Diyasa, and Kraugusteeliana, "'F.Q.A.M' Feyn-QLattice Automation Modelling: Python Module of Machine Learning for Data Classification in Water Potability," 2021, doi: 10.1109/ICIMCIS53775.2021.9699371.

[24] Fedesoriano, "Stroke Prediction Dataset," *11 clinical features for predicting stroke events*, 2020. https://www.kaggle.com/datasets/fedesoriano/stroke-prediction-dataset.

[25] F. Farhangi, "Investigating the role of data preprocessing, hyperparameters tuning, and type of machine learning algorithm in the improvement of drowsy EEG signal modeling," *Intell. Syst. with Appl.*, vol. 15, no. July, p. 200100, 2022, doi: 10.1016/j.iswa.2022.200100.

[26] Y. Fu, H. Liao, and L. Lv, "A comparative study of various methods of handling missing data in unsoda," *Agric.*, vol. 11, no. 8, 2021, doi: 10.3390/agriculture11080727.

[27] L. M. Matos, J. Azevedo, A. Matta, A. Pilastri, P. Cortez, and R. Mendes, "Categorical Attribute traNsformation Environment (CANE): A python module for categorical to numeric data preprocessing[Formula presented]," *Softw. Impacts*, vol. 13, no. July, p. 100359, 2022, doi: 10.1016/j.simpa.2022.100359.

[28] Q. H. Nguyen *et al.*, "Influence of data splitting on performance of machine learning models in prediction of shear strength of soil," *Math. Probl. Eng.*, vol. 2021, 2021, doi: 10.1155/2021/4832864.

[29] K. R. Broløs *et al.*, "An Approach to Symbolic Regression Using Feyn," 2021. [Online]. Available: http://arxiv.org/abs/2104.05417.

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

[31] W. Fierz, "A simplified method to approximate a ROC curve with a Bézier curve to calculate likelihood ratios of quantitative test results," *MethodsX*, vol. 7, p. 100915, 2020, doi: 10.1016/j.mex.2020.100915.

[32] A. Kulkarni, D. Chong, and F. A. Batarseh, *Foundations of data imbalance and solutions for a data democracy*. Elsevier Inc., 2020.

[33] D. Chicco, M. J. Warrens, and G. Jurman, "The coefficient of determination R-squared is more informative than SMAPE, MAE, MAPE, MSE and RMSE in regression analysis evaluation," *PeerJ Comput. Sci.*, vol. 7, pp. 1–24, 2021, doi: 10.7717/PEERJ-CS.623