Machine Learning-Based Stroke Prediction

Weijun Deng[1][a], Chen Li[2][b], Zhirui Yan[3][c] and Yuzhuo Yuan[4][d] *

[1] School of International Education, Guangdong University of Technology, Guangzhou, 511495, China

[2]School of Computer Science and Technology (School of Artificial Intelligence), Zhejiang Sci-Tech University, Hangzhou 310000, China

[3]School of Electronic Engineering (School of Artificial Intelligence), South China Agricultural University, Changsha, 510640, China

[4]Dundee International Institute, Central South University, Guangzhou, 410083, China

Keywords: Machine Learning, Logistic Regression, Stroke.

Abstract: The application of machine learning techniques in the medical diagnostic field has seen a gradual increase, with the search for an efficient and reasonably accurate prediction model becoming a focal point in related research areas. This study focuses on the comparison and evaluation of various machine learning models' performance on a stroke prediction dataset, aiming to identify the optimal prediction model. During the preliminary phase of the experiment, the dataset underwent preprocessing, which included handling missing values, label encoding of non-numeric data types, and feature selection based on the relevance between features and prediction labels. Moreover, models such as Logistic Regression, Decision Trees, XGBoost, and Random Forests were selected for in-depth analysis, and the Z-score method was employed for data normalization. Throughout the model tuning process, detailed model optimization was conducted through parameter adjustments and cross-validation methods. This study utilized AUC, precision0, and recall1 as evaluation metrics to conduct a comprehensive analysis of model performance, ultimately determining that the adjusted Random Forest and Logistic Regression models demonstrated the best performance in stroke prediction. The findings of this study provide an effective method for stroke prediction and offer guidance for future research in disease prediction using machine learning.

1 INTRODUCTION

Stroke is recognized as one of the leading causes of death and disability worldwide, and early prediction and accurate diagnosis are crucial to mitigating its harm. Advances in medical technology have provided new methods, particularly in an era where data analysis and processing techniques are increasingly mature, and the application of machine learning in disease prediction and medical diagnosis has gradually attracted attention. Compared to traditional statistical methods, such as conducting multiple linear regression analysis with SPSS software, machine learning offers a more flexible and efficient way of processing data. However, traditional statistical models, with their mature theoretical foundation and ease of interpretation, still hold their ground in certain domains.

In recent years, as a branch of machine learning, deep learning has shown superior performance in stroke prediction research, but its dependency on hardware and the opacity of models limit its application in certain areas. By contrast, machine learning models, with their efficiency in handling specific datasets, strong generalizability, and interpretability of prediction results, have become an important tool in stroke prediction research.

This research aims to explore the effectiveness of machine learning models in stroke prediction, with a particular focus on the performance of models such as Logistic Regression, XGBoost, and Decision Trees on specific datasets. By delving into different

[a] https://orcid.org/0009-0002-6216-5436
[b] https://orcid.org/0009-0007-5048-4217
[c] https://orcid.org/0009-0009-0494-4168
[d] https://orcid.org/0009-0005-7647-3888

Deng, W., Li, C., Yan, Z. and Yuan, Y.
Machine Learning-Based Stroke Prediction.
DOI: 10.5220/0012972300004508
Paper published under CC license (CC BY-NC-ND 4.0)
In *Proceedings of the 1st International Conference on Engineering Management, Information Technology and Intelligence (EMITI 2024)*, pages 754-761
ISBN: 978-989-758-713-9
Proceedings Copyright © 2024 by SCITEPRESS – Science and Technology Publications, Lda.

machine learning methodologies, including Logistic Regression models, researchers are committed to providing a reliable reference for the early prediction of stroke and offering new perspectives and methodological foundations for future research in this field. Facing the severe global health challenge of stroke, the goal of this project is not only to seek technological innovations and methodological advancements but also to provide practical tools for clinical use. This would facilitate early diagnosis and timely treatment, thereby improving patients' survival rates and quality of life.

2 RELATED WORK

Stroke, afflicts 17 million annually(Murphy,2020),is recognized as the second leading cause of death worldwide and a primary source of long-term disability (Silva, 2018), prompting extensive research into predictive methods for stroke using various approaches. There have been statistical researchers who utilized SPSS software and traditional statistical models, such as multiple linear regression, to study common causes of stroke, including factors like age, occupation, and climate conditions of the living environment (Li, 2016). These traditional statistical model studies possess mature techniques, comprehensive theories, and are easy to apply and interpret, yet they gradually show signs of obsolescence with the development of machine learning and deep learning technologies. This is due to reasons such as their relatively simplistic model metrics, lower prediction effectiveness; inability to self-optimize, weaker adaptability, and generalization capacity; and finally, traditional statistical models are constrained by human brain computational and analytical limitations, struggling with large-scale, high-dimensional data processing.

Deep learning has also garnered significant attention in recent predictive research, demonstrating superior performance in many studies. However, the reliance on hardware and the opaqueness of deep neural network (DNN) functions (Wu, 2022), along with the critical issue that deep neural networks can fail entirely in adverse dataset conditions (e.g., extreme imbalance between positive and negative instances), remain challenging to explain.

In contrast, machine learning models showcase high predictive result effectiveness and a comprehensive range of model metrics. Based on adaptive algorithms, they offer strong generalizability and versatility; they can process high-dimensional data efficiently and handle large datasets effectively. It is worth noting that most machine learning model research has matured, possessing a self-consistent and comprehensive theoretical foundation. Researchers have previously trained models using various sampling strategies in predictive studies based on machine learning models, including Logistic Regression, Gradient Boosting Machine, Extreme Gradient Boosting, Random Forests, Support Vector Machines, and Decision Trees. These studies have indicated a significant association between these machine learning models and laboratory variables in relation to stroke recurrence. The models demonstrated the stability of predicting stroke recurrence within a five-year time frame, highlighting the importance of laboratory variables in periodical predictions. Additionally, researchers have utilized various feature selection strategies, evaluating the performance of six interpretable algorithms, showcasing the potential of various machine learning models in predicting long-term stroke recurrence (Zhang 2021, Boukhennoufa 2022, Song 2022)

Beyond the application of foundational models, there have been many fascinating interdisciplinary studies in recent years. For instance, research by Pritam Chakraborty, Anjan Bandyopadhyay, Sricheta Parui, Sujata Swain from the Karolinska Institute of Industrial Technology combined machine learning and game theory in stroke prediction investigations (Chakraborty, 2024). Another study aimed at exploring methods for handling specific, representative stroke datasets, such as Han Zhaoyi and Lian Gaoshe's study from Taiyuan University of Technology, which achieved the highest efficiency in training imbalanced datasets with "SMOTEENN sampling + Recursive Feature Elimination with Random Forests(RFRFE) + XGBoost classification algorithm" (Han, 2023).

This research focuses on the study of machine learning models for stroke prediction.

3 METHODOLOGY

Based on the "Stroke Prediction Dataset," this study conducted a comprehensive analysis of a wide range of clinical patient characteristics and medical indicators using various machine learning models. The aim was to identify the most effective model for predicting stroke. In our research, we first preprocessed the data through label encoding. Subsequently, the experimental data underwent imbalanced learning and feature selection; finally, several machine learning models were constructed and trained, including the Logistic Regression model, Decision Tree model, XGBoost model, Gradient Boosting Decision Tree model, and Random Forest model, and their performance differences were compared. Figure 1 illustrates the workflow of our study.

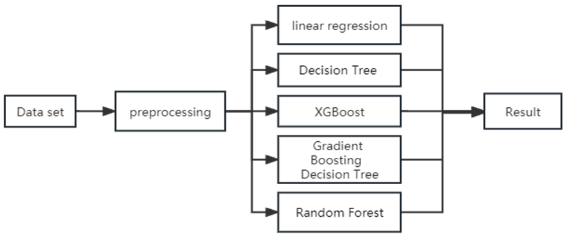

Figure 1: Workflow of our study (Photo/Picture credit: Original).

3.1 Preprocessing

Before establishing machine learning models with the dataset, the experiment first undertook data preprocessing. Given that the dataset used in the experiment contained a small amount of missing values, we opted to remove these missing values. Moreover, to deal with non-numeric data types, the study employed label encoding to transform such data. Subsequently, considering that some features in the dataset had low correlation with the prediction labels, we conducted feature selection, removing features with insufficient correlation coefficients to enhance the training efficiency and predictive performance of the model, thereby reducing the risk of overfitting. Additionally, in order to find the best-performing model, we applied and compared different feature selection methods (corr, LR, DT, RF) to the dataset and selected the model with the best performance.

During the process of training the machine learning models, we divided the original dataset into a training set (70%) and a test set (30%).

3.2 Model Selection and Construction

In this study, we selected several ensemble learning models for predicting stroke (an introduction to ensemble learning methods can be inserted here), including Logistic Regression, Decision Tree, XGBoost, Gradient Boosting Decision Tree, and Random Forest models for modeling and analyzing the stroke prediction dataset. This was done to explore and compare the performance differences between various machine learning models in stroke prediction.

● Logistic Regression

The Logistic Regression model is a type of generalized linear model that is commonly used for solving classification problems. It uses the sigmoid function to map the output of linear regression to the probability space, thereby facilitating the classification of categories 1 and 0. Below is the algorithmic formula of Logistic Regression.

$$P = \frac{1}{1+e^{-(k_0+k_1x_1+k_2x_2+\cdots+k_nx_n)}} \qquad (1)$$

In this formula, (b_0), (b_1), etc., represent the model parameters learned by the machine learning model from the training set. (x_1), (x_2) and so on, are the input features, such as age, gender, type of residence, etc. Through this formula, it is possible to determine the category of the input features.

● Decision Tree The Decision

Tree model is a common algorithm in machine learning, consisting of nodes and edges. The nodes represent attributes, while the edges represent branches leading to different outcomes. This model divides data based on its features, generating multiple nodes. Upon the completion of model training, the Decision Tree forms a tree-like structure. Through the evaluation of each node, the model classifies the data.

● XGBoost

The XGBoost model is an ensemble learning algorithm based on gradient boosting that corrects residuals through continuous iterations to make the final results more closely align with the true labels, thus enhancing model performance. The specific calculation formula is as follows.

$$\hat{y}_i = \sum_{k=1}^{K} j_k(x_i), j_k \in J \qquad (2)$$

In this formula, x_i represents the sample of class i, K is the number of trees, J is the set of all trees, and j_k is the structure and leaf weight function of the tree, each corresponding to an independent tree model. The final calculated result is the associated prediction value.

● Gradient Boosting Decision

Tree The Gradient Boosting Decision Tree is an iterative decision tree algorithm based on boosting ensemble learning that continuously fits residuals. It is known for its strong generalization capability and fast computation speed.

● Random Forest

The Random Forest is an ensemble classifier made up of many decision trees. Its operational principle involves randomly sampling multiple data subsets from the original dataset to construct several decision trees. Then, using the concept of ensemble learning, it integrates multiple trees to improve the model's predictive performance.

3.3 Evaluation Metrics

In this research, we utilized AUC, precision0, and recall1 as the evaluation metrics for model performance.

● AUC

AUC stands for the Area Under the ROC Curve, which can be used to assess the discriminatory power of a binary classification model. The larger the AUC value, the better the model's performance. Below is a formula for calculating the AUC.

$$AUC = \frac{\sum_{\mathrm{ins_i} \in \mathrm{positiveclass}} \mathrm{ranks}_{\mathrm{ins_i}} - \frac{\mathrm{M} \times (\mathrm{M}+1)}{2}}{\mathrm{M} \times \mathrm{N}} \quad (3)$$

During the AUC calculation process, after sorting the data from smallest to largest, the position of the first positive sample (starting from index 0) represents the number of times it scores higher than the negative samples. For the second positive sample, considering there's already one positive sample ahead, the number of negative samples is the position minus 1. Similarly, for the third positive sample, the number of negative samples before it is the position minus 2. For the (M^{th}) positive sample, the corresponding number of negative samples is its position number minus (M-1). Through this process, the numerator becomes the sum of all positive samples' position numbers, then subtract $(0 + 1 + 2 + \cdots + M - 1) = sum() - \frac{M \times (M-1)}{2}$.

Using the AUC metric, we can determine the model's predictive sensitivity and performance.

● precision0

precision0 represents the proportion of samples correctly predicted as negative among all samples predicted as negative. In this study, precision0 serves as an indicator to measure the extent to which conditions are mistakenly diagnosed as normal. The calculation formula is as follows.

$$\mathrm{Precision0} = \frac{\mathrm{True\ Negative}}{\mathrm{True\ Negative} + \mathrm{False\ Positive}} \quad (4)$$

The higher the precision0 value, the less frequently the model mistakenly judges patients as being in a normal condition.

● Recall1

recall1 refers to the model’s recall rate among positive cases, that is, the ratio of successfully predicted positive cases to all actual positive cases. In this study, recall1 is used to evaluate the degree to which each model fails to detect actual cases. The calculation formula is as follows.

$$Recall1 = \frac{Ture\ Positives}{True\ Positives + False\ Negatives} \quad (5)$$

When the model has fewer missed detections of actual cases, the value of recall1 will be higher. In the subsequent research, we will select the model with the highest recall1 and precision0 values through training and comparison.

4 EXPERIMENTAL SETUP AND RESULTS

4.1 Dataset Overview

The basis for this paper is the Stroke Prediction Dataset from Kaggle, which contains records of 5110 patients with various conditions. Each record comprises 12 attributes, as shown in Table 1 (Solorzano, 2020).

Table 1: Dataset attributes.

attribute	description
numbering	Unique identifiers
gender	Gender of the subject
age	Age of the subject
hypertension	0 means no hypertension and 1 indicates hypertension
heart disease	0 means no heart disease and 1 means heart disease
Ever married	Whether the subject is married or not
Job Type	Type of work
Type of residence	Subject's type of residence
Avg_glucose_level	The average level of glucose in the blood
Weight index	Weight divided by height squared （kg/m^2）
Smoking status	Subject's smoking status
Stroke	0 means no stroke and 1 means stroke

4.2 Experimental Setup

In this study, all algorithms and models were implemented in the IDE environment of PyCharm 3.8

(2018 edition), utilizing the Pandas and NumPy libraries for data processing of the stroke dataset, and sklearn, TensorFlow, xgboost, imblearn libraries for testing and evaluating machine learning models. Data visualization was carried out using the matplotlib and seaborn libraries.

The specific parameter settings for each model are as follows:

Logistic Regression Model: The regularization strength was set to 1.0, where a smaller regularization strength can enhance computational efficiency. The class weight was set to balanced by default. L2 regularization was adopted to include a penalty term to avoid overfitting.

Decision Tree Model: The splitting criterion for the decision tree was set to 'gini', which means using Gini impurity to choose the best splitting point. The maximum depth of the decision tree was set to 5. The minimum number of samples required to be at a leaf node was set to 1, and the minimum number of samples required to split a node was set to 2.

Random Forest Model: Similar to the aforementioned decision tree model, this random forest model also used 'gini' as the splitting criterion by default. The maximum depth of each decision tree was set to 5. The number of decision trees in the random forest was set to 100.

GBDT Model: The learning rate of the Gradient Boosting Decision Tree was set to 0.1. The loss function was set to 'log_loss', and the number of boosting trees was set to 100.

XGBoost Model: In this study, the objective function was chosen as the binary logistic regression function, with the maximum depth of each tree set to 5. The learning rate and the number of trees were not manually determined.

4.3 Model Evaluation

AUC, accuracy, precision of class 0, and recall of class 1 were used to evaluate the models used in the experiment. Results are shown in Figure 2 -Figure 5.

As illustrated in Figure 2, the Logistic Regression model and the Random Forest model achieved the highest AUC values, at 0.836 and 0.838, respectively. The Decision Tree model performed the worst, with an AUC value of 0.774. The AUC metric directly reflects a model's judgment and ranking abilities, hence the study concluded that the Logistic Regression and Random Forest models have better sensitivity and predictive performance based on the comparison.

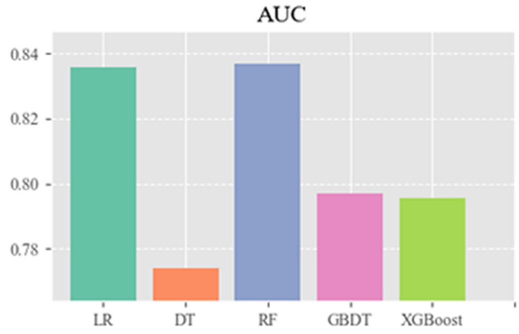

Figure 2: AUC for all 5 models (Photo/Picture credit: Original).

The accuracy of Logistic Regression and Random Forest is noticeably lower than that of GBDT and XGBoost (Figure 3). Generally, accuracy represents the correctness of a model's predictions in the test set; however, in cases of highly imbalanced data, the accuracy metric can become ineffective and misleading. For example, in the case of the dataset used in this experiment, if we design a simple algorithm that outputs "no stroke" regardless of the input parameters, this algorithm would lack any real predictive ability but could achieve an extremely high accuracy rate of 95.13%. Therefore, the judgment abilities of GBDT and XGBoost still require further investigation.

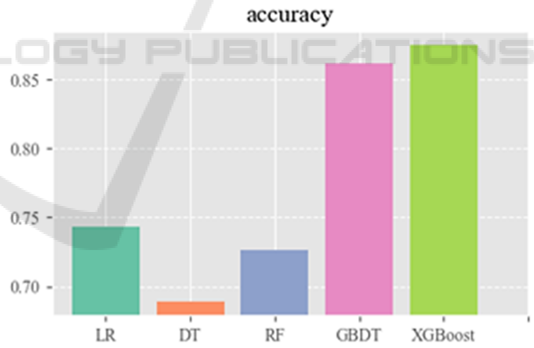

Figure 3: Accuracy of all 5 models(Photo/Picture credit: Original).

A higher precision0 indicates a lower probability of incorrectly predicting a subject as belonging to class 1 when they actually do not (i.e., a lower false positive rate for class 1) (Figure 4). Clearly, in stroke prediction, Logistic Regression and Random Forest models exhibit higher precision0 values, hence are more effective in avoiding the misclassification of subjects with stroke as not having a stroke. This capacity is crucial in medical diagnostics, where incorrectly identifying a patient's condition may result in no treatment being provided for a potentially serious condition, leading to increased risks for the patient. Therefore, in contexts where the cost of false positives is high—such as in medical diagnostic applications including stroke prediction—the precision of identifying negatives (precision0) is of paramount importance.

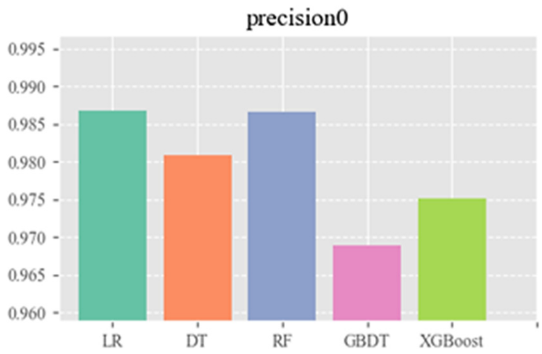

Figure 4: Accuracy of Class 0 (Photo/Picture credit: Original).

A higher recall1 indicates a model's capability to identify more subjects correctly as belonging to class 1 (in this case, individuals who have had a stroke). Evidently, in stroke prediction, both Logistic Regression and Random Forest models exhibit higher recall1 values, hence they are more efficient in comprehensively identifying individuals within the stroke-affected population. This ability ensures that fewer actual cases of stroke are missed, which is crucial for early intervention and treatment, potentially leading to better outcomes for patients.

On the other hand, the GBDT model exhibits poorer performance in this aspect. The lower recall1 value for GBDT suggests it may not be as effective in identifying all the true positive cases of stroke (Figure 5). This could lead to a higher number of stroke cases going undetected, a situation that is far from optimal in medical diagnostics where the early detection of conditions like stroke can significantly affect patient prognosis and treatment success.

Thus, while choosing a model for stroke prediction, it is essential to consider models that balance precision and recall effectively, aiming for models like Logistic Regression and Random Forest that demonstrate the capability to rightly identify individuals who have had a stroke, minimizing both false negatives and false positives.

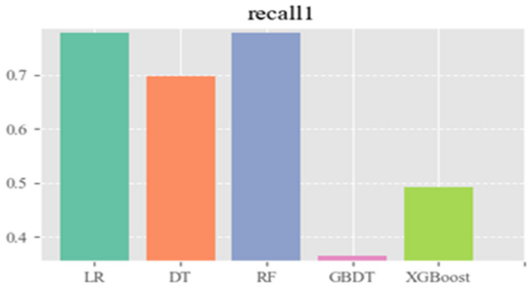

Figure 5: Category 1 recall (Photo/Picture credit: Original).

4.4 Analysis and Discussion

4.4.1 Discussion on the Superior Performance of Logistic Regression Model

Logistic Regression and Random Forest models achieved AUC values of 0.836 and 0.834, recall1 values of 0.778 and 0.746, and precision0 values of 0.987 and 0.985, respectively, with Logistic Regression proving more efficient. As a traditional classification model, Logistic Regression exhibited remarkable superiority in binary classification processing on this dataset, surpassing many advanced models. We attempt to analyze the reasons behind this.

- **Logistic Regression Model Excels at Handling Sparse Datasets**

Firstly, Logistic Regression's parameter estimation is based on the weights of non-zero feature values, making it inherently suitable for sparse datasets. During parameter estimation in Logistic Regression, only the weights corresponding to non-zero feature values are updated, while those for zero values remain unchanged.

Moreover, Logistic Regression can further promote sparsity through L1 regularization. L1 regularization incorporates an L1 norm penalty during model parameter estimation, pushing some feature weights towards zero, enabling feature selection and enhancing sparsity. For sparse datasets, L1 regularization helps the model automatically eliminate zero features that do not contribute to the predictive goal.

Lastly, Logistic Regression performs feature selection during model construction, automatically selecting features with strong predictive power for the target variable. For sparse datasets, with most feature values being zero, the model tends to choose non-zero features as predictors, enabling better modeling and prediction. Thus, Logistic Regression not only amplifies the influence of minority case data but also addresses the issue of weak feature selection to some extent.

According to the data, stroke incidence correlates significantly with age, a history of heart disease, and hypertension, with Pearson correlation coefficients of 0.23, 0.14, and 0.14, respectively. The data characteristics of hypertension and heart disease history, with $has : has\ not\ \approx 1 : 12$, are relatively sparse, fitting the criteria for a sparse dataset. The high performance of Logistic Regression in processing sparse datasets may be one reason for its outstanding performance on this dataset.

- **Logistic Regression Model Handles Independent Feature Column Datasets Well**

As indicated, the dataset's feature columns are relatively independent, with most Pearson correlation coefficients between features being small. Logistic Regression similarly excels at handling datasets with independent feature columns for the following reasons.

Coefficient estimates are accurate. Logistic Regression estimates feature coefficients using maximum likelihood estimation. When features are relatively independent, each feature's coefficient can be accurately estimated without interference or multicollinearity from other features.

Reduced variance inflation. When features exhibit collinearity or high correlation, Logistic Regression's coefficient estimates can become unstable and susceptible to variance inflation. Variance inflation refers to large variances in model coefficient estimates, making the model highly sensitive to small changes in input data.

The high efficacy of Logistic Regression in handling datasets with independent feature columns might be another reason for its superior performance on this dataset.

4.4.2 Discussion on the Poor Performance of GBDT Model

As mentioned, two features that significantly impact whether a stroke occurs—whether one suffers from hypertension and whether one has a history of heart disease—have relatively sparse positive and negative distributions in the dataset, affecting the GBDT model. GBDT, a tree-based classification method, struggles with sparse datasets since it may select these sparse features as split points while constructing trees. The presence of missing values complicates determining optimal split points, impacting model accuracy. Additionally, GBDT optimizes the model gradually through gradient boosting, adjusting sample weights with each tree's training. For sparse features, where most values are zero, their contribution might be underestimated, leading the model to overlook these features in favor of others with weaker correlations, thereby affecting predictive performance.

5 CONCLUSION

This study, through comparative analysis of applications on a specific dataset, investigated the performance of various machine learning models, including Logistic Regression, XGBoost, and Decision Trees, in predicting stroke. In evaluating model performance, we employed measures such as AUC, precision, and recall for a comprehensive assessment. The research findings demonstrate that the Logistic Regression model exhibited exceptional performance across these metrics in the dataset used.Furthermore, the research explored how handling imbalanced datasets, model tuning, and feature selection could enhance the precision and stability of machine learning model predictions. Experimental results indicate that micro-level adjustments, such as parameter tuning, do not significantly contribute to model improvement.

Detailed analysis of the stroke prediction models revealed that one of the primary reasons for the superior performance of the Logistic Regression model is its highly efficient handling of sparse datasets, along with its proficiency in managing datasets with relatively independent feature columns.

In summary, the Logistic Regression model showcased the best performance in this stroke prediction application, attributed to its efficient handling of sparse datasets and adaptability to datasets with relatively independent feature columns. This discovery underscores the importance of considering dataset characteristics when selecting an appropriate machine learning model for disease prediction. Future research could further explore the application of Logistic Regression models in predicting other types of diseases and also suggests attempting to combine the Logistic Regression model with other machine learning models to achieve higher prediction accuracy and performance.

AUTHORS CONTRIBUTION

All the authors contributed equally and their names were listed in alphabetical order.

REFERENCES

Boukhennoufa, I., et al. 2022. *Predicting the internal knee abduction impulse during walking using deep learning.* Frontiers in Bioengineering and Biotechnology, 10, 877347.

Chakraborty, P., Bandyopadhyay, A., Parui, S., Swain, S., Banerjee, P. S., Si, T., Mallik, S., Alfurhood, B. S., Al-Betar, M. A., Almomani, A. 2024. *OptiSelect and EnShap: Integrating Machine Learning and Game Theory for Ischemic stroke prediction.*

Murphy, S.J.X., & Werring, D.J. 2020*. Stroke: causes and clinical features.* Medicine, 48(9), 561-566*.*

Han, C., Lian, G. 2023. *Optimized Prediction Model for Imbalanced Samples Based on Stroke Data.* Journal of Shanxi Datong University (Natural Science Edition), (03), 31-35.

Li, Y. 2016. *Statistical Analysis of Factors Influencing the Onset of Stroke.* Journal of Chengdu University of Technology, 19(02), 49-52.

Silva, A. de B., Rigo, S. J., & Barbosa, J. L. V. 2018. *Examining Developments and Applications of Wearable Devices in Modern Society* . 134.

Solorzano, F. *Stroke Prediction Dataset*. Kaggle, 2020. Available at: https://www.kaggle.com /fedesoriano /stroke-prediction-dataset.

Song, X., et al. 2022. *Activities of daily living-based rehabilitation system for arm and hand motor function retraining after stroke.* IEEE Transactions on Neural Systems and Rehabilitation Engineering, 30, 621-631.

Wu, X., & Yang, B. 2022. *Ensemble Learning Based Models for House Price Prediction, Case Study: Miami, U.S.* In 2022 5th International Conference on Advanced Electronic Materials, Computers and Software Engineering 449-458.

Zhang Y 2021, *Prediction of long-term stroke recurrence using machine learning models*. DOI: 10.3390/JCM10061286.