A machine learning-based model for stroke prediction

Dongchen Wu[1], Xinfang Zhang[1,*] Xiaochen Zhu[2]

[1]School of Biomedical Engineering, Northeastern University, Shenyang, 110167, China
[2]School of Bioengineering, Qingdao University, Qingdao, 266071, China

[*]20227273@stu.neu.edu.cn

Abstract. In recent years, machine learning has highlighted good results in the early diagnosis and prediction of diseases. Stroke is a serious threat to human health. Early prediction of stroke is of great significance for treatment and intervention. This paper mainly investigates the application of different machine learning models in stroke prediction and compares the performance of each model. First, we collected some data, which contained 5110 entries or records and 12 different attributes. The dataset was then subjected to various preprocessing measures, such as eliminating data redundancy. Seven different machine learning models were used. Includes Logistic Regression, Support Vector Classifier (SVC, K-Neighbors Classifier, Decision Tree Classifier, Random Forest Classifier, XGBoost Classifier (XGBClassifier) and Deep Neural Networks (DNN). After model comparison, we found that when the dataset was extremely imbalanced, the AUC of DNN before feature selection was 82%, which was significantly better than other machine learning models. In addition, the characteristics of each model were analyzed to provide a reference for the selection of stroke prediction models. The results of this study provide value for the early diagnosis and intervention of stroke, and also provide new ideas for the application of machine learning algorithms in the field of stroke prediction.

Keywords: Machine learning, Deep learning, Stroke, Diagnosis, Prediction.

1. Introduction

Stroke is the second leading cause of death worldwide. The estimated cost of stroke worldwide is over US $721 billion, which was equivalent to 0.66% of global GDP [1]. The Global Burden of Disease (GBD) 2019 stroke burden estimates provide valuable insights into the alarming rise in the lifetime risk of stroke. These estimates reveal that over the past two decades, the likelihood of experiencing a stroke has surged by a staggering 50%. This means that in today's world, one out of every four individuals are affected by this debilitating condition at some point in their lives [2]. The prevalence and incidence of stroke have increased significantly in people under the age of 70. These findings highlight the importance of stroke control programs and the need for updated data on stroke incidence, prevalence, and mortality rates for effective prevention and control strategies.

The incidence of stroke is quite high. Each year, more than 12.2 million new stroke patients are reported globally [2]. Furthermore, 1 in 4 people over 25 are predicted to have a stroke at some point in their entire lives [2]. Secondly, stroke has a high mortality rate with six and a half million deaths each year, a number that is rapidly increasing. The narrow treatment window for stroke is the main cause of this high mortality rate. Treatment must be administered within a few hours after the interruption of blood flow supply to minimize brain tissue damage. Moreover, stroke has a high disability rate and significant impact. Stroke-induced brain damage can result in various neurological abnormalities, such as paralysis, speech disorders, and cognitive decline. Finally, the recurrence rate of stroke is high due to the enhanced dependence on blood supply in the area of vascular injury. Health can be adversely affected by unhealthy lifestyles such as smoking, drinking excessive alcohol and eating unhealthy foods and lack of exercise, are all risk factors for stroke recurrence. These habits not only increase the risk of stroke but may also accelerate the process of relapse.

Many factors are involved in the development of stroke. Every patient is different. As a result, predictive models may not be 100% accurate in predicting whether a stroke will occur. But with more data and better algorithms, the accuracy and reliability of predictive models will continue to improve. The study will use this data to compare different models. The majority of the aforementioned studies have focused on the performance of various models under conditions of balanced datasets. However, given that in real-world scenarios, data pertaining to strokes are predominantly imbalanced, our research delves into the performance of various machine learning models under conditions of data imbalance.

The aim of this study is to predict stroke, we chose machine learning for prediction. Machine learning is a core area of artificial intelligence (AI). The core of machine learning is to build a model that can learn from data. This model automatically improves its ability to predict or make decisions through continuous learning and adjustment. Both training and prediction of machine-learning models depend on data. The quality, quantity and diversity of the data have an important impact on the performance of the model. Models are the heart of machine learning: they learn from data and make predictions. Model selection and design are critical to the success of machine learning projects. According to each model, we predict that deep neural networks (DNN) have the best performance in the case of imbalanced datasets.

2. Literature Review

The advent of maximum procedures has led to a shift in non-invasive methods for auxiliary medical systems in stroke prevention. Stroke victims may still be survived if they are identified early enough, according to the research. One benefit is that bioelectrical pictures may be used to predict strokes, which is advantageous since non-invasive methods are highly useful in diagnosing strokes and can be better predicted at an early stage [3]. In 2020, stroke was made more predictable by using machine learning in the early prediction stage. In the beginning, people who wanted to prevent strokes utilized standard neural networks to extract variables from structured data, including blood pressure, heart rate, gender, age, and hypertension. By utilizing ML techniques in 2019, the stroke predictor (SPR) model improved its prediction accuracy to 96.97% [4]. The SPN approach is also beneficial since it analyses the risk levels associated with strokes using an improvised random forest. In 2022, a number of models are created and assessed in the study using machine learning (ML) in order to provide an effective structure for long-term danger prediction of stroke development [4]. A study's primary contribution is a stacking technique that performs well and is supported by a number of measures, including accuracy, precision, recall, F-measure, and AUC. The experiment's findings demonstrated that, with an accuracy of 98%, an AUC of 98.9%. The stacking classification works better than the other approaches [5]. Somya Srivastav examined and contrasted the results of numerous algorithms for machine learning in 2023. By using the "Brain stroke dataset" and the "Support Vector Machine (SVM)" technique for prediction, Biswas, N. attained 90% accuracy [4]. Employing the "Logistic regression" technique with the "Stroke dataset," the proposed model has a 95.02% accuracy rate [6]. With an accuracy output of 95.02%, the analysis shows that the logistic regression classifier performed better than other methods for predicting strokes. By expanding the dataset, this classifier may be further improved to produce even better and more useful results [7]. In the present study, to deal with imbalanced data, we proposed a DNN-based model for stroke prediction and compared its performance with that of other classical machine learning methods.

3. Method

3.1. The introduction of Dataset

The dataset utilized for predicting occurrences of brain stroke was sourced from "KAGGLE". It was provided in the format of a "comma-separated values" (CSV) file. This investigation made use of the "Stroke Prediction Dataset" which is a relatively small collection comprising 5110 entries or records and 12 different attributes [8]. Within this dataset, there are 249 records of brain stroke identified as "Yes" indicating confirmed cases, and 4861 instances labeled as "No", indicating the absence of a stroke. The ratio of positive to negative samples is shown in Figure 1. As can be seen, the dataset is highly imbalanced. The dataset features several variables: identifier (id), gender, age, presence of hypertension, existence of heart disease, marital status (ever married), type of employment (work type), type of residence (Residence type), average glucose level, body mass index (BMI), smoking habits (Smoking status), and stroke occurrence. In addition, this dataset provides additional test sets which consists of 2044 entries.

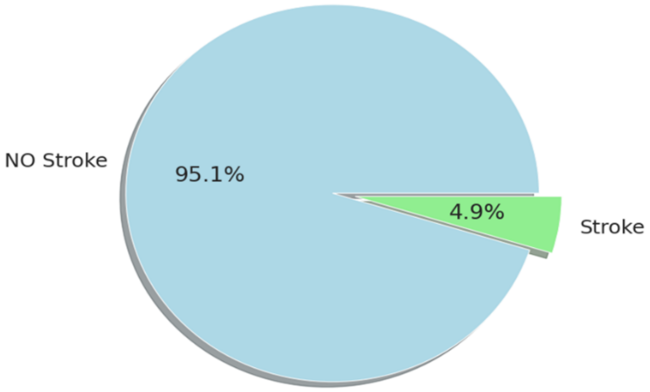

Figure 1. The ratio of positive to negative samples.

3.2. Preprocessing

3.2.1. Data cleaning

The dataset contains numerous duplicates, missing values, or features irrelevant to the prediction goal. Considering the practical situation and model performance, we have made the following adjustments: First, we removed the "ID" column to reduce dataset complexity and interference from irrelevant factors. Second, due to varying personal beliefs, the "gender" column includes "other" entries. Given their low proportion in the dataset and the prediction goal, we chose to delete all rows containing "other." Third, the "BMI" column has 201 missing values, representing 4% of the total. Considering its impact on model prediction, we opted for imputation [9] to fill in the "BMI" column.

3.2.2. Data Transformation

There are several columns in this dataset with features that are difficult to handle. To ensure the model runs smoothly, we performed type conversion on the following features. First, "Smoking status" has three categories: "never smoked," "formerly smoked," and "unknown." Since "unknown" accounts for 30% of the total values and cannot be directly deleted, we chose to apply one-hot encoding to this feature. Second, the "age" in this dataset is represented as a floating-point number. To reduce dataset complexity, we converted it to integer data. Finally, we applied one-hot encoding to the columns "residence type," "work type," and "ever married."

3.3. The processing of the dataset

Firstly, the dataset was divided into training and validation datasets with a ratio of 8:2 to be subjected to 5-fold cross-validation. The main purpose is to pick the model that performs best on the validation set. In addition, this dataset provides additional test sets. In the presence of severe data imbalance, we initially employed various machine learning models for prediction. To enhance the performance of each machine learning model, we further filtered the data based on Pearson's correlation coefficient, selecting features with a correlation higher than 0.1 (Figure 2). Subsequently, we applied the same procedure using DNN for prediction. Finally, the model classifies and predicts the test set of the dataset.

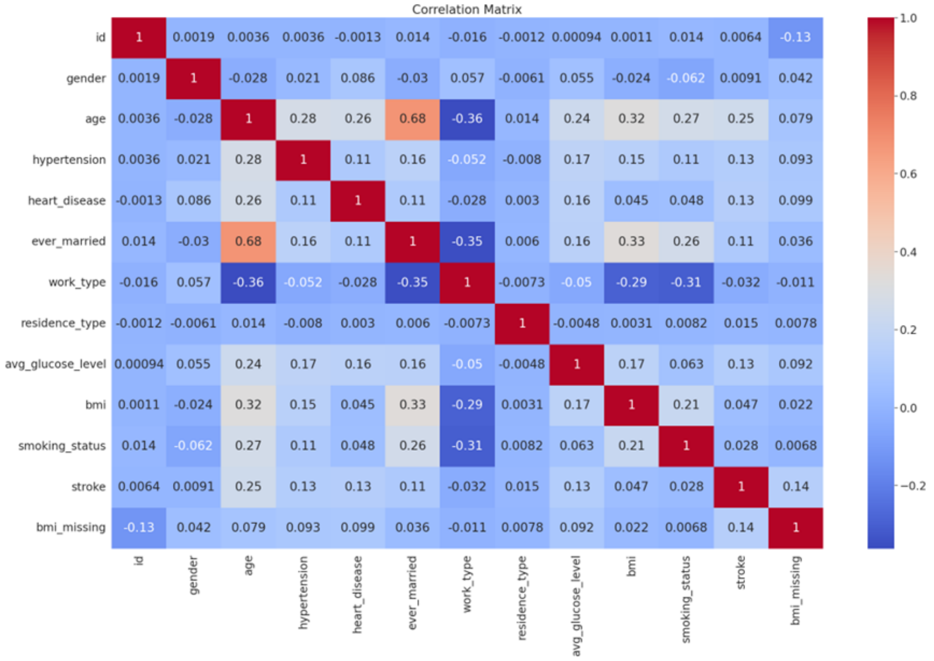

Figure 2. The correlation matrix.

3.4. Models

Seven models were selected for stroke prediction including K-Neighbors Classifier, Logistic Regression, Support Vector Classifier (SVC), Random Forest Classifier, Decision Tree Classifier, XGBoost Classifier (XGBClassifier) and DNN.

Deep neural network, also referred to as a multilayer perceptron (MLP), is a multilayer unsupervised neural network. Its design is inspired by the perceptron, but in the DNN, multiple hidden layers are introduced, making it more complex. The layers in a DNN are fully connected to each other, as in the perceptron. Neurons in each layer are connected to neurons in the next layer. Although the network may appear large and complex, its principle is like that of the perceptron when focusing on a small part of it [10]. In deep learning neural networks, the neurons in each layer will perform a certain nonlinear transformation of the input data and pass the output to the next layer. Through continuous iterative training, the neural network can optimize the weight parameters, thereby improving the modeling and classification accuracy of the data. Its advantages include automated feature extraction, flexibility, and high performance.

The logistic regression model assumes a specific data distribution and uses maximum likelihood estimation to estimate its parameters. It is a linear model commonly used for classification problems. The logistic regression model assumes that the data follows a Bernoulli distribution. The gradient descent method is used to calculate the parameters by maximizing the likelihood function. Logistic regression models are preferred due to their simplicity, parallelization, and interpretability [11]. The support vector machine classification algorithm is a two-class model that aims to find the linear classifier with the maximum interval on the feature space. SVMs can process both linearly separable and non-separable data. The SVM is a suitable method for classifying small and medium-sized data samples, as well as non-linear and high-dimensional problems. The K-Neighbors Classifier, a well-established and straightforward machine learning algorithm. Its concept is intuitive: if a sample's K nearest neighbors in the feature space belong to a certain category, then the sample also belongs to that category. The K-Neighbors Classifier is better suited for automatically classifying domains with a large sample size, whereas those with a small sample size are more prone to generating inaccurate classifications [12].

A predictive model that maps the relationship between object attributes and their values called the decision tree. The decision tree model has several advantages. It is easy to understand and interpret, can handle discrete or continuous data types, and can be used for classification and regression problems. Decision tree models have several disadvantages, including their tendency to overfit, sensitivity to small data changes, high training costs, instability, and suboptimal solutions [13]. Random forest is a classifier that comprises multiple decision trees. The output class is determined by the mode of the output class of individual trees, and it belongs to the Bagging type. By combining multiple weak classifiers, the result is voted or averaged, resulting in high accuracy, generalization performance, and good stability [14]. XGBoost is a machine learning system that is scalable, portable, and accurate. It has pushed the limits of machine learning computation, running more than ten times faster on a single machine than popular solutions at the time. It can even process billions of data in distributed systems.

3.5. Experimental Environment and Hyperparameter

The experiments in this paper were conducted on the Kaggle platform. We used two types of graphic processing units (GPUs) available on the platform: (NVIDIA P100 GPU 16GB and NVIDIA Tesla T4 GPU16GB), with a server processor Intel(R) Xeon(R) @ 2.00GHz, and default configurations for memory, storage, and other devices. We implemented the experiments using Python3.7.12 on the TensorFlow 2.15.0 deep learning framework, and conducted the experiments on Ubuntu 20.04.5 LTS.

The DNN model used the Adam optimization algorithm. Binary crossentropy was used as the loss function. The model employed the ReLU activation function for the first two dense layers and the sigmoid activation function for the output layer. The training process run for 20 epochs. Each batch during training consisted of 50 samples. The input shape was determined by X_train.shape [1], which is the number of features in the training data. A random seed of 1 was set for TensorFlow's random number generator to ensure reproducibility. These hyperparameters, along with the model architecture, play crucial roles in the model's performance on tasks such as accuracy, F1 score, recall, and AUC metrics. Other machine learning models used the default parameters of the relevant functions in scikit-learn.

3.6. Evaluation Metrics

To better evaluate the performance of various models on the "stroke" dataset, the following metrics are often used for evaluation:

$$Accuray = \frac{\mathrm{T}P + TN}{TP + TN + FP + FN} \tag{1}$$

$$Recall = \frac{TP}{TP + FN} \tag{2}$$

TP (True Positives) stands for instances correctly classified as positive while TN (True Negatives) are those accurately classified as negative. P (False Positives) represent instances wrongly classified as positive; FN (False Negatives) indicate instances wrongly classified as negative.

$$\mathrm{F1} - score = 2 \times \frac{Precision \times Recall}{Precision + Recall} \tag{3}$$

Precision refers to the proportion of true positive samples in those predicted to be positive, while Recall indicates the proportion of actual positive samples in the predicted positive samples relative to all positive samples in the dataset.

$$\mathrm{AUC} = \int_0^1 \mathrm{TPR3}(t)\ dt \tag{4}$$

AUC stands for Area Under the Curve. TPR(t) represents the True Positive Rate at threshold t, and dt represents a small increment for integration. This formula integrates over the range of 0 to 1, which corresponds to the full range of the ROC curve's x-axis.

4. Results

Considering the importance of accurately assessing model performance and the rigor in predicting disease outcomes. The performance metrics used for comparison were the AUC of the Receiver Operating Characteristic (ROC) and accuracy on the test set.

Before feature selection, the accuracy of various models ranged between 91%-94%, yet the AUC performance was markedly poor, with the best decision tree model achieving only a 59% AUC (Figure 3). After feature selection, there was a slight improvement in all metrics for all models, but the AUC remained significantly low. The machine learning models were still unable to adequately distinguish between positive and negative cases (Figure 4). Subsequent analysis using DNN revealed results as shown in Table 1. Under conditions of extreme dataset imbalance, the AUC of DNN before feature selection, at 82%, was substantially higher than that of other machine learning models. It was hypothesized that DNN performance would improve after feature selection; however, the AUC was slightly lower than before feature selection. Prior to feature selection, the precision of the various models was observed to be between 91%-94%, but the AUC performance was notably poor, with the best performing decision tree model only reaching an AUC of 59%. After feature selection, there was a marginal improvement across all evaluation metrics for all models, yet the AUC remained critically low. The machine learning models continued to struggle with effectively distinguishing between positive and negative instances. Following this, an analysis using DNN was conducted, with the results presented in Table 1. In the context of extreme dataset imbalance, the AUC for DNN prior to feature selection, at 82%, significantly outperformed that of other machine learning models. It was initially hypothesized that post-feature selection, the performance of DNN would enhance; however, the AUC slightly decreased compared to the pre-feature selection phase.

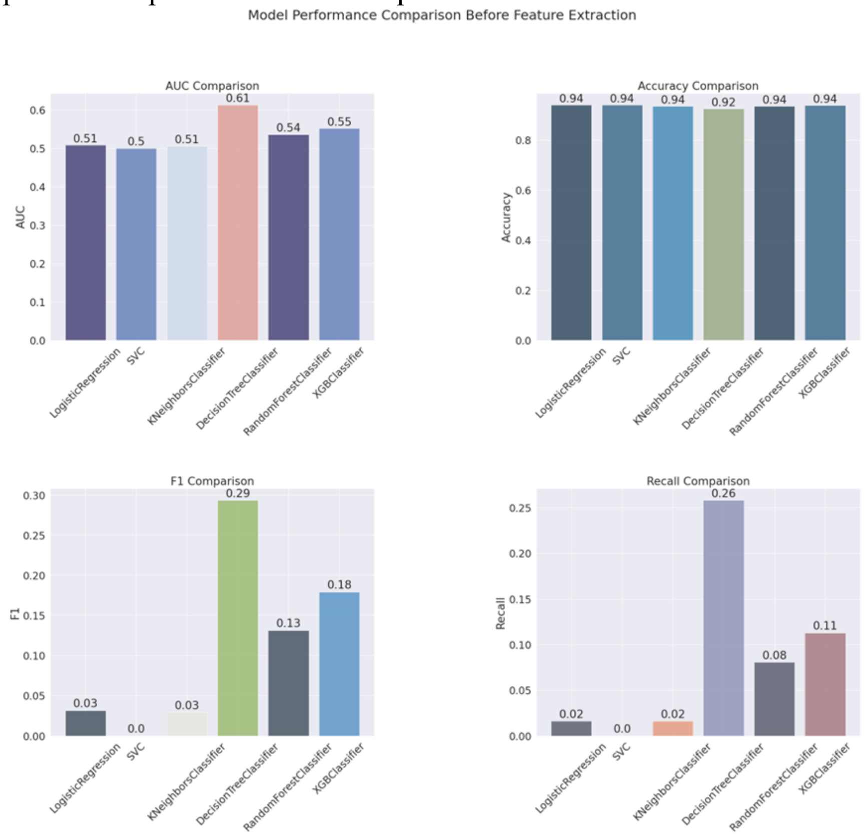

Figure 3. Model accuracy and AUC comparison before feature extraction.

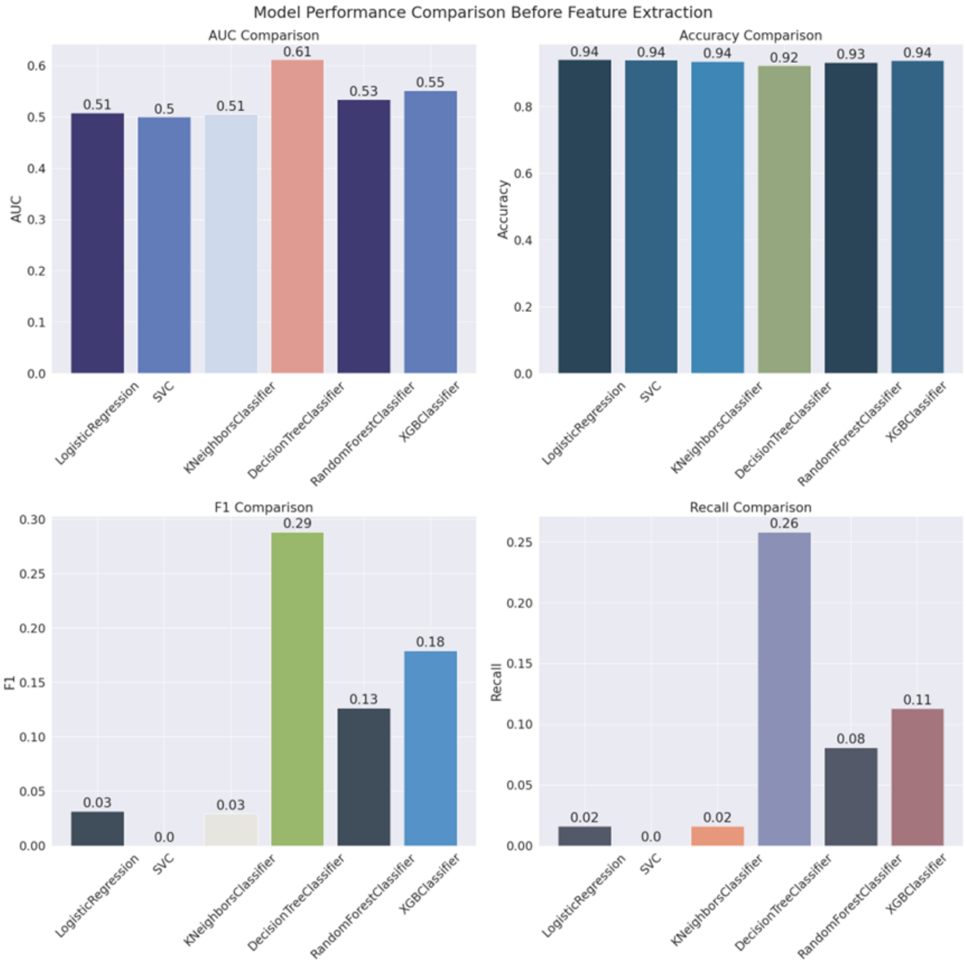

Figure 4. Model accuracy and AUC comparison after feature extraction.

Table 1. The results of DNN.

Metric	Before Feature Extraction	After Feature Extraction
Accuracy	0.9393	0.9393
AUC	0.8005	0.7218

5. Discussion

Compared with the traditional machine learning, DNN had more ability to model or abstract the representation of things and simulate more complex models because it had more layers of results. There were three main advantages. First, it's the strong learning ability. From training, testing to validation, DNN performed very well. Wide coverage and good adaptability are the second advantage. Its neural network had many layers and a very wide width, which can theoretically be mapped to any function, and many complex problems could be solved. Third, DNN is data-driven, with a high ceiling. DNN is highly dependent on data, and the larger the amount of data, the better its performance. In the image recognition, facial recognition, NLP and other parts of the task, its performance even exceeds the human, and could also be adjusted to further improve his upper limit.

From the perspective of comparing DNN with other machine learning models, under conditions of extreme dataset imbalance, the ANN's AUC of 82% and 79% was significantly higher than that of other machine learning models. Looking at it from the perspective of machine learning models excluding DNN, the decision tree model's AUC of 59% and 60% was higher than that of ensemble learning models like Random Forest and XGBClassifier. Comparing the machine learning models excluding DNN before and after feature selection, the AUC of DNN decreased from 82% to 79% after feature selection.

For comparing machine learning models excluding DNN, the decision tree model outperforms ensemble models like Random Forest and XGBClassifier, feature extraction in the context of data imbalance might impact the ability of deep learning models to distinguish between positive and negative instances [15]. As for the AUC of DNN decrease after feature selection, although no specific research had been found to explain this, it might be due to decision trees being prone to overfitting, especially when they were very deep. However, on certain datasets, a well-tuned decision tree might just fit the distribution of the data accurately, whereas ensemble models might lose performance in trying to reduce overfitting. Aside from these two issues, other experimental results largely aligned with our initial hypotheses.

6. Conclusion

This study focuses on the successful use of seven machine learning techniques for stroke prediction using imbalanced datasets. Data preprocessing and model comparison analyses were performed on extremely imbalanced data. The experimental results indicated that DNN outperformed other models with an AUC of 82% before feature selection. The implications of this study were important for stroke prevention and diagnosis. In the future, we will continue to optimize our project in many ways. We can use more detailed and comprehensive data to train the model and improve its applicability in selecting data sets. Additionally, we can optimize the model at a deeper level to further enhance its performance.

Authors Contribution

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

References

[1] Feigin V. L, Brainin M, Norrving B, et al. 2022 World Stroke Organization (WSO): Global Stroke Fact Sheet International Journal of Stroke 17 18-29
[2] Feigin V L, Stark B A, Johnson C O, Roth G A, Bisignano C, Abady G G and Hamidi S 2021 Global, regional, and national burden of stroke and its risk factors, 1990-2019: a systematic analysis for the Global Burden of Disease Study The Lancet Neurology 20 795-820
[3] Fernandez-Lozano C, Hervella P, Mato-Abad, V. et al. 2021 Machine Learning-Based Model Can Predict Stroke Outcome Scientific Reports 11 10071
[4] Vamsi B, Bhattacharyya D, Midhunchakkaravarthy D 2021 Detection of Brain Stroke Based on the Family History Using Machine Learning Techniques (Springer, Singapore)
[5] Dritsas E, Trigka M 2022 Stroke Risk Prediction with Machine Learning Techniques Sensors (Basel) 22 4670
[6] Santos, L. I., Camargos, M. O., D'Angelo, M. F. S. V., Mendes, J. B., deMedeiros, E. E. C.,Guimarães, A. L. S. and Palhares, R. M. 2022 Decision tree and artificial immune systems for stroke prediction in imbalanced data Expert Systems with Applications 191 116221
[7] S Srivastav, K Guleria and S. Sharma 2023 Machine Learning Models for Early Brain Stroke Prediction: A Performance Analogy World Conference on Communication & Computing 1-6
[8] Fedesoriano 2021 Stroke Prediction Dataset Kaggle Machine Learning Repository https://www.kaggle.com/datasets/jillanisofttech/brain-stroke-dataset
[9] A T Sree Dhevi 2014 Imputing missing values using Inverse Distance Weighted Interpolation for time series data Sixth International Conference on Advanced Computing 255-9
[10] Miikkulainen R, Liang J, Meyerson E, et al. 2024 Evolving deep neural networks Artificial intelligence in the age of neural networks and brain computing 269-287
[11] Lukman A F, Kibria B M G, Nziku C K, et al. 2023 K-L Estimator: Dealing with Multicollinearity in the Logistic Regression Model Mathematics 11 340
[12] Dinata R K, Adek R T, Hasdyna N, et al 2023 K-nearest neighbor classifier optimization using purity AIP Conference Proceedings 2431

[13] Zaitseva E, Rabcan J, Levashenko V, et al. 2023 A new method for analysis of Multi-State systems based on Multi-valued Decision Diagram under epistemic uncertainty Applied Soft Computing 134 109988

[14] Alice K, Deepa N, Devi T, et al. 2023 Importance analysis of decision making factors based on fuzzy decision trees Measurement: Sensors 25 100566

[15] H. He and E. A. Garcia 2009 Learning from Imbalanced Data IEEE Transactions on Knowledge and Data Engineering 21 1263-84