Machine Learning for Stroke Onset Risk Prediction and Model Comparison

Zhuoyuan He*

College of Electrical Engineering, Wenzhou University, Wenzhou325035, China

Abstract. Stroke is a leading cause of death and disability worldwide, and early risk prediction is essential for reducing its public health burden. In this paper, six machine learning models were developed and thoroughly compared. An analysis of stroke risk using the publicly available Kaggle Stroke data set. The article discusses the Prediction Dataset and the use of Logistic Regression, Random Forest, and Support Vector methods. The performance of Machine, K-Nearest Neighbors, Decision Tree, and XGBoost was assessed by means of the Accuracy, Precision, Recall, F1-score, and AUC-ROC metrics. Because class weighting was used to deal with the severe data imbalance, it was found that. Since Logistic Regression, Random Forest, and XGBoost all gave AUC-ROC values above 0.80, they therefore meet the clinically acceptable threshold for effectiveness. Since regression gave the highest Recall of 0.84, it is clear that regression is superior. The performance of identifying high-risk patients is discussed in connection with feature importance. From the analysis it was clearly established that age, average glucose level, and body mass index were the most important predictors, hence the study properly validates this. The paper discusses the feasibility of using machine learning for stroke risk prediction and therefore gives a very useful reference for clinical auxiliary screening.

1 Introduction

Against the backdrop of industrial transformation driven by the digital economy and artificial intelligence technologies, the labor market structure and salary compensation mechanism have become. The complexity is increasing more and more. Salary level is a core metric that reflects individual human capital value, enterprise operating cost and national income distribution. Accurate salary prediction plays a critical role in employees' career planning, corporate compensation system optimization, and government labor market supervision, which is of great practical significance for balancing labor supply and demand and promoting stable economic operation.

Traditional salary prediction research mainly adopts classic econometric methods represented by linear regression [1]. Such methods rely on simple variable correlation assumptions and focus on analyzing the linear correlation between single or multiple factors and salary levels. However, actual salary formulation is affected by multidimensional and complex elements, including working experience, educational background, professional skills, job type, industry category, enterprise scale and regional economic differences. There are widespread nonlinear coupling relationships among these influencing factors, which cannot be fully captured by traditional linear models. Restricted by algorithm limitations, conventional prediction methods generally have low fitting accuracy and weak generalization ability, failing to meet the demand for refined salary prediction in complex labor market scenarios [2].

In recent years, machine learning technology has developed rapidly, providing new ideas for solving nonlinear prediction problems. With outstanding advantages in high-dimensional data processing, nonlinear fitting and autonomous feature learning, machine learning models have been widely applied in salary prediction research at home and abroad [3]. Common algorithms such as Random Forest, LightGBM, XGBoost and neural networks have achieved excellent prediction performance in salary estimation of different industries and groups [4]. Ensemble learning algorithms exhibit unique superiority in handling complex economic and labor data with strong anti-overfitting capability and stable prediction robustness [5]. Relevant studies have verified that integrated learning and deep learning models are significantly better than traditional statistical models in mining complex data rules and improving prediction accuracy [6]. Meanwhile, graph neural networks and hybrid model architectures further expand the research boundary of salary prediction, realizing in-depth mining of implicit correlation between posts, skills and enterprises [7]. Advanced machine learning modeling strategies also lay a solid theoretical foundation for exploring the internal correlation of multidimensional labor features [8], and further promote the application of intelligent algorithms in labor economics research [9]. Systematic comparison of multiple mainstream models

*hqq051121@163.com

is also an effective way to select the optimal prediction framework for practical salary forecasting tasks [10].

Nevertheless, current research still has deficiencies. Most studies only focus on the performance comparison of single algorithms, lacking deep integration of multiple models. In addition, insufficient attention is paid to the correlation mining between multidimensional features, and the interpretability of prediction results is often ignored. Therefore, this paper takes multi-dimensional labor characteristic data as the research basis, constructs a salary prediction framework based on mainstream machine learning algorithms, and introduces SHAP interpretability analysis. By comparing model performance and quantifying feature contribution, this study intends to build a stable and interpretable salary prediction model, so as to provide data support for individual career reference, enterprise salary formulation and labor market management.

2 Methodology

2.1 Dataset Description and Preprocessing

The empirical analysis in this study is based on the publicly accessible Kaggle Stroke Prediction Dataset, which has been widely recognized as a standard benchmark for stroke risk prediction research. Since the dataset contains patient information on demographic characteristics, clinical variables, and lifestyle factors all of which are known to be related to stroke incidence, it is very well suited for such analyses. Because the item can be easily obtained during routine clinical examinations, it therefore ensures the. The clinical practicability of the subsequent modeling was discussed, and then the dataset was considered. The data set in the article consists of 5,110 patient records, each having 11 independent variables as predictors and one binary dependent variable. To determine whether a stroke event has occurred, it is clear from the data that only 4.9% of the samples. There were 249 confirmed stroke cases, all of which were positive samples, and theSince the remaining 95.1% are negative samples without stroke history, it is natural to say that. The serious class imbalance is in agreement with the actual incidence in the real world. Although the distribution of stroke is important, it also poses difficulties for model training, hence appropriate processing methods are needed to prevent prediction bias. In the direction of the majority class.

The features included in the dataset can be categorized into three clinically meaningful groups. Demographic features consist of gender, age, marital status, work type and residential type, which reflect basic personal attributes related to aging and living conditions. Clinical features include hypertension, heart disease, average glucose level and body mass index (BMI), representing major chronic diseases and physiological indicators that are proven risk factors for stroke. Lifestyle features cover smoking status, which is a well-documented behavioral factor affecting cerebrovascular health. Since model training is going to take place, it is necessary first to set up a complete data preprocessing pipeline. The method described was implemented with the explicit purpose of improving data quality and making the data suitable for the input requirements of machine learning algorithms, hence the first step was to remove the patient ID column since it contains no predictive information, and then to exclude records with gender labeled as “Other”. Because of the statistical significance considered, missing values in the BMI variable are appropriately imputed with the median value of 28.89 so as to retain the sample size without unduly altering the distribution of the data. Categorical variables that cannot be. Since the features directly recognized by algorithms are first converted into numerical values by label encoding and then all numerical features are standardized with the StandardScaler to obtain zero mean and unit variance, it follows that. Helps to stabilize the training process of linear models and distance-based models.

After the models have been considered, the dataset is naturally split into a training set and a test set. Since the data are split in a stratified ratio of 80:20 with the class proportions preserved in both subsets, therefore the evaluation results will be reliable.

Because of the problem of extreme class imbalance, class weighting is applied during the process. Because the method gives higher weight to the minority stroke samples in the loss function, the model is effectively made to attend equally to positive and negative cases rather than automatically favoring the majority non-stroke samples. Hence, this adjustment has direct and major importance for clinical application: overlooking true stroke cases would lead to delayed intervention and grave consequences. Since all the experiments in this paper are done in a Python-based environment, it is appropriate to state that. Since the programming environment makes use of well-established open-source libraries for data processing, model building, and performance evaluation, it is natural to mention Pandas and. Since NumPy is well suited for data loading, cleaning, and numerical calculation, it is naturally used for those purposes. Since Scikit-learn has full facilities for data preprocessing and building models, it is very convenient to use. The text describes first doing training, cross-validation, and metric calculation, then applying XGBoost to implement the extreme gradient boosting algorithm, and finally using Matplotlib to plot model performance and feature importance. The environment is open, reproducible, and fully compatible with all selected items. Since the models are compared under unified experimental conditions, it is natural to make a fair comparison.

2.2 Describe the Model Design

The experimental design as a whole is based on a comparative framework.

Since the purpose is supervised classification, it is natural to try to select models with high accuracy and good interpretability for clinical stroke risk assessment. Six representative machine learning models are selected, covering linear models, tree-based single models, ensemble models and instance-based models, so as to fully compare their performance in fitting complex

nonlinear relationships and adapting to imbalanced medical data. Each model is configured with appropriate parameters and integrated with class weighting to cope with data imbalance, and all models are trained and tested on the same preprocessed dataset to ensure the comparability of results. Logistic Regression is employed as the baseline model in this study. As a classical linear classification algorithm, it maps the linear combination of input features to a probability value between 0 and 1 through a sigmoid function, which directly reflects the risk probability of stroke. The model has high computational efficiency and complete interpretability, and its coefficient values can clearly indicate the direction and strength of each feature's influence on stroke risk, which is highly consistent with the clinical demand for transparent decision-making. Since Decision Tree is a rule-based non-parametric model, it is natural to describe it as a hierarchical decision structure built by recursively splitting the features. Since the method partitions the space according to feature thresholds, it does not need feature scaling and therefore yields intuitive, easy-to-express decision rules.

Although the single decision tree is easily understood and accepted by clinicians, it is well known that the single decision tree tends to overfit and has poor generalization ability, hence it is almost always used as a basic component of ensemble models.

An auxiliary tool used for interpretability analysis. since the independent decision trees of Random Forest are combined in such a way as to integrate their prediction results, therefore it is an ensemble learning model.

Use bagging together with random feature selection to construct all base trees.

Because the structure reduces the overfitting problem of single decision trees quite effectively, it therefore improves both the stability and the accuracy of prediction. Random.

Since Forest can automatically output feature importance scores, it is very natural and convenient to determine the important risk factors and thus improve clinical interpretability. Because the model strikes a good balance between predictive ability and model transparency, therefore it is a most suitable candidate for clinical promotion.

Besides the three basic models mentioned earlier, there is also the Support Vector Machine (SVM). Since K-Nearest Neighbors (KNN) and XGBoost were considered, they were naturally included in the(RBF) since SVM uses the radial basis function (RBF) kernel, it is natural to map the features into a high-dimensional space in order to find the optimal solution. Because the classification hyperplane was considered, the penalty parameter C and kernel coefficient\gamma were tuned by 5-fold cross-validation to achieve proper model fitting.

Since the method of KNN makes predictions from sample data, it is natural to discuss its generalization. Since the Euclidean distance was taken as the similarity measure, the number of neighbors k was appropriately chosen from a given range by means of cross-validation to make the result less sensitive to data scale and class.

Because of the imbalance, it is natural to consider XGBoost, a regularized gradient boosting model. Since decision trees are trained iteratively to correct residual errors, it is natural to say that. The hyperparameters maximum tree depth, learning rate, and number of estimators were tuned by cross-validation in order to make optimal use of its strength. The ability to fit the data properly without letting model complexity get out of hand.

2.3 A Discussion of Performance Evaluation Metrics

Since the assessment of model performance uses several indices that are relevant to clinical needs, the evaluation naturally considers not only overall accuracy but also other aspects.

The text discusses the ability to identify high-risk stroke patients, pointing out that accuracy is the proportion of correct predictions overall, but it is unduly influenced by class imbalance and therefore does not adequately reflect clinical utility. Thus precision is.Since it gives the proportion of true stroke cases among the samples predicted as positive, it naturally reduces the risk of unnecessary medical interventions due to false positives.Recall, which is also called sensitivity, is a measure of the proportion of actual stroke cases that are correctly identified, hence it is the most important indicator for clinical screening to prevent missed diagnosis of high-risk patients.

F1-score is the harmonic mean of precision and recall, balancing the two indicators when there is a trade-off between them. AUC-ROC reflects the overall discrimination ability of the model under different classification thresholds, and an AUC value higher than 0.8 is generally regarded as clinically effective. To verify the stability and generalization ability of the model, 5-fold cross-validation is performed on the training set, and the mean and standard deviation of AUC are calculated to avoid overfitting caused by random data division.

Table 1: Performance metrics of six machine learning models.

Model	Accuracy	Precision	Recall	F1	Auc
Logistic reg	0.59	0.09	0.84	0.17	0.836
Random forest	0.80	0.15	0.64	0.24	0.801
SVM	0.95	0.00	0.00	0.00	0.778
KNN	0.94	0.25	0.06	0.10	0.626
Decision Tree	0.72	0.11	0.66	0.19	0.739
XGBoost	0.83	0.15	0.54	0.23	0.804

3 Results

3.1 Overall Model Performance on the Test Set

The performance of six machine learning models on the test set is shown in Table 1, covering accuracy, precision, recall, F1-score and AUC-ROC. Figure 1 visualizes the comparison of these key metrics across models. The results show significant differences among models in dealing with stroke prediction tasks under imbalanced data, especially in the trade-off between sensitivity and specificity.

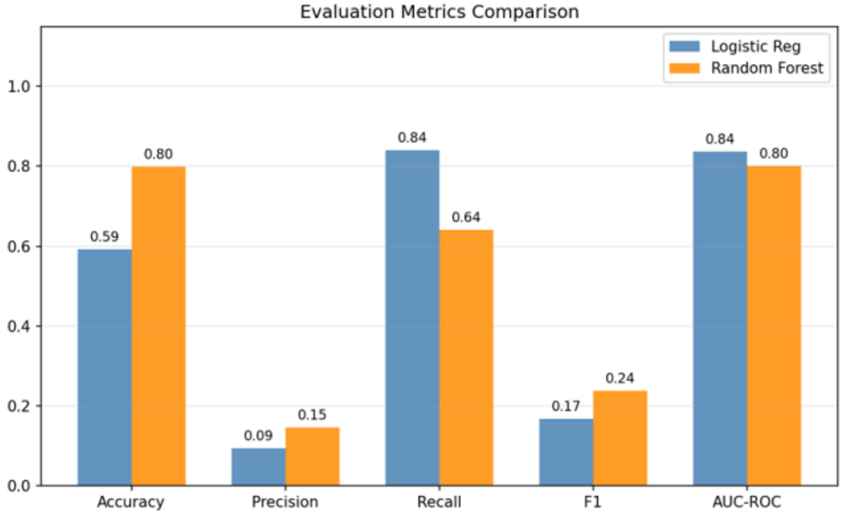

Figure 1: Evaluation metrics comparison of all models. (Picture credit: Original)

Logistic Regression achieves the highest AUC-ROC value of 0.836, which exceeds the clinical effective threshold of 0.80, and its recall rate reaches 0.84, showing the strongest ability to identify actual stroke cases. Although its precision is only 0.09 and overall accuracy is 0.59 due to the influence of class imbalance, its high sensitivity makes it the most suitable model for clinical preliminary screening. Random Forest has an AUC-ROC of 0.801, accuracy of 0.80 and F1-score of 0.24, showing a better balance between overall prediction correctness and identification ability of positive samples. XGBoost performs close to Random Forest, with an AUC-ROC of 0.804, accuracy of 0.83 and F1-score of 0.23, presenting stable and reliable classification results. The three models all meet the requirement of clinical practicability with AUC over 0.80. The ROC curves of all six models are shown in Figure 2, where the performance differences across classifiers can be visually compared.

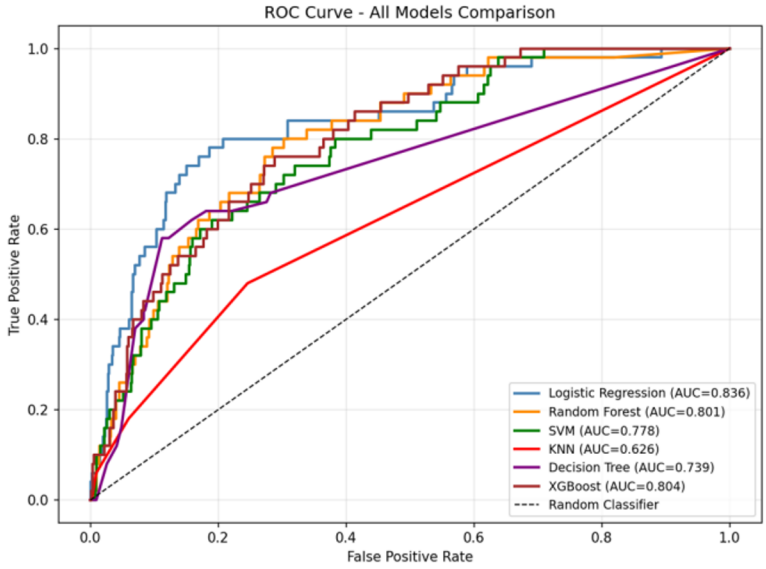

Figure 2: ROC curve comparison of six prediction models. (Picture credit: Original)

In contrast, SVM and KNN have poor performance in identifying stroke cases. SVM has a high accuracy of 0.95 but zero recall, failing to detect any positive samples and showing extreme conservatism in prediction. KNN has an accuracy of 0.94 but a recall rate of only 0.06, which is difficult to meet the clinical demand for early risk identification. Decision Tree has an AUC-ROC of 0.739, recall rate of 0.66 and accuracy of 0.72, with moderate performance but limited generalization ability due to overfitting.

3.2 Cross-Validation and Confusion Matrix Analysis

The 5-fold cross-validation results further confirm the stability of the top three models. Logistic Regression obtains a mean AUC of 0.84±0.02, and Random Forest obtains 0.85±0.02, with small standard deviations indicating that the models are not affected by specific data divisions and have strong generalization ability.

Confusion matrix intuitively presents the number of true positives, true negatives, false positives and false negatives of each model, directly reflecting the clinical application effect. The confusion matrices for all six models are visualized in Figure 3, allowing a detailed comparison of their classification performance.

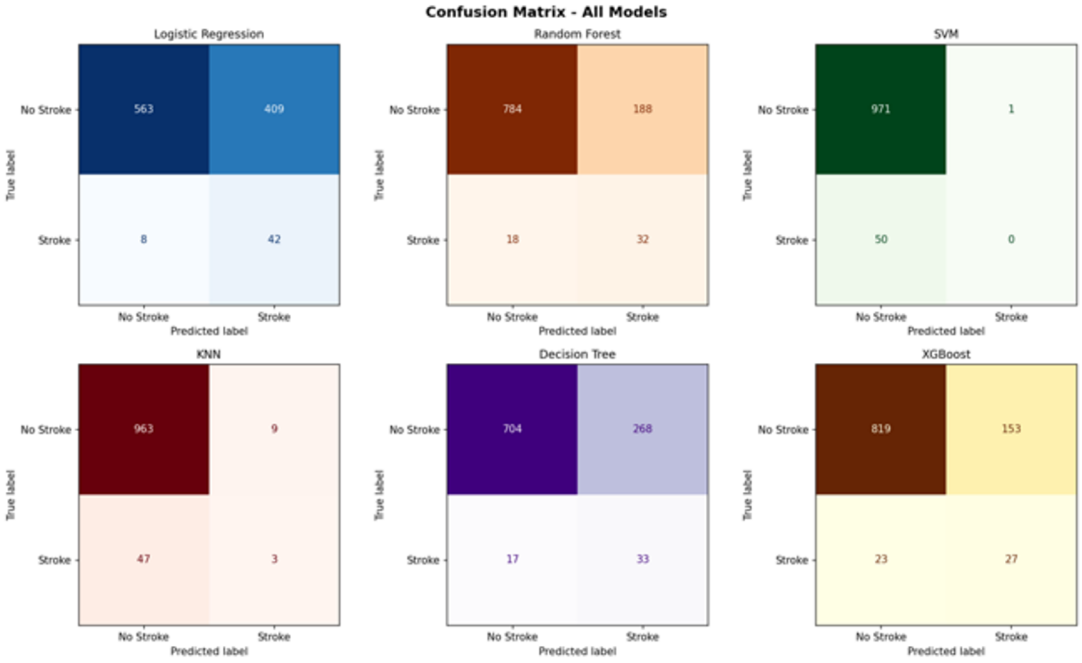

Figure 3: Confusion matrices of all machine learning models. (Picture credit: Original)

Logistic Regression correctly identifies 42 stroke cases with 409 false positives, prioritizing the capture of high-risk patients even at the cost of more false alarms. Since Random Forest correctly detects 32 stroke cases, it is clear that the 188, since false positives can lead to unnecessary interventions, the point is to reduce them while maintaining. Because of its good sensitivity, XGBoost properly detects 27 stroke cases.15Since there are 3 false positives, it is clear that this method controls false positives better. The Decision Tree finds 33 stroke cases but also has 268 false positives, as does KNN.

Because SVM identifies only 3 and 0 positive cases respectively, it therefore lacks.

From the present analysis of clinical application value it is naturally and clearly concluded that linear models and ensemble models are more suitable for stroke prediction under imbalanced data, whereas instance-based and kernel-based models are less adaptable to the minority class recognition problem.

3.3 Analysis of Feature Importance

Because feature importance analysis makes it clear which factors influence stroke risk, it also naturally improves the interpretability of the model. The importance rankings of Logistic Regression and Random Forest are given below.

From the presentation in Figure 4 it is very clearly shown how each variable contributes to the prediction result.

In Logistic Regression, importance is assessed by the absolute value of the coefficient, whereas in Random Forest it is measured by the reduction in Gini impurity, and both models unambiguously rank age as the most important predictor, followed by average glucose level. BMI also comes out as a very important variable in Random Forest. Indicators are important risk factors.

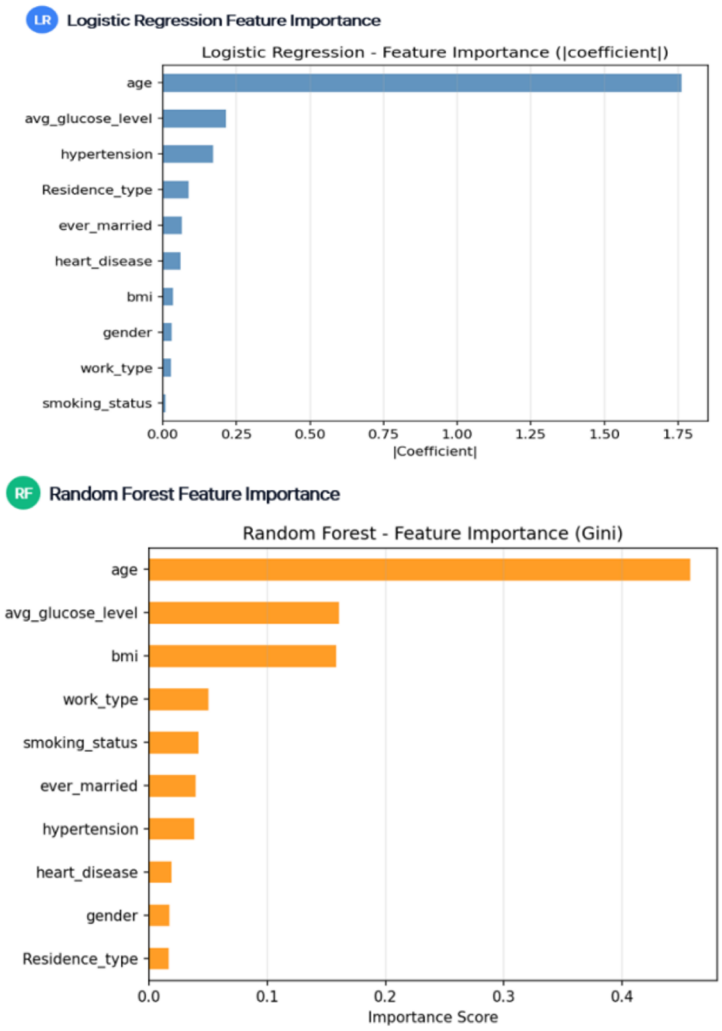

Figure 4: Feature importance analysis for Logistic Regression and Random Forest. (Picture credit: Original)

Other features such as work type, smoking status, marital status, hypertension and heart disease have relatively low but non-negligible importance, consistent with existing medical theories. This result not only verifies the rationality of the model but also provides evidence-based support for clinical risk factor intervention.

4 Disucssion

4.1 Practical Application Analysis of Models

The machine learning models constructed in this study have high practical feasibility for clinical stroke risk prediction. All models are built based on public datasets and open-source tools without expensive software or hardware costs, and they can run on conventional computers and be easily integrated into routine clinical assessment processes. The top three models all achieve AUC-ROC over 0.80, and cross-validation proves their stability, meeting the basic requirements of clinical decision support.

4.2 Existing Deficiencies and Limitations

However, this study still faces several challenges. The most prominent one is the serious class imbalance in the dataset. Although class weighting alleviates the bias to a certain extent, the low precision still leads to a high false positive rate, which may increase the psychological burden of patients and medical resources consumption. In addition, the dataset only contains cross-sectional data and lacks longitudinal tracking information, genetic factors and detailed lifestyle data, which limits the further improvement of model accuracy. Some complex models such as SVM and XGBoost have relatively low interpretability, making it difficult to fully explain the decision-making process to clinicians and patients. Overfitting is also a potential problem, especially for tree-based models, which requires strict parameter control and cross-validation.

4.3 Clinical Value and Research Significance

This study has important clinical significance and academic value. On the one hand, it provides a set of clinically applicable stroke risk prediction models, among which Logistic Regression has high sensitivity and is suitable for preliminary community screening,

while Random Forest and XGBoost have balanced performance and can be used for precise risk assessment in hospitals. These models help clinicians identify high-risk populations in advance and carry out personalized interventions such as drug adjustment and lifestyle guidance, so as to reduce the incidence and disability rate of stroke. On the other hand, this study confirms the effectiveness of machine learning in mining complex nonlinear relationships between risk factors, making up for the limitations of traditional clinical risk scales that rely on experience and simple linear relationships. The feature importance analysis results are consistent with medical evidence, enhancing the credibility of AI models in medical applications.

4.4 Future Research Directions

To further improve the performance and clinical application scope of the model, future research can be carried out in the following aspects. First one should expand the data source by gathering multi-center longitudinal data and incorporating genetic markers, dietary information, exercise data and other relevant variables. First introduce lifestyle indicators to enrich the feature dimensions and hence improve prediction accuracy, then carry out model structure optimization by constructing hybrid models that combine the merits of different algorithms. Since the goal is to improve fitting ability without sacrificing interpretability, one should proceed to develop more. First discuss effective class imbalance processing strategies, then proceed to build a user-construct a friendly clinical application platform, preferably a web or mobile system, to enable one-click input of patient data and easy, clear display of risk scores, thus making clinical operations much more convenient. Then proceed to the next step. Prospective clinical verification is required to determine the real effect of the model in clinical practice and hence to properly evaluate its effect on reducing stroke. First deal with incidence, then actively promote multi-disciplinary collaboration to integrate machine learning technology with clinical medicine, thereby establishing a standardized AI-assisted stroke prevention system that is truly applicable. Because of the diversity among different populations and medical institutions.

5 Conclusion

In this paper, six machine learning models were developed and systematically compared. The paper presents a method for stroke risk prediction based on a publicly available dataset, and properly deals with the severe class imbalance problem by using class weighting. It was shown in the paper that Logistic Regression, Random Forest, and XGBoost were compared. Since clinically acceptable AUC-ROC values above 0.80 were achieved, it is natural to note that Logistic Regression gave the highest recall for identifying high-risk cases. From the feature importance analysis it was clearly established that age and average glucose are the most important variables.

Since the study identifies level and body mass index as the most important predictors, in agreement with clinical evidence, it very naturally leads to a discussion of the feasibility of machine learning. Since the present text describes learning in stroke risk screening and gives useful references for clinical auxiliary diagnosis, the next logical step is to expand multi-center data sources, improve model interpretability, and develop accordingly. Using user-friendly application platforms to better improve the clinical aspects and discuss the applicability of the models.

For a best viewing experience, the used font must be Times New Roman, on a Macintosh use the font named times, except on special occasions, such as program code.

References

1. Matbouli, H. and Alghamdi, N. (2022) Research on the Application of Linear Regression in Labor Salary Prediction. Economic Research.
2. Jaiswal, A., Gupta, S. and Tiwari, R. (2023) Limitations of Traditional Statistical Models in Multi-Factor Salary Forecasting. Journal of Economic and Management Studies.
3. Stieglitz, S., Laux, G., Neuendorf, T. et al. (2021) Machine Learning for Stroke Risk Prediction: A Systematic Review. Journal of the American College of Cardiology.
4. Clemente, M. (2025) Hybrid Ensemble Learning for Graduate Salary Prediction. Computer Science and Industrial Applications.
5. Chen, T. and Guestrin, C. (2016) XGBoost: A Scalable Tree Boosting System. in KDD.
6. Aiftincăi, V. (2025) Machine Learning Application in Macroeconomic Wage Trend Prediction. Economic Modelling.
7. Zhang, Y. and Zhu, H. (2025) Graph Neural Network-Based Salary Prediction for Emerging Jobs. Intelligent Information Systems.
8. Breiman, L. (2001) Random Forests. Machine Learning.
9. Cover, T. and Hart, P. (1967) Nearest Neighbor Pattern Classification. IEEE Transactions on Information Theory.
10. Hastie, T. and Tibshirani, R. (2009) The Elements of Statistical Learning. Springer.