Dean&Francis
ISSN 2959-6157

Stacked Ensemble with Voting-Classifier for Stroke Prediction

Tianhao Yang[1, *]

[1]School of Internation, Beijing University of Posts and Telecommunication, Beijing, 100000, China
*Corresponding author: pro845009@outlook.com

Abstract:

Stroke is one of the major causes of death nowadays. It occurs rapidly and has severe sequelae, seriously affecting the quality of life of patients. Therefore, early identification of high-risk individuals and intervention can effectively prevent strokes. Machine learning technology has shown great potential in the field of disease prediction. Many medical fields have begun to widely use machine learning to assist in diagnosis. However, there are numerous machine learning algorithms, making it difficult to choose, and different algorithms perform differently on different problems. A single model is unable to balance the advantages and disadvantages of performance to make more accurate predictions. Therefore, this study aims to develop a stacked ensemble model with Voting-Classifier as the meta-model to combine the advantages of models that perform well on this problem, and use real data as the training set to identify high-risk individuals. Moreover, the model performance should be more accurate and more robust than that of a single model.

Keywords: Machine learning; stacked ensemble model; Voting-Classifier; prediction.

1. Introduction

According to data from the World Health Organization (WHO), approximately 7 million people died from stroke in 2021, accounting for more than 10% of total global deaths (68 million), making it the third leading cause of mortality worldwide. Furthermore, as a prevalent cerebrovascular disorder, stroke is often associated with severe sequelae, including hemiplegia, facial paralysis, aphasia, and dementia [1]. These sequelae will significantly reduce the patient's self-care ability and increase the likelihood of them developing psychological problems [2]. However, identifying high-risk individuals and implementing preventive measures in advance could potentially prevent stroke or reduce its severity. Therefore, the effective prediction of stroke holds significant value in its prevention and early intervention.

After many years of development, today's technology has undergone tremendous changes, especially in the field of artificial intelligence. Machine learning technology, as a part of artificial intelligence, has successfully introduced new approaches into the medical field. By analyzing, modeling, and computing large datasets, it can reveal hidden patterns and holds significant promise for applications such as disease prediction and assisted diagnosis [3]. In fact, typical classification algorithms, including decision

trees, are widely utilized in disease prediction and auxiliary diagnosis [4]. For instance, machine learning has already demonstrated its application value in the early diagnosis and screening of esophageal cancer [5]. However, the reliability and accuracy of the algorithms pose a major challenge for the application of machine learning in the medical field [6]. Therefore, it is of vital importance to develop reliable and accurate algorithms.

To accurately predict the occurrence of stroke, it is essential to consider well-established risk factors such as age, gender, genetic predisposition, hypertension, cardiovascular disease, smoking status, and body mass index (BMI) [7]. However, an excessive number of variables not only complicates the predictive process but also increases the cost of clinical assessments. Therefore, it is necessary to develop a predictive approach that maintains high accuracy while selecting only those factors that are easily measurable and universally applicable across populations.

In the process of feature selection, reference can be made to the Shi and Liu's study which utilized Least Absolute Shrinkage and Selection Operator (Lasso) regression to identify key features that contribute effectively to stroke prediction [8]. By proactively filtering out non-informative features, this method significantly reduces redundant computations, expedites model convergence, and minimizes the adverse effects of noise on model performance.

The selection of an appropriate algorithm plays a critical role in developing effective predictive models, as different machine learning algorithms exhibit distinct advantages and limitations. In their comprehensive study comparing seven prominent algorithms - including random forest, decision tree, K-nearest neighbor, adaptive boosting, gradient boosting, logistic regression, and support vector machine - for stroke prediction, Begum demonstrated that random forest consistently outperformed other methods, achieving superior overall performance [9].

Begum's research demonstrated the superior performance of the random forest algorithm for this specific prediction task. However, models based on a single algorithm may exhibit performance instability. In contrast, ensemble algorithms typically offer stronger learning capabilities and greater stability compared to individual models [10]. Therefore, this study employs a stacked ensemble model to optimize the prediction model.

Building upon prior research, this work introduces a stacked ensemble framework that integrates heterogeneous base models through a meta-learner. The hierarchical architecture not only consolidates the strengths of individual algorithms but also mitigates their weaknesses, resulting in a system with superior robustness, generalizability, and capability to process high-dimensional feature interactions without compromising reliability.

2. Methods

2.1 Data Source

The data used in this paper is the Brain stroke prediction dataset on Kaggle, which is processed from the Brain stroke prediction dataset by Izzet Turkalp Akbasli's various methods. The original data set is collected by the Electronic Health Record (EHR) controlled by McKinsey Company, but the BMI in the data set has NAN value, and simple processing will introduce a lot of noise. Therefore, the dataset processed by Izzet Turkalp Akbasli is also selected for the smooth progress of the research.

2.2 Data Introduction

There are over 5,000 entries in this dataset, each with 11 attributes, which are described in Table 1. The author wanted to take advantage of these properties, try to use a stacked ensemble approach to build a model to predict the likelihood of stroke, and see if the performance of the model can surpass that of a single model.

Table 1. Variable Attributes

Attribute	Type	Range
gender	Categorical	Male or Female
age	Numeric	0-82
hypertension	Categorical	0: no hypertension, 1: have hypertension
heart_disease	Categorical	0: no heart diseases 1: have a heart disease
ever_married	Categorical	No or Yes
work_type	Categorical	"children", "Govt_jov", "Never_worked", "Private" or "Self-employed"
Residence_type	Categorical	"Rural" or "Urban"
avg_glucose_level	Numeric	55.1-272
bmi	Numeric	14-48.9
smoking_status	Categorical	"formerly smoked", "never smoked", "smokes" or "Unknown"

ISSN 2959-6157

2.3 Data Preprocessing

Before training, this paper needs to do some preprocessing on the data. The author found that there is a "Unkown" value in smoking_status, which means that the patient's information is not available, and smoking doubles the risk of stroke [7]. So the author decided to discard all the "Unkown" data. At the same time, the attribute values in the data set are not all numerical types, so the label coding is used for the data before training to cope with this situation. However, this is not enough, because this is an extremely imbalanced dataset. The stroke=1 in the dataset only makes up about 5% of the total dataset, and the proportion rises to 6% by dropping all the "Unkown". Therefore, many studies have employed the Synthetic Minority Over-sampling Technique - Edited Nearest Neighbors (SMOTE-ENN) to handle the data. SMOTEENN is a technique that combines SMOTE oversampling and ENN undersampling to generate minority samples and clean up noisy samples to improve the overall quality of the data. However, since SOMTE relies on the K-nearest neighbor algorithm to synthesize the minority class, merely using ENN to clean the synthesized data would still lead to an inflated performance data of the KNN model. Therefore, the authors adopted the Synthetic Minority Over-sampling Technique - Weighted Edited Nearest Neighbor (SMOTE-WENN) to handle the dataset. This method can retain more overlapping regions of positive examples while avoiding the excessive deletion of useful samples.

2.4 Method Introduction

To predict strokes, the author chose to construct a stacked ensemble model, which is an advanced ensemble learning method. This method takes the prediction results of multiple basic models as a new dataset, and then uses this dataset to train a meta-model for the final prediction. This allows us to combine the strengths of different models and ultimately improve the overall performance. The base models of Stacking Ensemble Model constructed by the author are Random Forest (RF), eXtreme Gradient Boosting (XGBoost) and Extra Trees (ET). The reason for choosing these three models is simple: RF has been proven to perform well in predicting stroke, XGBoost provides a different perspective while performing well, and ET model outperforms most of the models in the author's study.

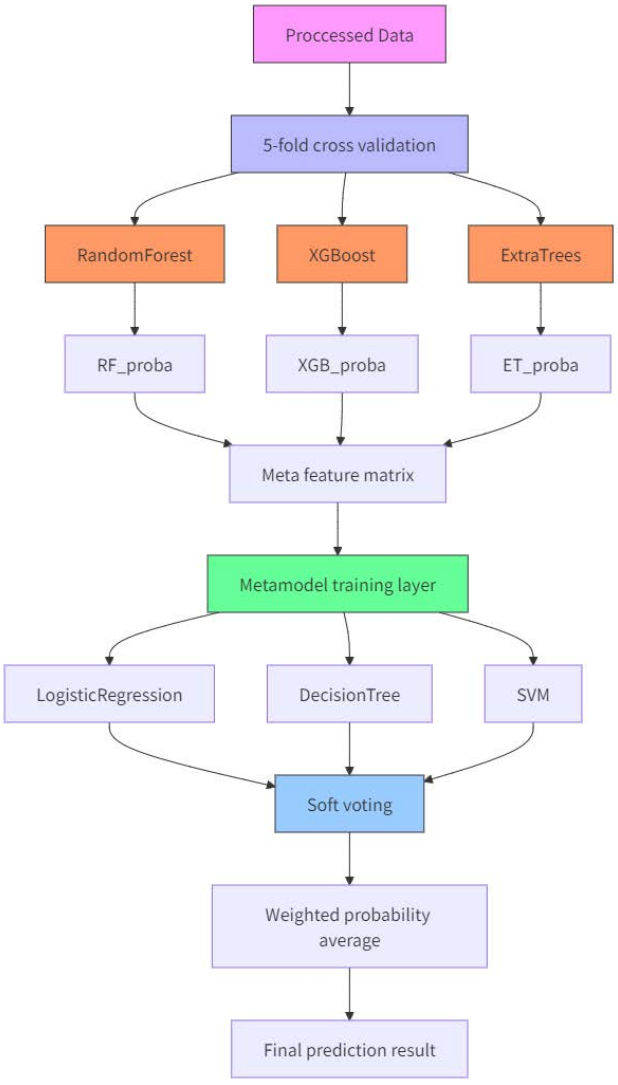

Fig. 1 Flow Chart (Picture credit: Original)

Instead of choosing a single model for the meta-model, the authors chose Voting-Classifier as the meta-model. In Voting-Classifier, Logistic Regression (LR), eXtreme Gradient Boosting (XGBoost) are used to perform soft voting together to increase the robustness and anti-noise ability of the model. The operation flow chart of the model is shown in Figure 1.

3. Results and Discussion

3.1 Evaluation Metrics

In order to better demonstrate and compare the performance of the model, the author employed metrics such as Area Under the ROC Curve (AUC), Average Precision (AP), Precision, Recall, F1-Score, and Accuracy. These metrics are all related to stroke (Accuracy represents the overall accuracy of the model in predicting No Stroke and Stroke). Among these metrics, the author will place

greater emphasis on AUC and AP, as AUC can reflect the comprehensive performance of the model, and in the case of data imbalance, AP (calculated from the area under the Precision-Recall Curve) is a more superior metric for model evaluation compared to AUC [9].

3.2 Base Models Performance

In order to ensure that the final stacked ensemble model performs well, the authors trained multiple models to find the appropriate basic models. The author evaluated eight models: Decision Tree (DT), Random Forest (RF), Logistic Regression (LR), Support Vector Machine (SVM), Extreme Gradient Boosting (XGBoost), Naive Bayes (NB), K-Nearest Neighbors (KNN), and Extra Trees (ET). Their performance is shown in Table 2.

Table 2. Model Performance Comparison(With STOME-WENN)

Model	AUC	AP	Precision	Recall	F1	Accuracy
DT	0.903	0.836	0.861	0.937	0.897	0.901
RF	0.989	0.988	0.914	0.969	0.941	0.944
LR	0.873	0.833	0.732	0.81	0.769	0.775
SVM	0.832	0.733	0.701	0.825	0.758	0.757
XGBoost	0.988	0.984	0.891	0.982	0.934	0.936
NB	0.852	0.799	0.716	0.815	0.762	0.765
KNN	0.961	0.92	0.822	0.993	0.9	0.898
ET	0.986	0.986	0.905	0.957	0.93	0.934

The data in the table shows that the AUC and AP metrics of the three basic models (RF, XGBoost, and ET) have performed exceptionally well. Moreover, their accuracy is also outstanding among these models. Therefore, the author chose these three models as the basic models for the stacked ensemble model. Coincidentally, all these three models are based on decision tree models. The feature importance of these models can be observed, and the feature importance of XGBoost is different from the other two models, providing a different perspective for the final stacked ensemble model (Figure 2, 3 and 4).

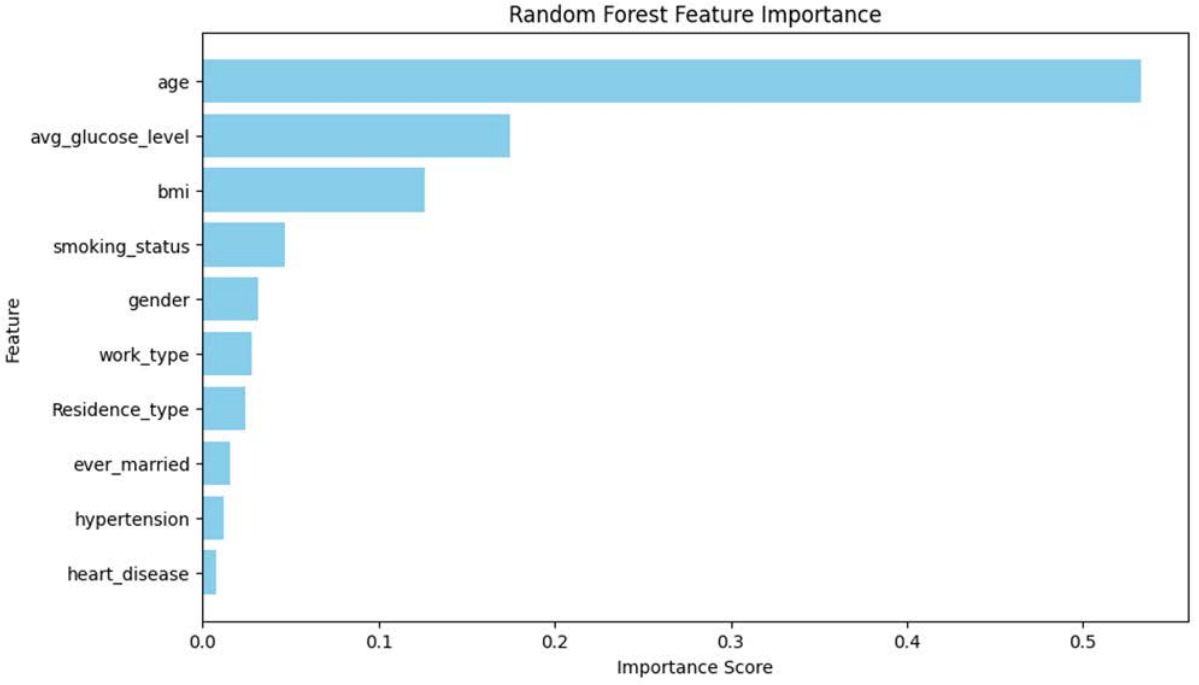

Fig. 2 Random Forest Feature Importance (Picture credit: Original)

ISSN 2959-6157

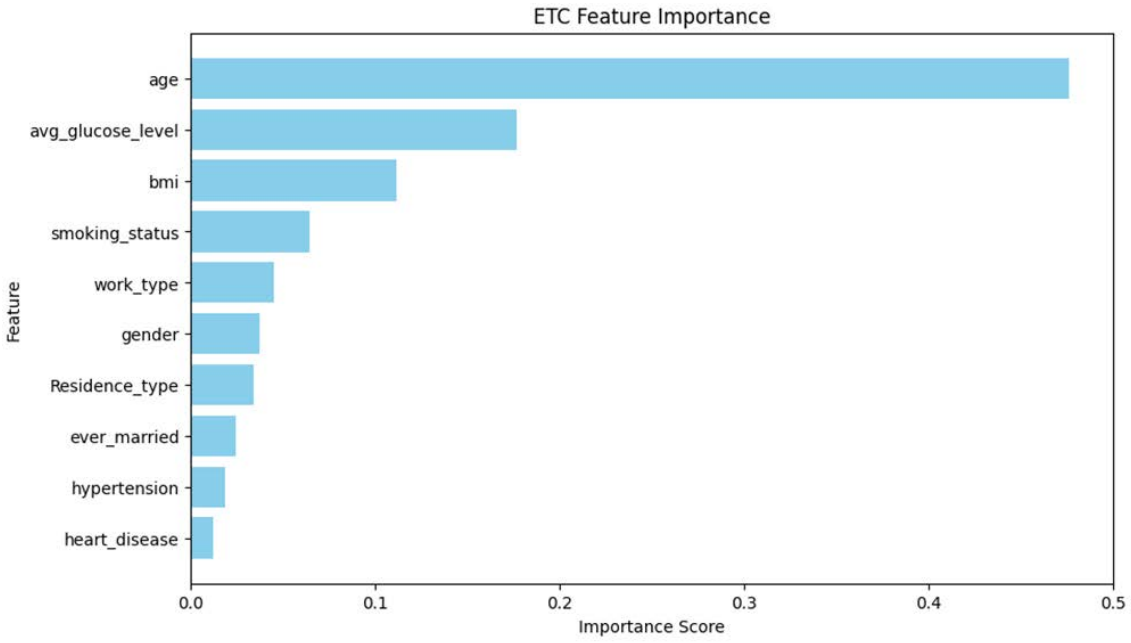

Fig. 3 ETC Feature Importance (Picture credit: Original)

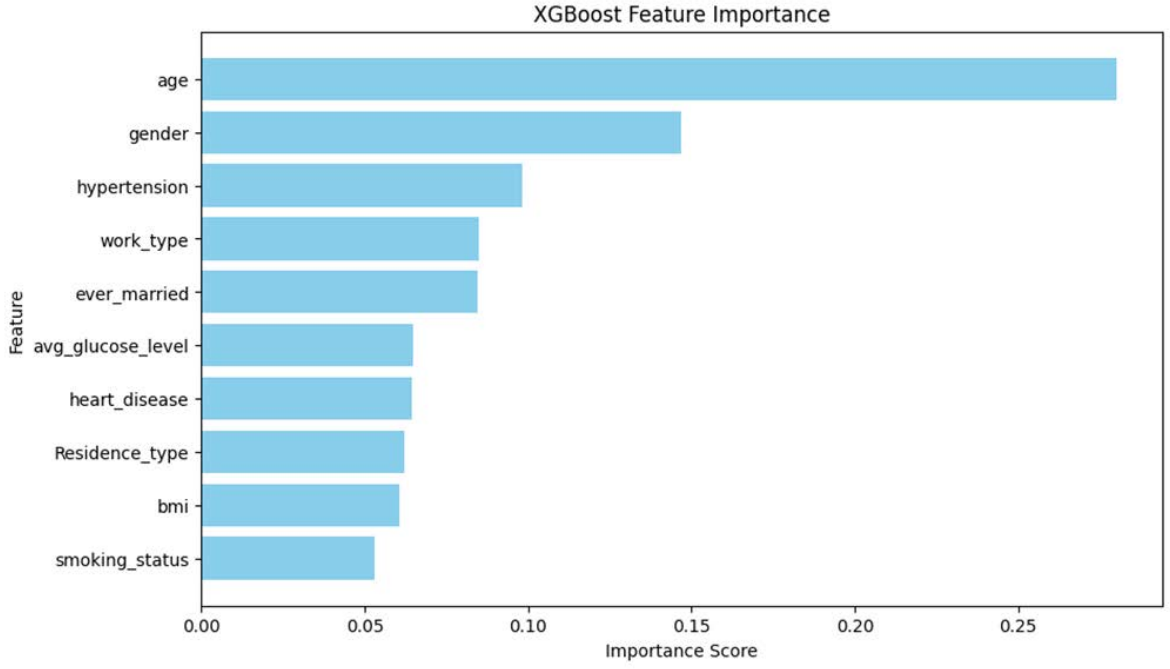

Fig. 4 XGBoost Feature Importance (Picture credit: Original)

3.3 Stacked Ensemble Model Performance

Based on the above three excellent models as the basic models, the author selected the soft voting model composed of the LR and XGBoost models as the meta-model. The author believes that these two models simultaneously offer both a linear perspective and a non-linear perspective. Moreover, XGBoost can identify the underlying patterns predicted by the base models, ultimately enhancing the performance of the final stacked ensemble model.

In order to compare the advantages of using the voting model as the meta-model, the author also constructed a stacked ensemble model that solely used LR as the meta-model. The performance of these models compared to other single models is as follows (Figure 5, 6).

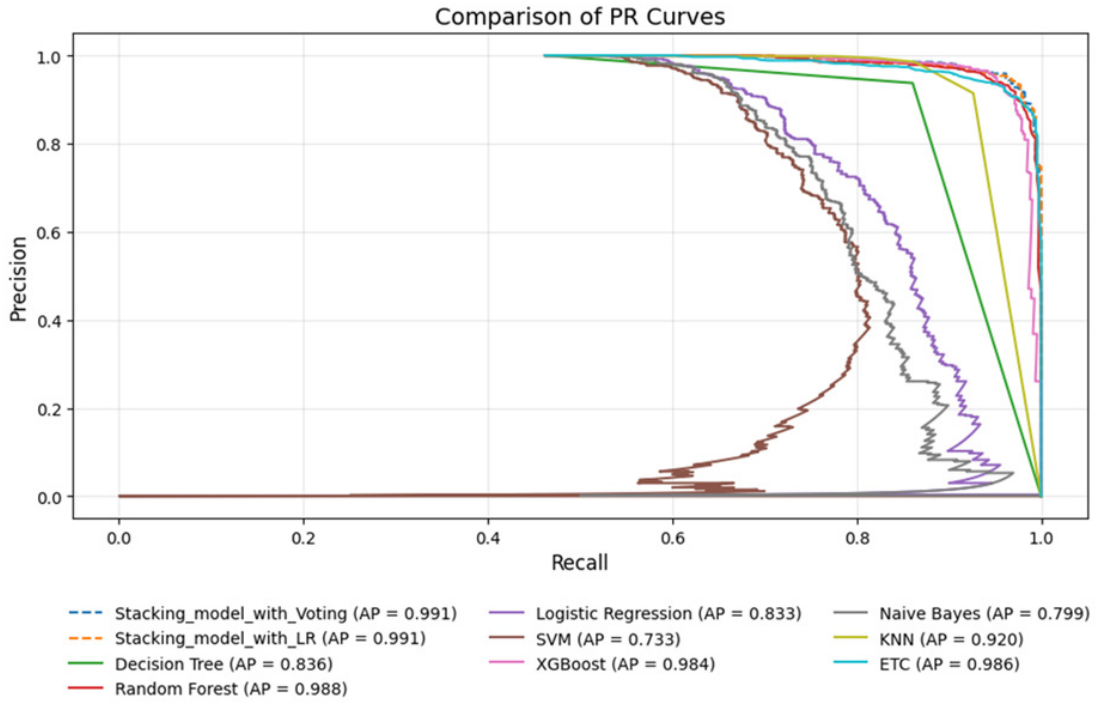

Fig. 5 Precision-Recall Curves (Picture credit: Original)

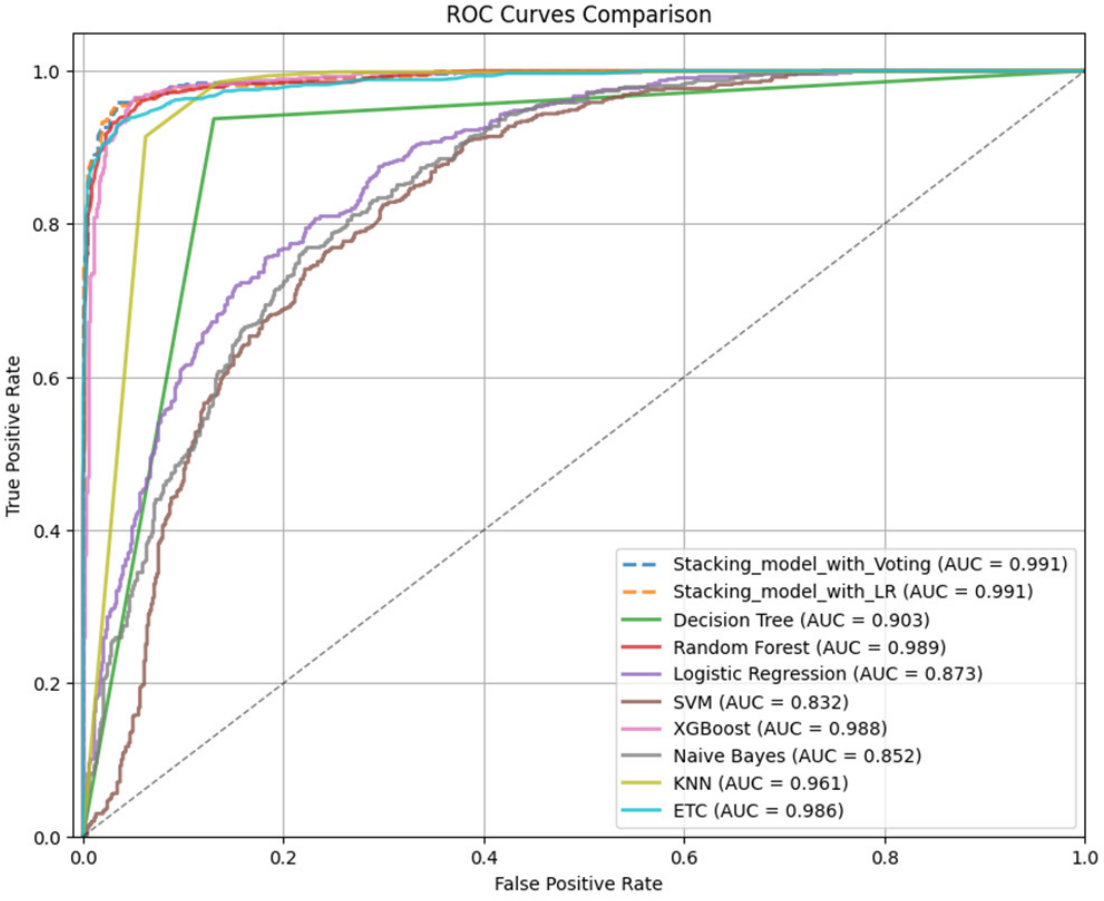

Fig. 6 ROC Curves (Picture credit: Original)

ISSN 2959-6157

Table 3. All Models Performance Comparison (With STOME-WENN)

Model	AUC	AP	Precision	Recall	F1	Accuracy
Stacking_model_with_Voting	0.991	0.991	0.949	0.959	0.954	0.957
Stacking_model_with_LR	0.991	0.991	0.924	0.964	0.943	0.947
DT	0.903	0.836	0.861	0.937	0.897	0.901
RF	0.989	0.988	0.914	0.969	0.941	0.944
LR	0.873	0.833	0.732	0.81	0.769	0.775
SVM	0.832	0.733	0.701	0.825	0.758	0.757
XGBoost	0.988	0.984	0.891	0.982	0.934	0.936
NB	0.852	0.799	0.716	0.815	0.762	0.765
KNN	0.961	0.92	0.822	0.993	0.9	0.898
ET	0.986	0.986	0.905	0.957	0.93	0.934

From Table 3 and the two figures (Figure 5, 6), the performance of the stacked ensemble model is significantly superior to that of the single model, and this is reflected in the three metrics of AUC, AP, and Accuracy. By comparing the two stacked ensemble models, it can be observed that their AUC and AP performance show almost no difference. However, the Accuracy of the model that uses Voting-Classifier is slightly higher than that of the stacked ensemble model that uses a single model as the meta-model.

By comparing the above results, it is easy to see that it is difficult to distinguish the performance differences between the two stacked ensemble models on the dataset processed by SMOTE-WENN. Therefore, the author separately compared the two models on the original dataset that was not processed (The dataset will undergo stratified sampling to maintain the original distribution), and the results are as follows (table 4).

Table 4. Stacking Models Performance Comparison (Without STOME-WENN)

Model	AUC	AP	Precision	Recall	F1	Accuracy
Stacking_model_with_Voting	0.797	0.379	0.406	0.289	0.338	0.93
Stacking_model_with_LR	0.798	0.35	0.193	0.644	0.297	0.811

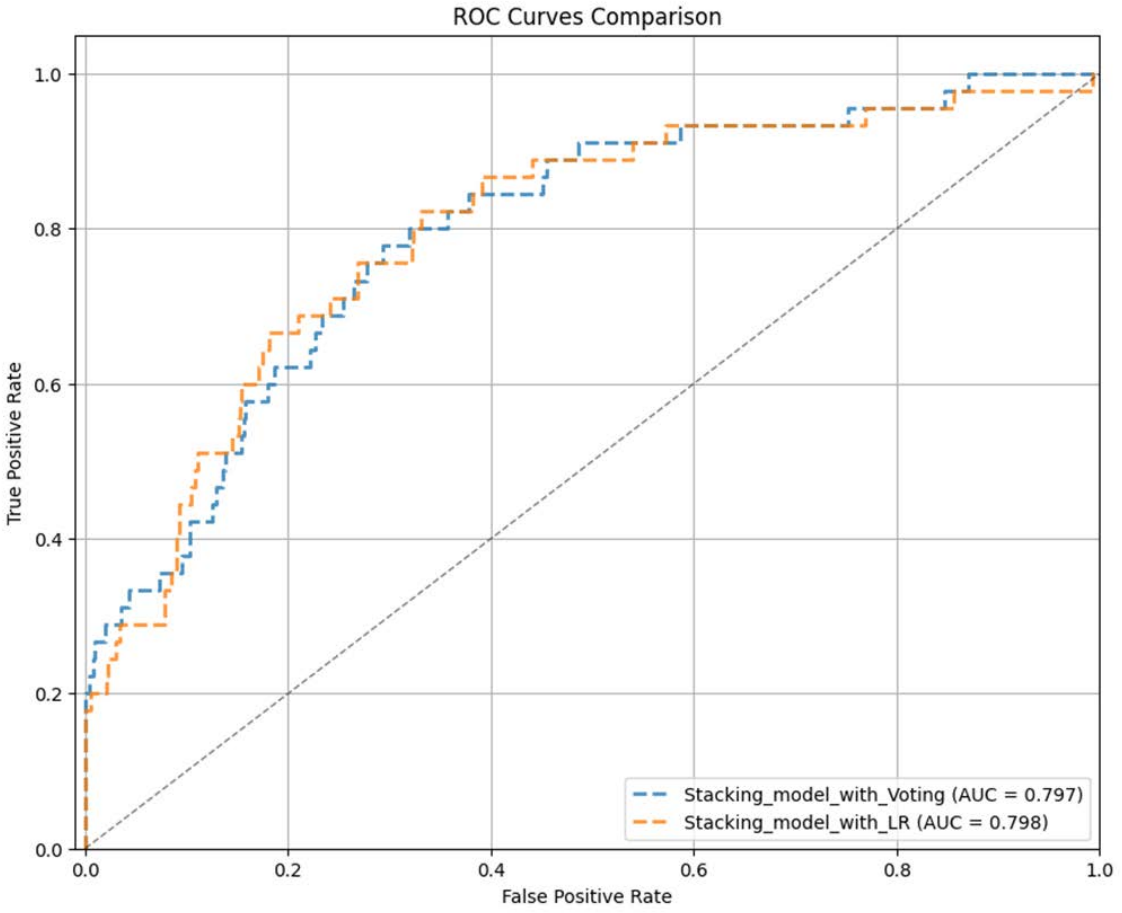

Fig. 7 Stacking Model ROC Curves (Picture credit: Original)

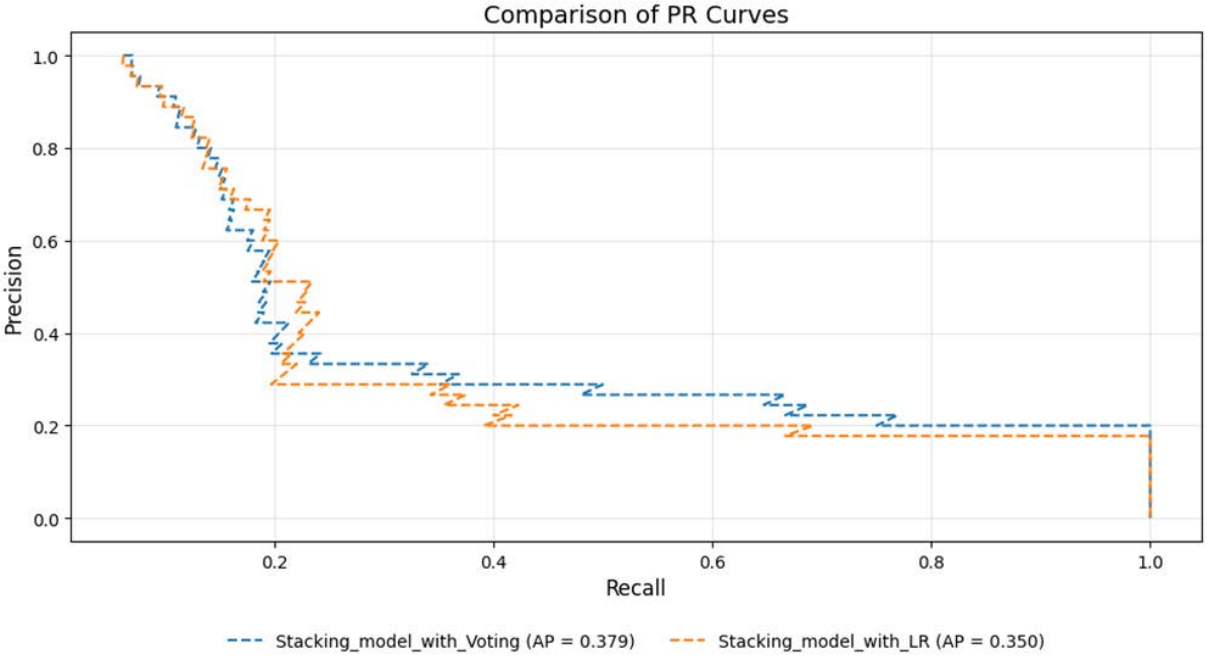

Fig. 8 Stacking Model PR Curves (Picture credit: Original)

ISSN 2959-6157

From the above figures (Figure 7, 8) and table 4, it is not difficult to observe that in extremely unbalanced datasets, although the AUC performance of the quantity model shows almost no difference, the stacked ensemble model using VotingClassifier performs better on AP and has a higher accuracy. From this, it can be seen that the robustness and generalization ability of the stacked ensemble model using Voting-Classifier is indeed better than that of the stacked ensemble model using a single model as the meta-model.

4. Conclusion

This study ultimately shows that using a stacked ensemble model is more effective than using a single model for predicting stroke, and the AUC reached 0.99 (when AUC is greater than 0.9, it is generally considered that the model performs well), and the AP also reached 0.99 simultaneously. The author believes that these metrics have demonstrated that the model has been able to perform the task of predicting stroke and has met the required accuracy standards in the medical field. The robustness and generalization ability of the stacked ensemble model using Voting-Classifier as the meta-model are also higher than those of the stacked model using a single model as the meta-model. In summary, the author successfully designed a stacked ensemble model with Voting-Classifier as the meta-model, and the performance of the model met the expectations.

References

[1] Zhang Y, Li F, Wang P, et al. WANG Ping‘s experience in treating apoplexy sequela from regulating and replenishing Yuan Qi. Chinese Journal of Traditional Chinese Medicine, 2021, 36(2): 866-868.

[2] He Y. Application and effect analysis of rehabilitation nursing model in the rehabilitation process of apoplexy sequelae patients. Smart Healthcare, 2025, 11(02): 103-105.

[3] Lan X, et al. Application of machine learning algorithm in medical field. Chinese Journal of Medical Instrumentation, 2019, 40(3): 93-97.

[4] Vappa K, et al. Machine learning method for knowledge discovery experimented with otoneurdigical data. Comput Methods Programs Biomed, 2008, 91(2) :155.

[5] Wang Y Q, et al. Research Progress of Machine Learning in the Diagnosis and Treatment of Esophageal Cancer. Computer Science, Advance online publication, 2025.

[6] Xu J. Artificial intelligence empowers technological innovation in medical: Transformation and challenges. Science and Technology Entrepreneurship Monthly, 2025, 38(6): 113-124.

[7] Murphy S J, Werring D J. Stroke: causes and clinical features. Medicine (Abingdon), 2020, 48(9): 561-566.

[8] Shi X, Liu Y. Machine learning for predicting stroke risk in heavy smokers-NHANES Study. China Digital Medicine, 2025, 20(3): 26-35.

[9] Begum A, et al. Effective stroke prediction using machine learning algorithms. Aust. J. Eng. Innov. Technol., 2024, 6(2), 26-36.

[10] Ding W W, Fang J T. Early warning of stroke risk based on optimized random forest and XGBoost model. Journal of Hubei University(Natural Science), 2025, 47(6).