frontiers | Frontiers in Physiology

TYPE Original Research
PUBLISHED 11 November 2024
DOI 10.3389/fphys.2024.1357123

OPEN ACCESS

EDITED BY
Raimond L. Winslow,
Northeastern University, United States

REVIEWED BY
Ming Huang,
Nagoya City University, Japan
Eric S. Ho,
Lafayette College, United States

*CORRESPONDENCE
Zhi-Ping Liu,
zpliu@sdu.edu.cn
Rui Gao,
gaorui@sdu.edu.cn

RECEIVED 20 December 2023
ACCEPTED 23 October 2024
PUBLISHED 11 November 2024

CITATION
Zhang D, Yu N, Yang X, De Marinis Y, Liu Z-P and Gao R (2024) SRPNet: stroke risk prediction based on two-level feature selection and deep fusion network.
Front. Physiol. 15:1357123.
doi: 10.3389/fphys.2024.1357123

SRPNet: stroke risk prediction based on two-level feature selection and deep fusion network

Daoliang Zhang[1], Na Yu[1], Xiaodan Yang[2], Yang De Marinis[1,3], Zhi-Ping Liu[1]* and Rui Gao[1]*

[1]School of Control Science and Engineering, Shandong University, Jinan, China, [2]Department of Rehabilitation Medicine, Affiliated Hospital of Jining Medical University, Jining, China, [3]Department of Clinical Sciences, Lund University, Malmö, Sweden

Background: Stroke is one of the major chronic non-communicable diseases (NCDs) with high morbidity, disability and mortality. The key to preventing stroke lies in controlling risk factors. However, screening risk factors and quantifying stroke risk levels remain challenging.

Methods: A novel prediction model for stroke risk based on two-level feature selection and deep fusion network (SRPNet) is proposed to solve the problem mentioned above. First, the two-level feature selection method is used to screen comprehensive features related to stroke risk, enabling accurate identification of significant risk factors while eliminating redundant information. Next, the deep fusion network integrating Transformer and fully connected neural network (FCN) is utilized to establish the risk prediction model SRPNet for stroke patients.

Results: We evaluate the performance of the SRPNet using screening data from the China Stroke Data Center (CSDC), and further validate its effectiveness with census data on stroke collected in affiliated hospital of Jining Medical University. The experimental results demonstrate that the SRPNet model selects features closely related to stroke and achieves superior risk prediction performance over benchmark methods.

Conclusions: SRPNet can rapidly identify high-quality stroke risk factors, improve the accuracy of stroke prediction, and provide a powerful tool for clinical diagnosis.

KEYWORDS

stroke risk prediction, feature selection, deep fusion network, transformer, stroke risk factors

1 Introduction

Stroke is a global public health issue, ranking as the second leading cause of death and the third leading cause of disability worldwide (Owolabi et al., 2022). Moreover, the incidence of stroke is increasing in recent years, and the burden of stroke poses a huge challenge to low- and middle-income countries (Owolabi et al., 2021). However, the complexity, suddenness, and significant differences in clinical manifestations of stroke have brought great difficulties to treatment. It is widely acknowledged that stroke is preventable and controllable (Johnson et al., 2019). Therefore, active intervention on risk factors of stroke and

accurate prediction of stroke risk through early screening can assist doctors and patients in implementing appropriate preventive and therapeutic measures, significantly reducing the harm caused by stroke.

So far, some studies employed traditional medical statistical methods to predict stroke risk (Wang et al., 2022; Abraham et al., 2021). These methods typically relied on a series of risk factors to construct mathematical models for calculating risk scores. However, these methods were time-consuming and labor-intensive, and ignored the complex nonlinear relationships and interactions among features, resulting in limited prediction performances (Obermeyer and Emanuel, 2016). With the rapid development of artificial intelligence, machine learning methods provide new solutions for stroke risk prediction. The machine learning methods can process complex screening data, and reveal patterns and associations hidden within large-scale data, thereby enhancing the accuracy of stroke risk prediction.

A better understanding of risk factors is critical for stroke diagnostic evaluation and treatment decision. In fact, controlling the risk factors (such as hypertension and diabetes) can reduce the risk of stroke. Qi et al. (2020) used multi-variable Cox regression analysis to obtain the features associated with the occurrence of stroke and its subtypes in China by introducing socioeconomic and other related factors. Abraham et al. (2019) employed elastic-net logistic regression to screen for genetic risk factors of stroke. Hunter and Kelleher (2023) used data from NHLBI Biologic Specimen and cardiac studies as risk factors, and studied the effect of age on stroke risk factors through a logistic regression algorithm. Maalouf et al. (2023) developed the regression model to find that negative emotions could increase stroke risk. Generally, stroke is a complex disease, and it is difficult to predict stroke risk via a single feature. However, having too many types of features may lead to redundant information and increase diagnostic costs. Furthermore, different risk factors contribute differently to stroke occurrence. More importantly, considering the association relationship among features is expected to be beneficial for the early stroke screening. Therefore, there is an urgent need to develop effective feature selection methods for predicting stroke risk.

Currently, numerous studies have been devoted to stroke risk prediction using machine learning techniques. For example, Li et al. (2019b) applied the Bayesian network model to estimate the incidence of stroke, revealing the relationship between combinations of multiple risk factors and stroke. Nwosu et al. (2019) analyzed the electronic health records of patients using neural networks, decision trees, and random forests to determine the impact of risk factors on stroke prediction. Arafa et al. (2022) developed a stroke risk prediction method for urban Japanese based on the Cox proportional hazards model, incorporating cardiovascular risk factors. Dritsas and Trigka (2022) designed an ensemble learning method for long-term stroke risk prediction. Liu et al. (2019) first adopted the random forest regression algorithm to impute missing data, and then used the deep neural network (DNN) to predict stroke on imbalanced physiological data. Although the above methods achieved promising results, the model structures they employed are relatively disconnected between features and algorithms, and the generalization ability of these models needs to be improved.

Here we propose a novel prediction model based on two-level feature selection and deep fusion network, termed SRPNet, for inferring stroke risk. In particular, two-level feature selection can comprehensively search for significant features related to stroke risk. We first apply multiple methods including Pearson correlation, chi-square test, Lasso and elastic net to select risk factors respectively, and combine the obtained risk factors as a candidate feature set. We then traverse all candidate risk factor combinations in the feature set by seven machine learning methods, such as support vector machine (SVM), k-nearest neighbor (KNN), decision tree (DT), gradient boosting decision tree (GBDT), random forest (RF), Gaussian Naive Bayes (GaussianNB) and AdaBoost, to identify the most important features associated with stroke. This enables evaluating the correlations between features and eliminating redundant information, providing reliable risk factors for stroke screening program. Next, the proposed deep fusion network integrates Transformer (Vaswani et al., 2017) and fully connected neural network (FCN) (Long et al., 2015) to establish a risk prediction model for stroke patients. This prediction model utilizes the attention mechanism of Transformer to explore hidden relationships among risk factors, and adopts FCN to better capture the nonlinear relationships among features. The experimental results indicate that SRPNet improves the accuracy and efficiency of stroke screening, and its performance is superior to existing benchmark methods. This work provides assistance for clinical diagnosis, and alleviates the burden of stroke.

2 Materials and methods

2.1 Datasets

The CSDC database covers 6 provinces, 41 hospitals and 12 population cohorts in China (Yu et al., 2016). The CSDC database facilitates stroke-related decision-making, research, and public health services through a comprehensive system. It collects and analyzes patient data, including risk factors, medical history, and sociodemographic information, ensuring that each subject has a unique record. A two-stage stratified cluster sampling method was employed during the data screening process (Li et al., 2019a). First, more than 200 screening areas were selected based on the local population size and the total number of counties. Then, urban communities and townships were used as the primary sampling units (PSUs) according to the geographical location and the recommendations from the local hospitals. In each PSU, all residents aged 40 and above were surveyed using cluster sampling during the initial screening period. Doctors assessed each patient's condition, categorizing them as low risk, medium risk, high risk, transient ischemic attack (TIA), or stroke. The CSDC dataset comprises 862,244 middle-aged residents. Table 1 shows the detailed features of the CSDC dataset.

The in-house data is sourced from the medical records of 49 patients at affiliated hospital of Jining Medical University in 2023. It includes 14 features such as gender, age group, ethnic groups, marital status, occupation, education level, hypertension, atrial fibrillation, smoking, hyperlipidemia, diabetes, overweight, and family history of stroke. Each patient has been diagnosed by a physician and classified as either having suffered a stroke or being in good health. The summary information for these two datasets is listed in Table 2.

TABLE 1 Summary of specific features in the CSDC dataset.

Risk factors	Statistics	Abbreviation	Risk factors	Statistics	Abbreviation
Age group	54.48 ± 11.25	AG	Diabetes	49,674/812,570	Diabetes
Gender (male/female)	397,765/464,479	Gender	Lack of exercise	169,500/692,744	LE
Ethnic groups (minorities/majority)	2,716/859,528	EG	Overweight	148,834/713,410	Overweight
Occupation (mental/manual)	147,585/650,577	Occupation	Number of Marriages[c]	0.94 ± 0.23	NM
Education status[a]	1.79 ± 0.93	ES	Marital status	783,723/78,521	MS
Family history of stroke/hypertension/coronary heart disease[b]	60,320/801,924	FHS/HYP/CHD	Marriage_other	3,537/858,707	MO
History of stroke	16,862/845,382	HS	Provincial GDP	39.12 ± 15.88	PGDP
Hypertension	182,800/679,444	HYP	Province longitude	112.94 ± 6.61	PLO
Atrial fibrillation	23,445/838,799	AF	Province latitude	35.18 ± 2.96	PLA
Low-Density Lipoprotein Cholesterol	270,313/591,931	LDL-C	Province precipitation	721.27 ± 202.66	PP
Province's highest temperature	26.83 ± 2.08	PHT	Province's highest humidity	78.60 ± 4.67	PHH
Province's lowest temperature	−0.09 ± 3.32	PLT	Province's lowest humidity	55.36 ± 12.80	PLH
Smoking	155,982/706,262	Smoking	Category[d]	384,272/477,972	Category

[a]The education status is divided into five levels, where 0 indicates illiteracy, 1 represents primary education, 2 represents secondary education, 3 represents higher education, and 4 represents postgraduate education.
[b]The counts of subjects with and without family history of stroke/hypertension/coronary heart disease.
[c]The number of marriages represents the number of times a subject has been married, with 0 for single, 1 for once married, and 2 for remarried.
[d]The variable category represents the categories of random grouping.

TABLE 2 The detailed information of datasets.

Datasets	# of samples	# of features	Phenotypes of samples
CSDC	862,244	26	Low risk (612,819), Medium risk (124,103), High risk (85,155), TIA (23,305), Stroke (16,862)
In-house data	49	14	Health (24), Stroke (25)

2.2 Overview of SRPNet

The SRPNet model mainly consists of two modules: two-level feature selection, and deep fusion network. The overall framework is illustrated in Figure 1. Since the dataset contains text information, the SRPNet firstly performs data preprocessing, which involves digitizing the textual information and normalizing the data. To eliminate low-correlation and redundant features, the two-level feature selection method is employed to identify comprehensive features associated with stroke. Finally, the deep fusion network, which adaptively fuse Transformer and FCN by attention mechanism, takes the obtained significant features as input to provide accurate stroke risk prediction results for stroke patients.

2.3 Data preprocessing

Based on the stroke risk researches (Tian et al., 2019; Guan et al., 2019), we used text information digitization to convert non-numeric features into numeric vectors suitable for machine learning or deep learning methods. Occupations are divided into mental workers and manual workers. For the marital status, we characterize it by whether the respondent is currently married and the number of marriage times. Based on the location information of the respondents, we convert it to the local climate, such as maximum temperature, minimum temperature, precipitation, humidity, etc. All of which are closely related to stroke. For the remaining features, we also use similar knowledge-based feature engineering for feature representation. Data normalization (Park et al., 2022) is used to

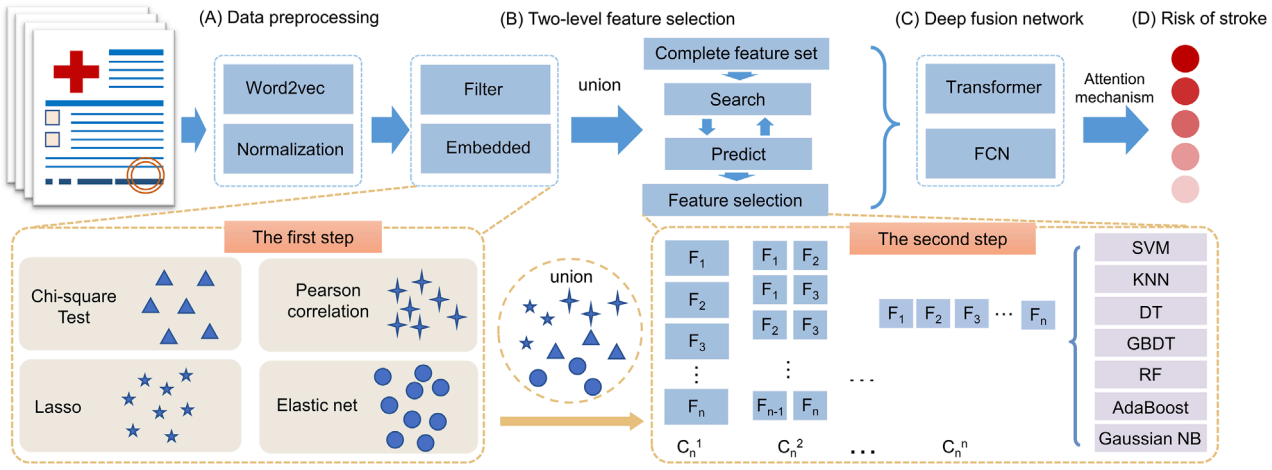

FIGURE 1
The entire framework of SRPNet. **(A)** Preprocess the input data. **(B)** Select significant features using the two-level feature selection. **(C)** Predict stroke risk based on the deep fusion network. **(D)** Output the stroke prediction results.

scale data elements to the (0,1) interval, which helps improve the effectiveness and reliability of model training. The normalization formula is defined as follows Equation 1:

$$x' = \frac{x - \min(x)}{\max(x) - \min(x)}. \quad (1)$$

2.4 Two-level feature selection

In this section, the two-level feature selection method that contains two-step feature selection processes will be introduced. The first step of feature selection involves four distinct methods, which are Pearson correlation, chi-square test, Lasso, and elastic net. The union of selected features from each method forms a set of candidate features. In the second step, based on the seven machine learning models, such as SVM, KNN, GBDT, RF, DT, AdaBoost and GaussianNB, we evaluate all possible combinations of candidate features via grid search. Each combination is scored based on its performance in the given models. It allows us to determine the optimal combination of features that are most predictive of stroke risk.

The two-step approach provides a rigorous feature selection process by multiple machine learning methods. The first step reduces the number of features based on statistical tests of relevance. The second step further refines the features by evaluating prediction performance in representative machine learning models. This ensures that the most informative and generalizable features have been selected for predicting stroke risk.

2.4.1 The first step of feature selection

We employ four feature selection methods, including chi-square test, Pearson correlation, Lasso and elastic net, to assess the correlation between features and disease risk from different perspectives. The chi-square test and Pearson correlation prefer to filter out features, which have the advantage of high computational efficiency while not being prone to overfitting. However, their over-reliance on filter thresholds may overlook many important features. On the other hand, Lasso and elastic net are embedded feature selection methods that select salient features while accounting for feature correlations by calculating feature weights. Therefore, we combined the filter and embedded methods to comprehensively screen for the important features related to stroke risk factors. For details, we provide brief introductions to the chi-square test, Pearson correlation, Lasso, elastic net.

Chi-square test (Sharpe, 2015). The chi-square test is used to check the correlation of the independent variable with the dependent variable. We use the chi-square test to delete the features with small changes. The formula of chi-square test is described as Equation 2:

$$\chi^2 = \sum \frac{(A - E)^2}{E}, \quad (2)$$

where A is the observed value of the feature, and E is the expected value of the feature. The assumption of chi-square test is that features are independent. The larger result of the chi-square test means the higher correlation between features.

Pearson correlation (Cohen et al., 2009). We use Pearson correlation coefficient to measure the linear correlation between features and disease risk. When all the features have been scaled to (0,1), the most important feature should have the highest coefficient, and the irrelevant feature should have a coefficient whose value is close to zero. The Pearson correlation coefficient can be determined by Equation 3:

$$\rho_{x_1,x_2} = \frac{\mathrm{cov}(x_1,x_2)}{\sigma x_1 \sigma x_2} = \frac{E(x_1x_2) - E(x_1)E(x_2)}{\sqrt{E(x_1^2) - E^2(x_1)}\sqrt{E(x_2^2) - E^2(x_2)}}, \quad (3)$$

where x_1 and x_2 represent the different feature, respectively. $\mathrm{cov}(x_1,x_2)$ denotes the covariance of x_1 and x_2. σx_1 denotes the standard deviation of x_1, and σx_2 denotes the standard deviation of x_2.

Lasso (Nusinovici et al., 2020). Lasso built upon logistic regression analysis techniques, serves to select the most crucial features while reducing model complexity through the shrinkage of feature weights. Specifically, lasso introduces L_1 regularization into the loss function of a linear regression model, minimizing the mean squared error between predicted values and actual observations. The Lasso loss function is given by Equation 4:

$$\min\theta\left\{\frac{1}{2n}\sum_{i=1}^{n}\left(y_i-\theta_0-\sum_{j=1}^{p}x_{ij}\theta_j\right)^2+\lambda\sum_{j=1}^{p}\left|\theta_j\right|\right\}, \tag{4}$$

where $\theta=\left(\theta_0,\theta_1,...,\theta_p\right)$ denote coefficients that we need to compute. y_i is the label that takes a value of 0 or 1. λ is a positive tuning parameter used to balance the loss term and penalty term. x_{ij} represents the value of the j-th feature of the i-th sample.

Elastic net (Zhang et al., 2017). Since Lasso regression sometimes performs poorly in inter-correlated features, the elastic net was proposed to overcome this limitation. Elastic net regularization combines L_1 penalty with L_2 penalty together to select better relevant features simultaneously. The elastic net is defined as Equation 5:

$$\min\theta\left\{\frac{1}{2n}\sum_{i=1}^{n}\left(y_i-\theta_0-\sum_{j=1}^{p}x_{ij}\theta_j\right)^2+\lambda\sum_{j=1}^{p}\left|\theta_j\right|+(1-\lambda)\sum_{j=1}^{p}\theta_j^2\right\}, \tag{5}$$

where $\lambda\in[0,1]$ used to balance the L_1 penalty and L_2 penalty. The L_2 penalty of regularization term is defined as $\varphi(\theta;\lambda)=(1-\lambda)\sum_{j=1}^{p}\theta_j^2$, which is known as Ridge regression.

2.4.2 The second step of feature selection

Although we have selected the important risk factors at the first step feature selection, the filter and embedded methods have the shortcomings of excessive threshold reliance and simply correlation consideration. To capture the deep correlation between features, we use seven machine learning methods to conduct the second step feature selection, which traverse all candidate features combinations based on the result of first step feature selection.

The candidate feature combinations consist of all possible permutations of the features selected during the process of feature selection. Assume there are n features that are selected. Then totally there are 2^n-1 candidate feature combinations. Next, we traverse all candidate combinations using seven machine learning methods and select the optimal feature combination based on the classification performance of these seven different classifiers. As we know, machine learning methods are based on specific theoretical assumptions. Therefore, employing different machine learning methods can increase the diversity of feature selection. Brief introductions of the seven machine learning methods are presented in Table 3. These algorithms have their own advantages and disadvantages, allowing us to thoroughly consider different scenarios in the feature selection process.

2.5 Deep fusion network

The complexity and diversity of stroke data require predicting stroke risk from multiple perspectives to enhance model robustness. Common predictive models, such as the Transformer, exhibit complex structures and excel at adapting to high-dimensional data, thus improving prediction performance. However, it often suffers from overfitting issues when dealing with small-scale datasets. In contrast, the FCN model has a simple structure and fast training speed, yielding exceptional performance on small-scale datasets. We utilize the advantages of both above predictive models and propose a deep fusion network method that can provide accurate the stroke risk prediction. As shown in Figure 2, deep fusion network integrates the Transform and the FCN, in which the dependencies between stroke risk factors are captured by the attention mechanism of the Transformer, and the complex nonlinear relationship is fitted by deep network structure of the FCN.

Transformer (Vaswani et al., 2017). Due to the powerful representation ability, Transformer can realize the outstanding performance in prediction tasks which is based on the self-attention mechanism. As observed in Figure 2, given an input $X\in\mathbb{R}^{n\times c}$, where n represents the number of patients (or patches) and c represents the embedded feature dimension for every patient. The self-attention mechanism can be defined as Equation 6:

$$Attention(Q,K,V)=\text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V, \tag{6}$$

where d_k is the input matrix embedding dimension. The matrix Q, K, V can be computed by the input matrix and linear transformation matrix W^Q, W^K, W^V, respectively. Then, we can get the value of Q, K, and V by computing $Q=XW^Q$, $K=XW^K$, and $V=XW^V$, where $W^Q\in\mathbb{R}^{c\times q}$, $W^K\in\mathbb{R}^{c\times q}$, $W^V\in\mathbb{R}^{c\times q}$, q denotes the linear mapping dimension.

Fully connected neural network. The FCN, also known as a Multilayer Perceptron (MLP), is a widely used artificial neural network structure in medical data analysis. It offers the advantages of fast training speed and robust modeling capabilities as the network depth increases. The stroke risk prediction model we designed includes one input layer, one hidden layer and one output layer. The calculation formula of each layer of network is defined by Equation 7:

$$y=\sigma(Wx+b), \tag{7}$$

where σ denotes the ReLU activation function. x is the input of the neuronal node, y is the output of the neuronal node. W and b denote weight and bias, respectively, which are learnable parameters.

Attention mechanism. In the stroke risk prediction task, the Transformer and the FCN extract clinical features at different levels and make distinct contributions to the prediction. Therefore, we introduce an attention mechanism to adaptively learn the importance of latent embeddings. Specifically, for the feature H_t extracted by Transformer, we apply a non-linear transformation and employ the shared attention vector W_t to obtain the attention coefficient a_t, namely, Equation 8:

$$a_t=\text{softmax}\left(W_t\cdot\sigma\left(WH_t+b\right)\right), \tag{8}$$

where σ denotes the tanh activation, W denotes a trainable weight matrix, and b denotes a bias vector. Similarly, we can calculate

TABLE 3 The overview of seven machine learning methods.

Methods	Theory	Advantages	Disadvantages
SVM (Noble, 2006)	Find the optimal hyperplane	Handling the interaction of nonlinear features	Difficulty in selecting the kernel function
KNN (Cunningham and Delany, 2021)	Find the k nearest neighbors	No assumptions, insensitive to outliers	Cannot handle imbalanced data
DT (Al Snousy et al., 2011)	Divides feature subspaces	Handle Boolean and numeric data simultaneously	Prone to overfitting
GBDT (Zhou et al., 2020)	Iteratively train decision trees	Strong interpretability	Difficulty tuning parameters
RF (Cutler et al., 2012)	Integrate several DTs	Strong generalization ability	Poor interpretability
AdaBoost (Schapire, 2013)	Integrated learning strategy	Prevent overfitting	Sensitive to outlier
GaussianNB (Liu et al., 2023)	Based on independence assumption	No need to tune parameters	Not suitable for high-dimensional data

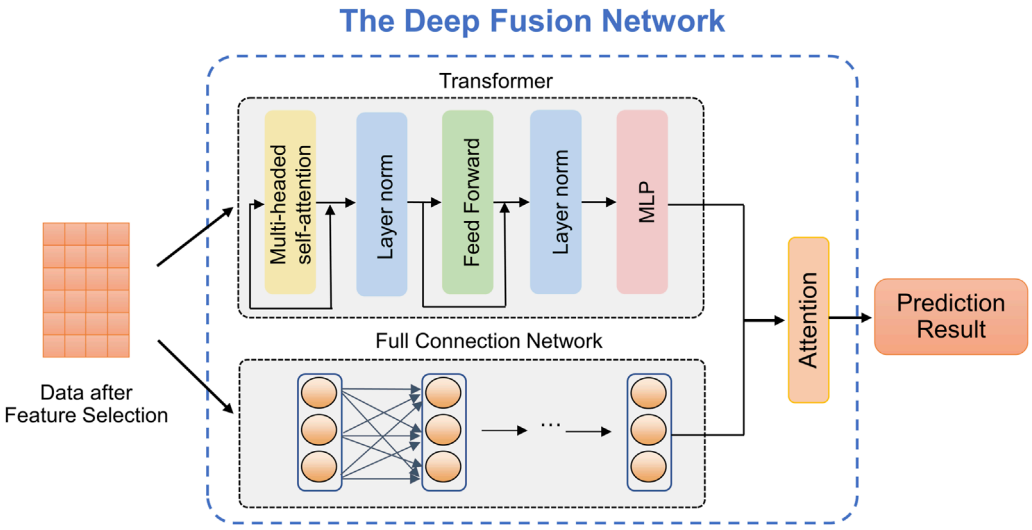

FIGURE 2
The structure of the proposed deep fusion network.

attention coefficients a_c for the features H_c extracted by the FCN. We combine these embeddings to obtain the final embedding H, Equation 9:

$$H = L(a_t \cdot H_t + a_c \cdot H_c), \tag{9}$$

where L denotes the single linear layer.

2.6 Evaluation metrics

Here we employ four evaluation metrics to assess the predictive performance of the model, including micro precision, micro F1-score, macro precision, and Cohen's Kappa coefficient (Younas et al., 2023). The definitions of these metrics are given as follows.

The micro average approach amalgamates performance measures across all samples. Specifically, for each class g_i within the set $G = \{1, ..., K\}$, where K denotes the total number of classes, a dedicated confusion matrix is constructed. In this context, the i-th matrix designates the g_i class as the positive class, while considering the remaining classes g_j with $j \neq i$ as the negative classes. The micro precision and micro F1-score are computed by Equations 10 and 11:

$$P_{micro} = \frac{\sum_{i=1}^{|G|} TP_i}{\sum_{i=1}^{|G|} TP_i + FP_i}, \tag{10}$$

$$F1_{micro} = \frac{2\sum_{i=1}^{|G|} TP_i}{2\sum_{i=1}^{|G|} TP_i + \sum_{i=1}^{|G|} FP_i + \sum_{i=1}^{|G|} FN_i}, \tag{11}$$

where TP represents the number of positive samples correctly predicted to be positive samples, FP represents the number of negative samples incorrectly predicted to be positive samples, FN represents the number of positive samples incorrectly predicted to be negative samples.

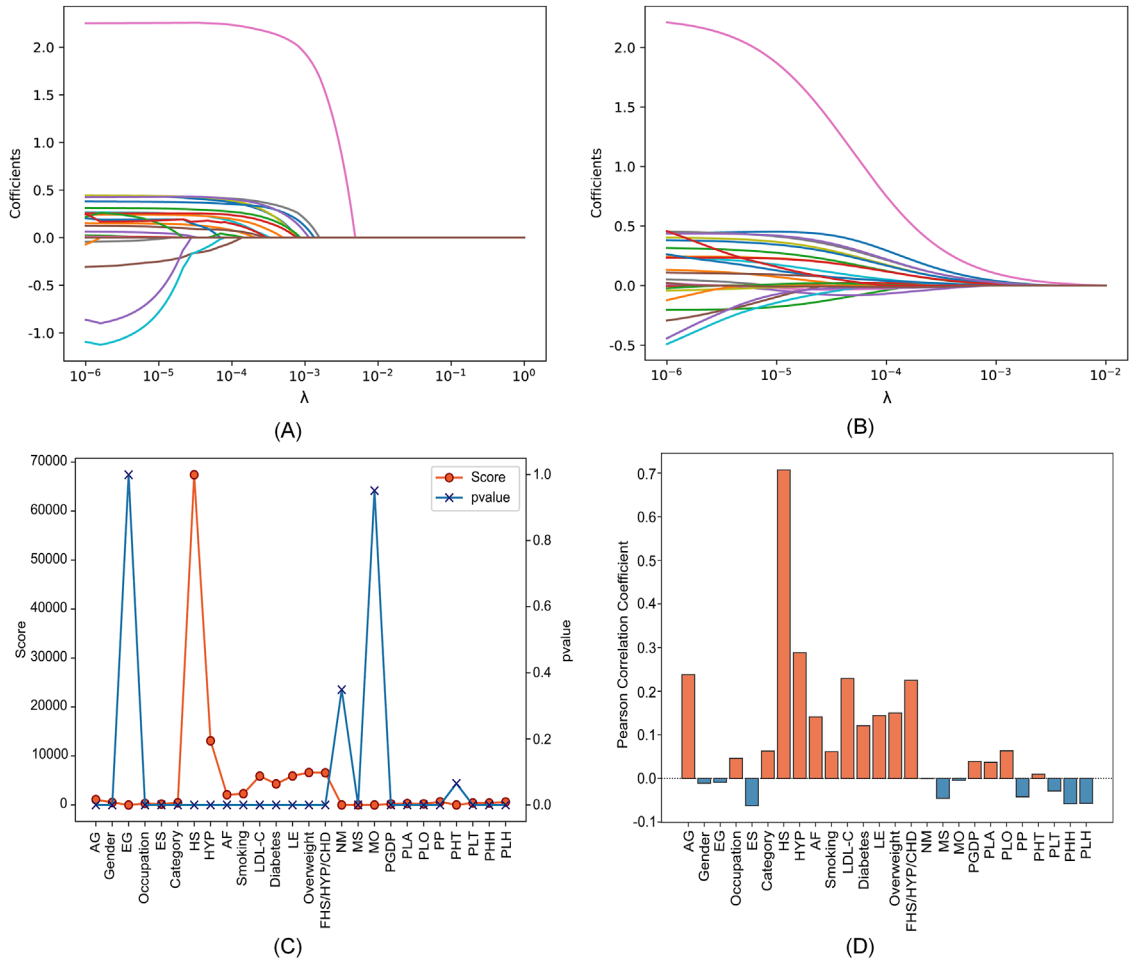

FIGURE 3
Feature selection results with different parameters in four feature selection methods. **(A)** Lasso. **(B)** Elastic net. **(C)** Chi-square test. **(D)** Pearson correlation.

TABLE 4 Feature selection results of four methods.

Methods	Features
Lasso	AG, Gender, Occupation, HS, HYP, AF, Smoking, LDL-C, Diabetes, LE, Overweight, FHS/HYP/CHD
Elastic net	AG, Gender, Occupation, ES, HS, HYP, AF, Smoking, LDL-C, Diabetes, LE, Overweight, FHS/HYP/CHD, PLT, PLH
Chi-square Test	HS, HYP, AF, LDL-C, Diabetes, LE, Overweight, FHS/HYP/CHD
Pearson correlation	AG, ES, HS, HYP, AF, Smoking, LDL-C, Diabetes, LE, Overweight, FHS/HYP/CHD, MS

Micro average tends to provide misleading results in the case of imbalanced data, as it doesn't take the predictive performance of each specific class into account. In contrast, macro average computes averages through the individual performance of each class. The macro precision is defined as Equation 12:

$$P_{macro} = \frac{1}{|G|}\sum_{i=1}^{|G|}\frac{TP_i}{TP_i + FP_i}. \quad (12)$$

Cohen's Kappa Coefficient is employed for assessing performance in situations of imbalanced class distribution, which is denoted by Equation 13:

$$kappa(k) = \frac{p_o - p_e}{1 - p_e}, \quad (13)$$

where p_o denotes the overall model accuracy, and p_e denotes the agreement expected by chance between the model's predictions and the actual class values (McHugh, 2012).

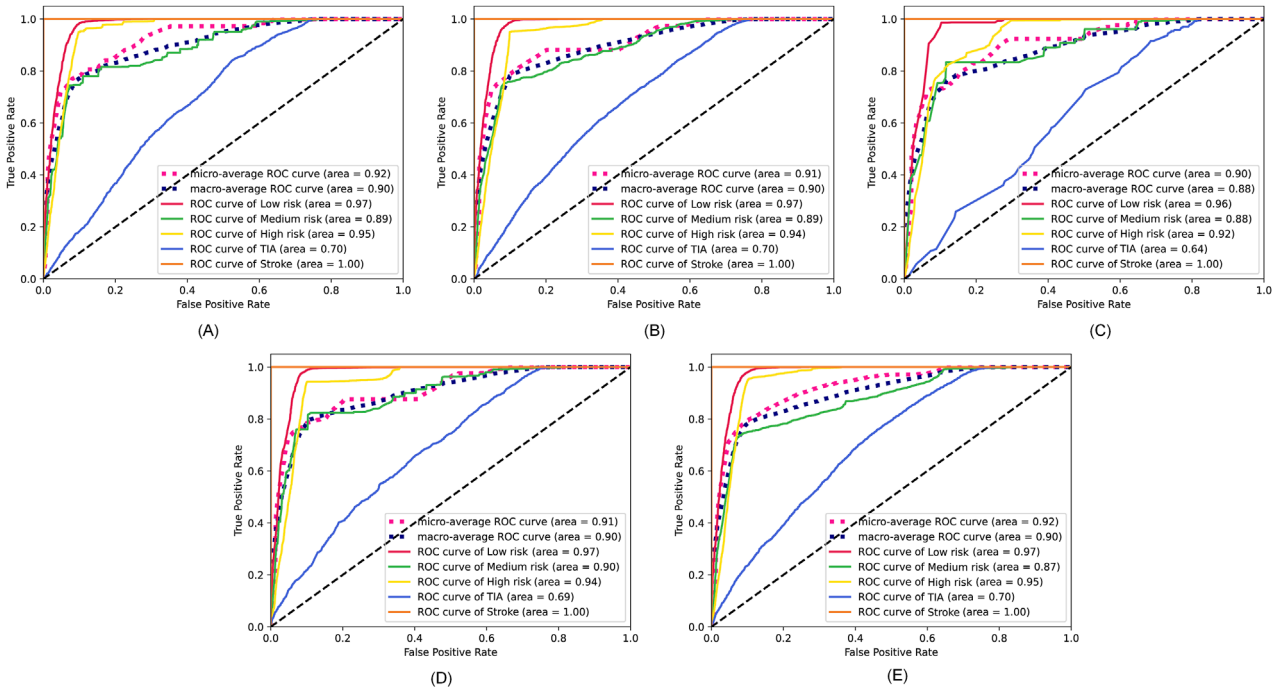

FIGURE 4
ROC curves corresponding to the feature sets selected by the four methods and the candidate feature set. **(A)** Lasso. **(B)** Elastic net. **(C)** Chi-square test. **(D)** Pearson correlation. **(E)** The robust candidate feature set.

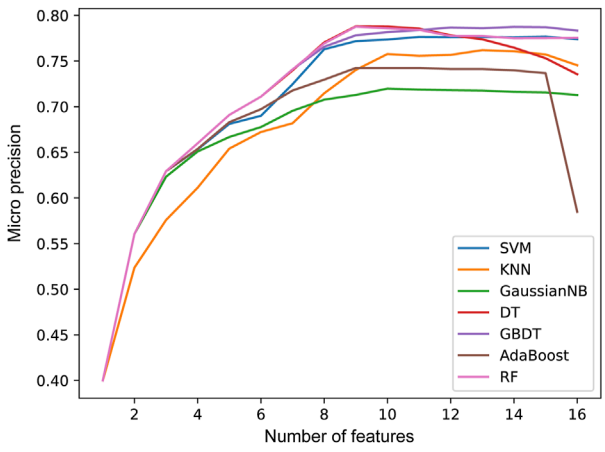

FIGURE 5
The prediction performance micro precision in seven machine learning methods with different number of features.

2.7 Implementation details

The stroke risk prediction model was built and trained using the PyTorch. Experiments were conducted on a PC with Intel(R) Xeon(R) Gold 6258R CPU @ 2.70 GHz and NVIDIA QuADro GV100 GPU. We trained the model with the Adam optimizer (Kingma and Ba, 2014) with default parameters and a fixed learning rate of 0.001. And we randomly select 80% of the samples from whole dataset for training, and the remaining 20% for testing. The maximum number of epochs employed for training is 100. The datasets and source codes are publicly available on GitHub: https://github.com/zhangdaoliang/SRPNet.

3 Results and discussion

3.1 Two-level feature selection results

We utilized a dataset from the CSDC database, consisting of 862,244 samples, with each sample originally having 26 distinct features. The proposed two-level feature selection method was used to screen out significant stroke features, which has a positive effect on improving the performance of the prediction model. In the first step of feature selection, we employed Lasso, elastic net, chi-square test and Pearson correlation methods for the initial screening of stroke-related factors. Here, we consider using $\alpha = 0.5$ for the elastic net. Figure 3 shows that the impact of different parameters contained in these methods on the feature selection results. We can observe in Figures 3A, B the paths of regression coefficient changes based on Lasso and elastic net, with each curve corresponding to one feature variable. Figures 3C, D demonstrate the correlation of each feature with stroke. It is worth noting that we tend to select features with higher scores and $p \leq 0.05$ in the chi-square test (Pandis, 2016). According to Figure 3, Lasso, elastic net, chi-square test and Pearson correlation methods select 13, 15, 8 and 12 features, respectively.

The specific feature selection results of each method are shown in Table 4. Subsequently, we took the union of features

TABLE 5 Comparison of stroke risk prediction results for the seven methods.

Methods	Selected features				All features			
	Micro F1-score	Micro precision	Macro precision	Cohen's Kappa coefficient	Micro F1-score	Micro precision	Macro precision	Cohen's Kappa coefficient
C5.0	0.9470	0.9470	0.7369	0.8828	0.9149	0.9149	0.7288	0.8068
RF	0.9478	0.9478	0.7672	0.8853	0.9167	0.9167	0.7170	0.8172
FCN	0.9257	0.9257	0.7119	0.8335	0.8906	0.8906	0.6811	0.7449
CNN	0.9421	0.9422	0.7316	0.8716	0.9371	0.9371	0.7310	0.8591
LSTM	0.9424	0.9424	0.8144	0.8723	0.9399	0.9399	0.7328	0.8654
Transformer	0.9480	0.9480	0.7449	0.8846	0.9198	0.9198	0.7396	0.8176
SRPNet	**0.9618**	**0.9618**	**0.8642**	**0.9165**	**0.9511**	**0.9511**	**0.8126**	**0.8920**

Note: The best experimental results are highlighted in bold.

TABLE 6 Prediction results based on our in-house dataset.

Methods	Micro F1-score	Micro precision	Macro precision	Cohen's Kappa coefficient
C5.0	0.9000	0.9000	0.8750	0.7826
RF	0.9000	0.9000	0.9285	0.7826
FCN	0.9953	0.9953	0.9933	0.9917
CNN	0.8000	0.8000	0.8571	0.6000
LSTM	0.8000	0.8000	0.8000	0.6000
Transformer	0.9000	0.9000	0.9283	0.7826
SRPNet	**0.9978**	**0.9978**	**0.9954**	**0.9929**

Note: The best experimental results are highlighted in bold.

selected by the four methods as the robust candidate feature set, which includes 16 features, i.e., AG, Gender, Smoking, MS, Occupation, ES, HS, HYP, AF, LDL-C, Diabetes, LE, Overweight, FHS/HYP/CHD, PLT, and PLH. The receiver operating characteristic (ROC) curves (Fan et al., 2006) corresponding to different feature sets are shown in Figure 4. We find that using the candidate feature set achieves better prediction results than features selected by individual methods. It illustrates that the first step of feature selection is of great significance for stroke risk diagnosis.

In the second step of feature selection, we eliminate risk factors with strong correlation between features. Based on the results of the first step of feature selection, we iterate through all candidate feature combinations. All feature combinations are evaluated under different machine learning methods as classifiers. The optimal feature combinations for different number of features are determined with respect to the evaluation results. Figure 5 shows the performance of the method with different numbers of feature variables. We see that as the number of features increases, the micro precision of most machine learning methods gradually improves and tends to stabilize. However, the performance of the DT and AdaBoost methods decreases significantly when the number of features is 9 and 15 respectively. When the number of features reaches 12, all seven machine learning methods overall achieve the best performance. Finally, we obtained risk factors that are highly relevant to stroke patients and have no redundant information among features, including Smoking, Occupation, ES, HS, HYP, AF, LDL-C, Diabetes, LE, Overweight, FHS/HYP/CHD, and PLT. It is worth noting that traditional methods consider age and gender to be strongly correlated with stroke risk (Howard et al., 2023; Ospel et al., 2023). However, two-level feature selection has removed them due to their redundancy with occupation and other risk factors. In contrast, the PLT features reflecting the climate of the patient's location are preserved, and it has been confirmed that low temperatures are associated with an increased risk of stroke (Chen et al., 2013). This indicates that SRPNet could provide new insights for future risk screening.

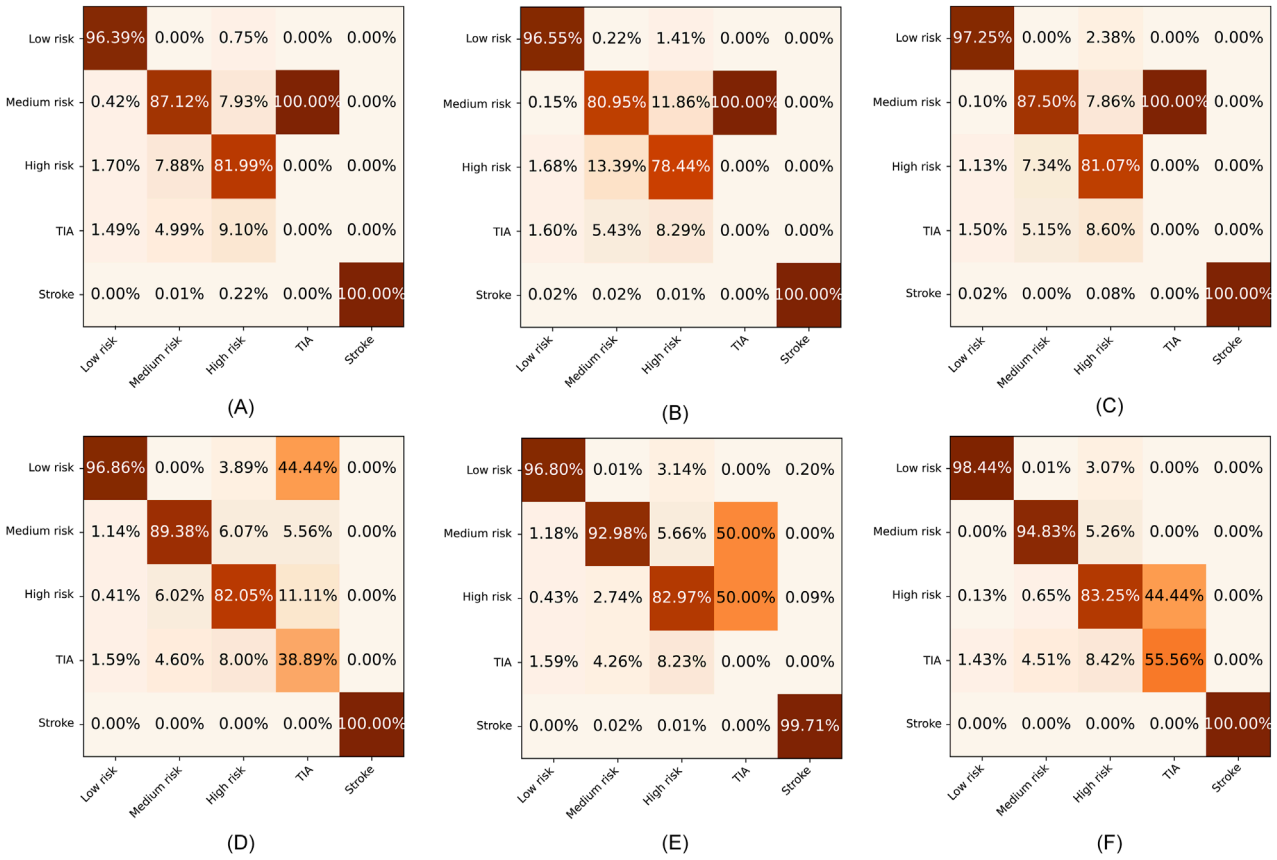

FIGURE 6
The confusion matrices for the six methods on the CSDC dataset. **(A)** C5.0. **(B)** FCN. **(C)** CNN. **(D)** LSTM. **(E)** Transformer. **(F)** SRPNet.

3.2 Stroke risk prediction results

In this section, we validate the effectiveness of the SRPNet model on the CSDC dataset. Decision tree C5.0 (C5.0) (Ahmadi et al., 2018), random forests (RF) (Breiman, 2001),FCN, one-dimensional convolutional neural network (CNN), long short-term memory network (LSTM) and Transformer are used as comparison methods to predict stroke risk. Table 5 shows the prediction performance of the seven methods on the original CSDC data (all features) and the data after two-level feature selection (selected features). We can find that SRPNet model obtains the best prediction results in terms of the four evaluation metrics. The performance of all predictors after two-level feature selection is significantly better than their performance when using all features. This demonstrates that the two-level feature selection can effectively filter weak and redundant information, thus improving the results of all predictors. On the selected feature data, SRPNet outperformes FCN and Transformer by approximately 1.4%, 1.4%, 12% and 3.2% on metrics micro F1-score, micro precision, macro precision, Cohen's Kappa coefficient. This reflects that deep fusion network can better explore potential relationships between risk factors. In summary, the proposed SRPNet model is reasonable and effective for predicting stroke risk.

Furthermore, to make the results more convincing, we evaluated six predictors on in-house data from affiliated hospital of Jining Medical University. The experimental results are recorded in Table 6. We can draw the similar conclusion that the proposed SRPNet model is an ideal and effective prediction tool of stroke risk. To explore the features that play a dominant role in precise classification, we removed each feature and obtained the prediction results for stroke risk. We found that after removing the hypertension (HYP) feature resulted in micro F1-score, micro precision, macro precision, and Cohen's Kappa coefficient of 0.7, 0.7, 0.83, and 0.28 respectively, which had the greatest impact on stroke prediction performance. Secondly, gender and age also significantly influenced stroke classification, while they are identified as redundant features in the CSDC dataset. The reason is that the analysis conducted on the CSDC dataset involves complex stroke risk prediction, focusing on differences between multiple risk levels, whereas the in-house dataset only focuses on whether someone has a stroke, conducting a simple stroke prediction analysis. Understanding these risk factors can assist doctors in making quick and accurate stroke diagnoses.

To further evaluate the superiority of SRPNet, we visualize the confusion matrices obtained by the six methods on the CSDC dataset and the in-house dataset in Figures 6, 7, where the columns and rows are the predicted labels and true labels, respectively. It shows that compared to other methods, The SRPNet method wins in all categories in terms of prediction accuracy. Additionally, we discover that the history of stroke (HS) feature and the hypertension (HYP) feature significantly enhance the ability of almost all algorithms in Figure 6 to detect stroke effectively.

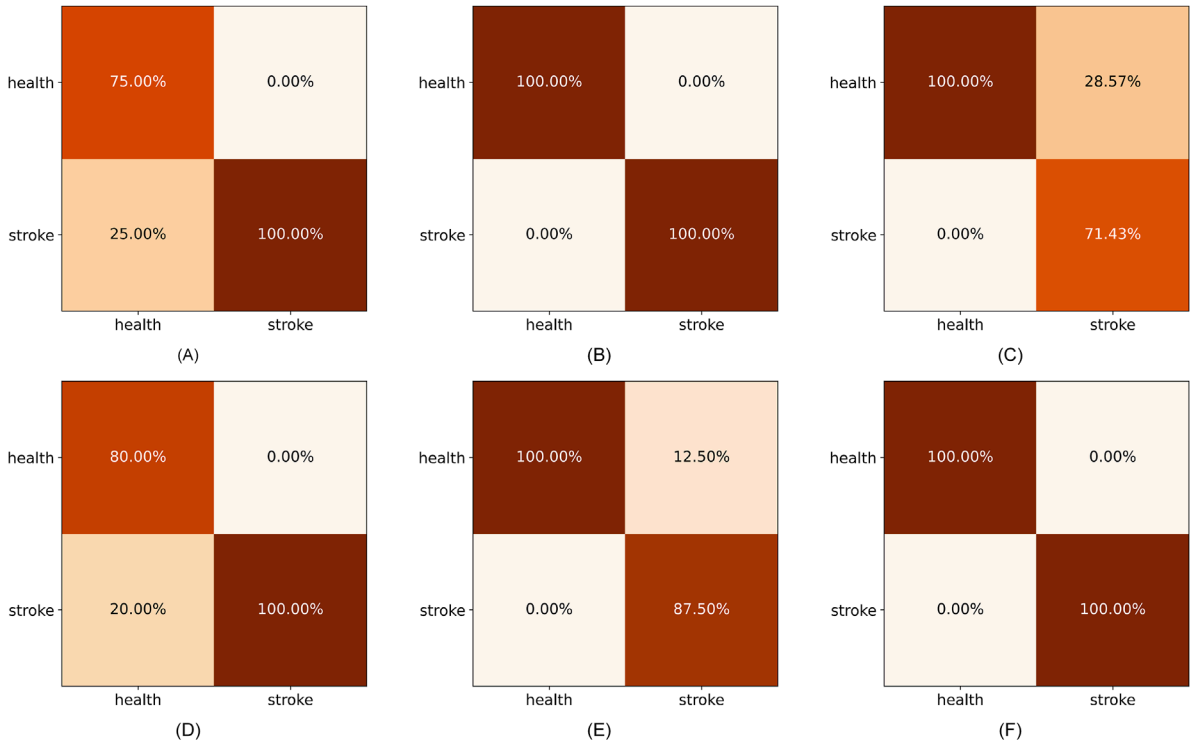

FIGURE 7
The confusion matrices for the six methods on the in-house dataset. **(A)** C5.0. **(B)** FCN. **(C)** CNN. **(D)** LSTM. **(E)** Transformer. **(F)** SRPNet.

4 Conclusion

In this paper, a novel prediction model based on two-level feature selection and deep fusion network is proposed for stroke risk prediction. Compared with traditional feature selection methods, the proposed two-level feature selection method not only focuses on the importance of individual these features, but also eliminates redundant information among important features. Furthermore, the proposed deep fusion network harnesses Transformer and fully connected networks to capture feature dependencies and model the non-linear relationships among features, respectively. Experimental results on the CSDC database and in-house dataset demonstrate that our proposed prediction model outperforms other representative methods. This prediction model can rapidly identify high-quality stroke risk factors and improve the accuracy of stroke prediction for patients, thereby effectively assisting doctors in formulating rational diagnosis and treatment plans.

The features included in the CSDC database and in-house dataset are limited. In the future, we will collect more clinical indicator features related to stroke for model training and testing. And we will also work on applying the proposed model to predict other diseases, demonstrating its generalizability. It's worth noting that researchers have the flexibility to substitute the feature selection method used in SRPNet with other methods that are frequently applied in the context of medical information, tailored to their specific requirements.

Data availability statement

The datasets presented in this study can be found in online repositories. The names of the repository/repositories and accession number(s) can be found below: https://github.com/zhangdaoliang/SRPNet.

Author contributions

DZ: Conceptualization, Data curation, Methodology, Software, Validation, Writing–original draft, Writing–review and editing. NY: Conceptualization, Investigation, Validation, Writing–review and editing. XY: Conceptualization, Data curation, Validation, Writing–review and editing. YD: Funding acquisition, Validation, Writing–review and editing. Z-PL: Conceptualization, Funding acquisition, Supervision, Validation, Writing–review and editing. RG: Conceptualization, Funding acquisition, Project administration, Supervision, Validation, Writing–review and editing.

Funding

The author(s) declare that financial support was received for the research, authorship, and/or publication of this article. This research was supported by the National Natural Science Foundation of China (NSFC) (Grant Nos U1806202, 62373216), the Fundamental Research Funds for the Central Universities (2022JC008), and the Program of Qilu Young Scholars of Shandong University.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

The author(s) declared that they were an editorial board member of Frontiers, at the time of submission. This had no impact on the peer review process and the final decision.

Publisher's note

References

Abraham, G., Malik, R., Yonova-Doing, E., Salim, A., Wang, T., Danesh, J., et al. (2019). Genomic risk score offers predictive performance comparable to clinical risk factors for ischaemic stroke. *Nat. Commun.* 10, 5819. doi:10.1038/s41467-019-13848-1

Abraham, G., Rutten-Jacobs, L., and Inouye, M. (2021). Risk prediction using polygenic risk scores for prevention of stroke and other cardiovascular diseases. *Stroke* 52, 2983–2991. doi:10.1161/STROKEAHA.120.032619

Ahmadi, E., Weckman, G. R., and Masel, D. T. (2018). Decision making model to predict presence of coronary artery disease using neural network and C5. 0 decision tree. *J. Ambient Intell. Humaniz. Comput.* 9, 999–1011. doi:10.1007/s12652-017-0499-z

Al Snousy, M. B., El-Deeb, H. M., Badran, K., and Al Khlil, I. A. (2011). Suite of decision tree-based classification algorithms on cancer gene expression data. *Egypt. Inf. J.* 12, 73–82. doi:10.1016/j.eij.2011.04.003

Arafa, A., Kokubo, Y., Sheerah, H. A., Sakai, Y., Watanabe, E., Li, J., et al. (2022). Developing a stroke risk prediction model using cardiovascular risk factors: the Suita Study. *Cerebrovasc. Dis.* 51, 323–330. doi:10.1159/000520100

Breiman, L. J. M. L. (2001). Random forests. *Mach. Learn.* 45, 5–32. doi:10.1023/a:1010933404324

Chen, R., Wang, C., Meng, X., Chen, H., Thach, T. Q., Wong, C.-M., et al. (2013). Both low and high temperature may increase the risk of stroke mortality. *Neurology* 81, 1064–1070. doi:10.1212/WNL.0b013e3182a4a43c

Cohen, I., Huang, Y., Chen, J., Benesty, J., Benesty, J., Chen, J., et al. (2009). Pearson correlation coefficient. *Noise Reduct. speech Process.*, 1–4. doi:10.1007/978-3-642-00296-0_5

Cunningham, P., and Delany, S. J. (2021). K-nearest neighbour classifiers-a tutorial. *ACM Comput. Surv. (CSUR)* 54, 1–25. doi:10.1145/3459665

Cutler, A., Cutler, D. R., and Stevens, J. R. (2012). *Random forests*. Editors C. Zhang, and Y. Q. Ma (Springer, New York: Ensemble Machine Learning), 157–175. doi:10.1007/978-1-4419-9326-7_5

Dritsas, E., and Trigka, M. (2022). Stroke risk prediction with machine learning techniques. *Sensors* 22, 4670. doi:10.3390/s22134670

Fan, J., Upadhye, S., and Worster, A. (2006). Understanding receiver operating characteristic (ROC) curves. *Can. J. Emerg. Med.* 8, 19–20. doi:10.1017/s1481803500013336

Guan, W., Clay, S. J., Sloan, G. J., and Pretlow, L. G. (2019). Effects of barometric pressure and temperature on acute ischemic stroke hospitalization in Augusta, GA. *Transl. Stroke Res.* 10, 259–264. doi:10.1007/s12975-018-0640-0

Howard, G., Banach, M., Kissela, B., Cushman, M., Muntner, P., Judd, S. E., et al. (2023). Age-related differences in the role of risk factors for ischemic stroke. *Neurology* 100, e1444-e1453. doi:10.1212/WNL.0000000000206837

Hunter, E., and Kelleher, J. D. (2023). Determining the proportionality of ischemic stroke risk factors to age. *J. Cardiovasc. Dev. Dis.* 10, 42. doi:10.3390/jcdd10020042

Johnson, C. O., Nguyen, M., Roth, G. A., Nichols, E., Alam, T., Abate, D., et al. (2019). Global, regional, and national burden of stroke, 1990–2016: a systematic analysis for the Global Burden of Disease Study 2016. *Lancet Neurology* 18, 439–458. doi:10.1016/S1474-4422(19)30034-1

Kingma, D. P., and Ba, J. (2014). Adam: a method for stochastic optimization. *arXiv Prepr. arXiv:1412.6980*.

Li, X., Bian, D., Yu, J., Li, M., and Zhao, D. (2019a). Using machine learning models to improve stroke risk level classification methods of China national stroke screening. *BMC Med. Inf. Decis. Mak.* 19, 261–267. doi:10.1186/s12911-019-0998-2

Li, X., Pang, J., Li, M., and Zhao, D. (2019b). Discover high-risk factor combinations using Bayesian network from cohort data of National Stoke Screening in China. *BMC Med. Inf. Decis. Mak.* 19, 67–68. doi:10.1186/s12911-019-0753-8

Liu, D., Lin, Z., and Jia, C. (2023). NeuroCNN_GNB: an ensemble model to predict neuropeptides based on a convolution neural network and Gaussian naive Bayes. *Front. Genet.* 14, 1226905. doi:10.3389/fgene.2023.1226905

Liu, T., Fan, W., and Wu, C. (2019). A hybrid machine learning approach to cerebral stroke prediction based on imbalanced medical dataset. *Artif. Intell. Med.* 101, 101723. doi:10.1016/j.artmed.2019.101723

Long, J., Shelhamer, E., and Darrell, T. (2015). "Fully convolutional networks for semantic segmentation," in Proceedings of the IEEE conference on computer vision and pattern recognition, 3431–3440.

Maalouf, E., Hallit, S., Salameh, P., and Hosseini, H. (2023). Depression, anxiety, insomnia, stress, and the way of coping emotions as risk factors for ischemic stroke and their influence on stroke severity: a case–control study in Lebanon. *Front. psychiatry* 14, 1097873. doi:10.3389/fpsyt.2023.1097873

Mchugh, M. L. (2012). Interrater reliability: the kappa statistic. *Biochem. medica* 22, 276–282. doi:10.11613/bm.2012.031

Noble, W. S. (2006). What is a support vector machine? *Nat. Biotechnol.* 24, 1565–1567. doi:10.1038/nbt1206-1565

Nusinovici, S., Tham, Y. C., Yan, M. Y. C., Ting, D. S. W., Li, J., Sabanayagam, C., et al. (2020). Logistic regression was as good as machine learning for predicting major chronic diseases. *J. Clin. Epidemiol.* 122, 56–69. doi:10.1016/j.jclinepi.2020.03.002

Nwosu, C. S., Dev, S., Bhardwaj, P., Veeravalli, B., and John, D. (2019). "Predicting stroke from electronic health records," in 2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC) (IEEE), 5704–5707.

Obermeyer, Z., and Emanuel, E. J. (2016). Predicting the future—big data, machine learning, and clinical medicine. *N. Engl. J. Med.* 375, 1216–1219. doi:10.1056/NEJMp1606181

Ospel, J., Singh, N., Ganesh, A., and Goyal, M. J. J. O. S. (2023). Sex and gender differences in stroke and their practical implications in acute care. *J. Stroke* 25, 16–25. doi:10.5853/jos.2022.04077

Owolabi, M. O., Thrift, A. G., Mahal, A., Ishida, M., Martins, S., Johnson, W. D., et al. (2022). Primary stroke prevention worldwide: translating evidence into action. *Lancet Public Health* 7, e74–e85. doi:10.1016/S2468-2667(21)00230-9

Owolabi, M. O., Thrift, A. G., Martins, S., Johnson, W., Pandian, J., Abd-Allah, F., et al. (2021). The state of stroke services across the globe: report of world stroke organization–world health organization surveys. *Int. J. Stroke* 16, 889–901. doi:10.1177/17474930211019568

Pandis, N. (2016). The chi-square test. *Am. J. Of Orthod. And Dentofac. Orthop.* 150, 898–899. doi:10.1016/j.ajodo.2016.08.009

Park, H. W., Pitti, T., Madhavan, T., Jeon, Y.-J., and Manavalan, B. (2022). MLACP 2.0: an updated machine learning tool for anticancer peptide prediction. *Comput. Struct. Biotechnol. J.* 20, 4473–4480. doi:10.1016/j.csbj.2022.07.043

Qi, W., Ma, J., Guan, T., Zhao, D., Abu-Hanna, A., Schut, M., et al. (2020). Risk factors for incident stroke and its subtypes in China: a prospective study. *J. Am. Heart Assoc.* 9, e016352. doi:10.1161/JAHA.120.016352

Schapire, R. E. (2013). "Explaining adaboost," in *Empirical inference: festschrift in honor of vladimir N. Vapnik* (Springer Berlin Heidelberg), 37–52.

Sharpe, D. (2015). Chi-square test is statistically significant: now what? *Pract. Assess. Res. Eval.* 20, 8. doi:10.7275/tbfa-x148

Tian, Y., Liu, H., Si, Y., Cao, Y., Song, J., Li, M., et al. (2019). Association between temperature variability and daily hospital admissions for cause-specific cardiovascular disease in urban China: a national time-series study. *PLoS Med.* 16, e1002738. doi:10.1371/journal.pmed.1002738

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., et al. (2017). Attention is all you need. *Adv. neural Inf. Process. Syst.* 30.

Wang, Q., Zhang, L., Li, Y., Tang, X., Yao, Y., and Fang, Q. (2022). Development of stroke predictive model in community-dwelling population: a longitudinal cohort

study in Southeast China. *Front. Aging Neurosci.* 14, 1036215. doi:10.3389/fnagi.2022.1036215

Younas, F., Usman, M., and Yan, W. Q. (2023). A deep ensemble learning method for colorectal polyp classification with optimized network parameters. *Appl. Intell.* 53, 2410–2433. doi:10.1007/s10489-022-03689-9

Yu, J., Mao, H., Li, M., Ye, D., and Zhao, D. (2016). "CSDC—a nationwide screening platform for stroke control and prevention in China," in 2016 38th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC) (IEEE), 2974–2977.

Zhang, Z., Lai, Z., Xu, Y., Shao, L., Wu, J., and Xie, G.-S. (2017). Discriminative elastic-net regularized linear regression. *IEEE Trans. Image Process.* 26, 1466–1481. doi:10.1109/TIP.2017.2651396

Zhou, S., Wang, S., Wu, Q., Azim, R., and Li, W. (2020). Predicting potential miRNA-disease associations by combining gradient boosting decision tree with logistic regression. *Comput. Biol. Chem.* 85, 107200. doi:10.1016/j.compbiolchem.2020.107200