Big Data and Information Analytics

BDIA, 9: 400–426.
DOI: 10.3934/bdia.2025020
Received: 19 November 2025
Revised: 07 December 2025
Accepted: 10 December 2025
Published: 18 December 2025

https://www.aimspress.com/journal/BDIA

Research article

Association of blood heavy metals with risk of stroke: A machine learning-based study from NHANES 2017–2020

KeXin Li[1], FengQi Liu[2] and Enxiao Zhu[1,*]

[1] Henan Academy of Big Data, School of Mathematics and Statistics, Zhengzhou University, Zhengzhou 450001, China
[2] Department of Dermatology, The Second Affiliated Hospital of Xi'an Jiaotong University, Xi'an 710004, China

* **Correspondence:** Email: 1850916858@qq.com.

Abstract: Stroke is a condition in which the brain and spinal cord are damaged by abnormal blood supply and is one of the leading causes of disability and death worldwide. In order to meet the urgent public health needs arising from the rapidly increasing incidence of stroke, it is important to predict and diagnose stroke in advance. Environmental heavy metals have been implicated in the risk of stroke, but their role in early prediction is still understudied. The purpose of this study is to incorporate the heavy metal content in environmental exposure into the stroke risk prediction model and to analyze the stroke association of these heavy metal characteristics. The main research contents of this paper are as follows: First, the data were extracted from the NHANES database, and feature vectorization, K-nearest neighbor imputation, and normalization were performed. Due to the imbalance of data classes, five methods were used to deal with the problem, and the results show that the cost-sensitive method has the best effect, with an accuracy of 0.98. Through feature selection and correlation analysis, 14-dimensional important features were selected, including blood lead and blood manganese, two heavy metals. Second, this paper evaluates and compares a variety of traditional and ensemble machine learning models, such as random forest (RF), gradient-boosted decision trees (GDBT) and XGBoost. The results showed that the random forest model performed the best, with an accuracy of 0.96, a percision of 0.93, a recall rate of 0.98, and an F1 score of 0.95. Combined with Shapley additive explanations (SHAP) theory, the prediction results were explained and the influence of each feature on the prediction results was analyzed. The results showed that blood lead level was significantly associated with stroke (95% confidence interval: 0.227–0.522; p-value < 0.001), while blood manganese was significantly negatively correlated with stroke risk (95% confidence interval: 0.131–

0.022; p-value < 0.006). In addition, the above results were verified by plotting the dose-response curve. These findings suggest that environmental heavy metal exposure has important value in stroke prediction.

Keywords: machine learning; heavy metals; stroke risk prediction

1. Introduction

Stroke is a condition of brain and spinal cord injury caused by abnormal blood supply and is divided into ischemic stroke (IS) [1], hemorrhagic stroke (HE) [2], and transient ischemic attack (TIA) [3,4]. Figure 1 shows the stroke mortality rate in the United States from 2014 to 2020. It is clearly observable that since 2018, the growth rate of the mortality rate has been very fast, reaching a peak of about 40% in 2020. Globally, especially in developing countries, the number of deaths due to stroke is growing rapidly [5]. According to the World Stroke Organization (WSO), stroke is the second leading cause of disability worldwide, and stroke cases have doubled every decade over the past 40 years [6]. However, due to the lack of specific treatments, early diagnosis is crucial. If stroke is detected or diagnosed early, death and serious brain damage can be largely prevented.

In recent years, as environmental issues have become a hot topic of global concern, heavy metal pollution and its impact on human health have become an important branch of public health research [7–9]. Heavy metals can enter the human body through food, water, air and accumulate in the body, potentially affecting health [10]. Although there is a large body of literature supporting the relationship between environmental exposure to heavy metals and a variety of adverse health outcomes, the effective integration and application of heavy metal signatures in the field of early prediction of stroke is still rarely explored.

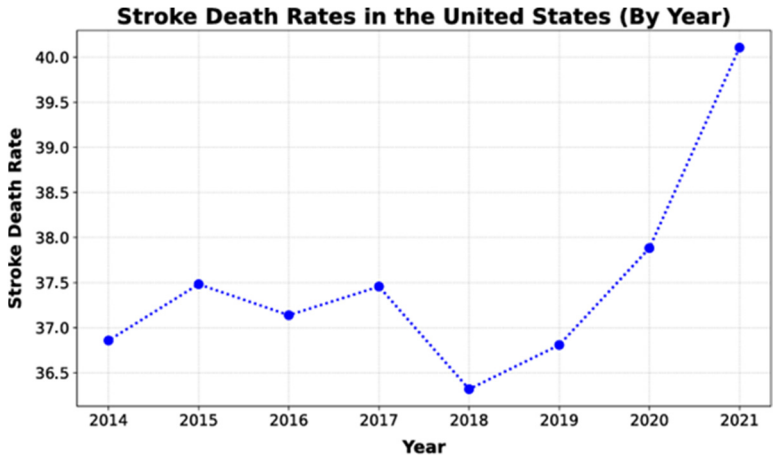

Figure 1. Stroke mortality rate by state in the United States.

Previous studies have used machine learning(ML) techniques to predict stroke: Amini et al. (2013) used decision trees (DT) and K-nearest neighbor imputation (KNN) to predict stroke, and the DT algorithm had the best accuracy (95%) [11]. Sung et al. (2015) developed the stroke severity index (SSI) using data mining methods and linear regression [12]. Adam et al. (2016) developed a

classification model for ischemic stroke using DT and KNN [13]. Sailasya et al. (2021) built a machine learning model for predicting stroke in which the accuracy of naive Bayes is about 82% [14]. Dev et al. (2022) utilized principal component analysis and perception neural networks for stroke prediction in electronic health records [15]. Most of these studies do not conduct detailed research on features and lack in-depth exploration of the correlation and importance between different features, which may limit the performance of the model. In addition, the lack of public datasets makes it difficult for other researchers to reproduce and validate the results, limiting the development and collaboration of the field.

In addition, existing studies have revealed the association between exposure to heavy metals (e.g., lead, cadmium, manganese) and the risk of cardiovascular disease; for example, Cao et al. (2023) found that higher blood lead and cadmium concentrations were significantly associated with increased stroke mortality [16]. Menke et al. (2006) found that blood lead levels were significantly associated with stroke mortality [17]. Meishuo et al. (2022) showed a prospective association between dietary manganese intake and the risk of cardiovascular mortality in the Japanese population [18].

The purpose of this study is to develop an early prediction algorithm for stroke based on machine learning, which innovatively uses heavy metal content as a potential early predictor. By identifying people at high risk of stroke and administering preventive treatment, it provides a new perspective on individualized stroke prevention and a deeper understanding of the link between environmental factors and brain health. This method is expected to improve the accuracy of stroke prediction, open up a new field of stroke pathophysiology, provide innovative tools for early warning, and promote the development of stroke prevention and treatment strategies.

2. Materials and methods

This section talks about the material and methods used for the preliminary study. It consists of three subsections which respectively discuss dataset description, model setup, and statistical parameters.

2.1. Data

The National Health and Nutrition Examination Survey (NHANES) database created by the Centers for Disease Control and Prevention (CDC) was used. NHANES is an ongoing repeated cross-sectional study using stratified, multistage sampling that assesses the health and nutritional status of adults and children in the United States over two-year cycles [19]. The survey, which began in the 1960s and consisted mainly of interviews and physical examinations, was conducted every two years starting in 1999, with approximately 5000 participants selected annually for data collection [19–21]. The NHANES database is open-access and detailed information can be obtained from the NHANES website (https://wwwn.cdc.gov/nchs/nhanes).

In this paper, demographic data, laboratory data, and questionnaire survey data from NHANES were included for four consecutive survey periods (2017–2020). After downloading the data from the NHANES website, we converted the data from XPT to CSV (comma separated values) format using R and connected all files via primary key (SEQN). Participants under 20 years of age with missing stroke data and uncertain whether they had a stroke were excluded from this study. In addition, participants with missing blood heavy metal data were also excluded. After screening, a total of 2735 subjects were included in this study. Figure 2 shows the screening process.

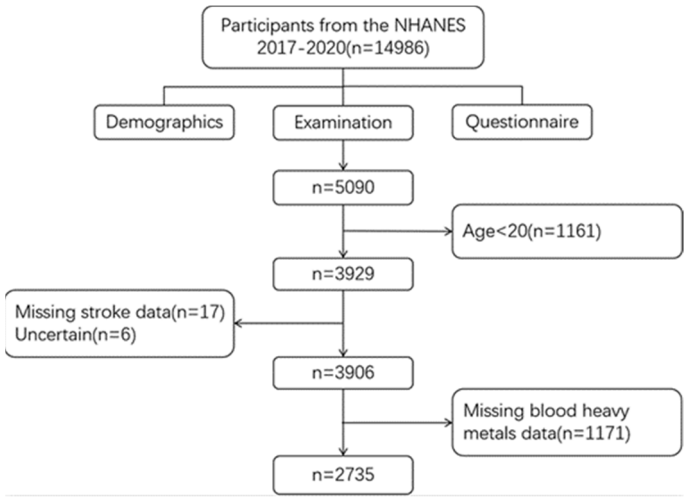

Figure 2. Flowchart of study subject screening.

The study variables for early stroke prediction are based on previous prediction models, literature, and the potential risk factors associated with stroke patients identified from consultation with clinical medical experts. In this paper, five heavy metals are innovatively included as variables in the eigenvector, and these risk factors are objective parameters that can be measured normally during medical treatment. In this paper, these characteristics were obtained from the NHANES database, and a total of 81 study variables were extracted, including 80 independent variables and 1 target variable. The target variable is used to distinguish whether a participant has had a stroke or not. Table 1 lists the characteristic variables included in the study dataset as well as the indicators and interpretations of the characteristic variables.

Table 1. The types of features and metrics contained in the dataset and their definitions.

Type	Indicators and Interpretations
Questionnaire	Stroke, High blood pressure, Cholesterol, Diabetes, Height, Weight, Smoke, Hepatitis B, Hepatitis C, Weak or failing kidneys, Kidney stones, Sleep hours-weekdays, Sleep hours-weekends, Snore, Snort or stop breathing, Trouble sleeping, Feel overly sleepy during day, Excessive drinking, Asthma, Anemia, Overweight, Blood transfusion, Arthritis, Congestive heart failure, Coronary heart disease, Angina, Heart attack, Thyroid problem, COPD/emphysema/ChB, Liver condition, Abdominal pain past 12 mos, Gallstones, Gallbladder surgery, Cancer or malignancy, Close relative had asthma, Close relative had diabetes, Close relative had heart attack, now controlling or losing weight, now increasing exercise, now reducing salt in diet, now reducing fat in diet
Demographics	Gender, Age, Race, Education level, Marital status
Examination	Blood lead, Blood cadmium, Blood mercury, Blood selenium, Blood manganese, Albumin, Creatinine, Albumin creatinine ratio, White blood cell count, Lymphocyte percent, Monocyte percent, Segmented neutrophils percent, Eosinophils percent, Basophils percent, Lymphocyte number, Monocyte number, Segmented neutrophils number, Eosinophils number, Basophils number, Red blood cell count, Hemoglobin, Hematocrit, Mean cell volume, Mean Cell Hgb Conc, Mean cell hemoglobin, RDW, Platelet count, Mean platelet volume, Nucleated red blood cells, Glycohemoglobin, Fasting Glucose, HS CReactive Protein, Insulin, Total Cholesterol

For continuous variables, the Shapiro-Wilk test is used to evaluate whether the variables conform to a normal distribution, and the distribution density function of the continuous variables is plotted, as shown in Figure 3. According to the test results, except for the five continuous variables, Albumin, Albumin-creatinine ratio, Lymphocyte number, HS CReactive Protein, and Insulin, the continuous variables did not conform to the normal distribution. When continuous variables do not conform to the normal distribution, it may lead to bias in the estimation of fit model parameters, failure of confidence intervals and hypothesis tests, and a decrease in the prediction accuracy of the model. To address these challenges, there are several approaches that can be taken. One approach is to use non-parametric models, which do not depend on the distribution of the data and are therefore adaptable to non-normally distributed data. Another strategy is to transform the variables, using such methods as logarithmic transformations, power function transformations, and Box-Cox transformations to improve the distribution of the data and make it closer to the normal distribution. In addition, statistical models that are specifically suitable for non-normal data, such as generalized linear models (GLM) or generalized additive models (GAM), can be selected to more flexibly adapt to the distribution characteristics of the data. For machine learning models, some algorithms have loose assumptions about data distribution, such as decision trees, random forests, etc., which have a certain degree of robustness and can process non-normally distributed data to a certain extent.

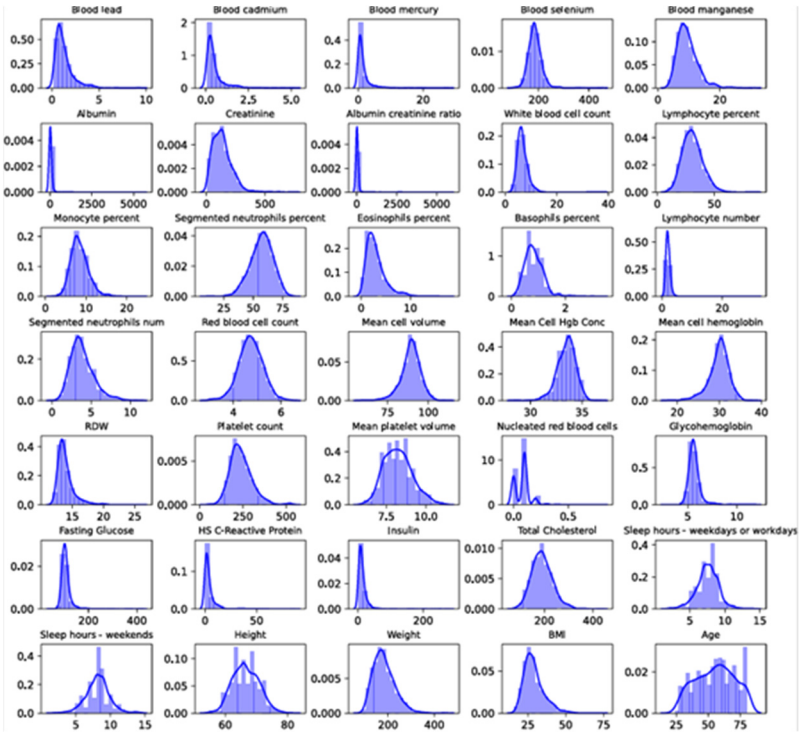

Figure 3. Density plot of continuous variable distribution.

2.2. Feature engineering

Feature engineering is an important part of machine learning applications which includes data preprocessing, feature selection, and dimensionality reduction, and which plays a key role in the performance of machine learning algorithms [22]. In order to extract and select the most valuable

information for stroke risk prediction from the raw data, feature engineering is necessary.

2.2.1. Data preprocessing

In machine learning modeling, the validity and reliability of the model are significantly affected by the integrity of the data. Many algorithms do not have the ability to handle missing data on their own, which makes the handling of missing values in the preprocessing phase critical. In view of the large number of missing samples, this paper chooses to directly eliminate them. For features with a small proportion of missing data and practical significance, KNN is used to fill in the missing data [23], which is a statistical technique that uses the corresponding eigenvalues of similar observations in the data to fill in the missing data. In this paper, the Euclidean distance is used to measure the similarity, and K nearest instance vectors are found, and the weighted average is used for continuous variables and majority voting for categorical variables.

In the field of machine learning, when dealing with continuous numeric features, it is common to encounter situations where different features have different magnitudes. To overcome this, normalizing the data is a common pre-processing step. In this article, we use min-max scaling to deal with continuous variables. This technique is a method of linear transformation based on the extreme values of the feature data, remapping all data values to a range of 0 to 1. In this way, we are able to ensure the scale consistency of features and facilitate the fast and efficient convergence of models based on gradient descent optimization algorithms.

Because the label values for the binary classification features in the dataset are "1" and "2", where "1" means "yes" and "2" means "no", in order to facilitate subsequent analysis, the label value indicating "no" was changed to "0". In addition to dichotomous features, there are also multiple categorical features in the dataset, such as race (race), education level (education level), and marital status (marital stage). In order to ensure that these multiple categorical features can be properly processed in the analysis process, one-hot encoding is used for transformation.

2.2.2. Feature filtering

Feature screening can reduce the complexity of the model, improve the explainability of the model, reduce data noise, and save computing resources, thereby improving the accuracy and reliability of the early stroke prediction model. Feature selection methods can generally be divided into three categories: filter, wrapper [24], and embedded [25]. In this paper, one of the three methods is selected for feature screening.

Logistic regression (LR) is used as a filter-based feature selection method, utilizing statistical analysis to evaluate the correlation between features and the target variable. We employed univariate LR for feature selection and obtained the importance scores of the features. To determine the optimal number of features, we incrementally added features and assessed the model's performance using 10-fold cross-validation to ensure robustness. We calculated the average score for each feature subset and plotted the relationship between the number of features and the cross-validation score, as shown in Figure 4. The plot reveals that the highest cross-validation score is achieved with 16 features, reducing the required features from 81 to 16 without sacrificing model accuracy. Additionally, we visualized feature importance using a bar chart with F-values as the metric, illustrated in Figure 5.

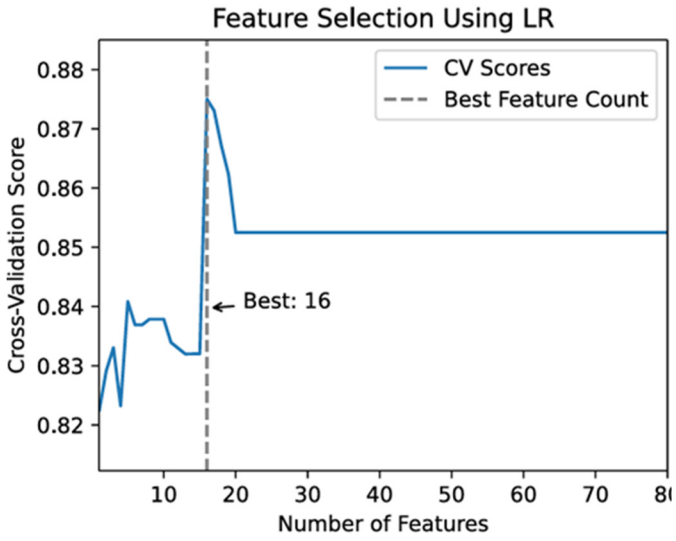

Figure 4. A plot of the relationship between the number of features and the cross-validation score.

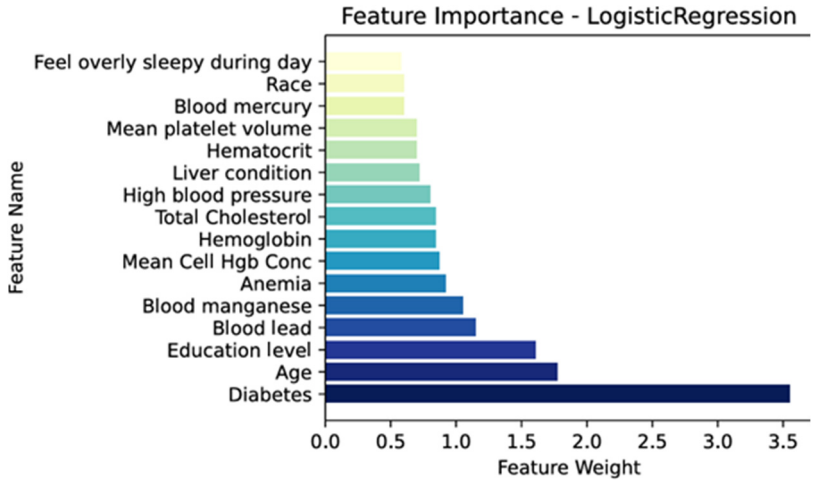

Figure 5. Feature importance map.

Elastic net regression is a hybrid model that combines lasso regression and ridge regression, which is trained by L1 regularization and L2 regularization as an embedded feature selection method. Its hybrid parameter (r) controls the proportion of L1 and L2 regularization, and the elastic network has better flexibility than using lasso or ridge regression alone, balancing feature selection and model complexity. The loss function is shown in Eq (1):

$$\mathcal{L}(\beta) = \frac{1}{2n}\sum_{i=1}^{n}\left(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\,\beta_j\right)^2 + \lambda\left[\frac{(1-\alpha)}{2}\sum_{j=1}^{p}\beta_j^2 + \alpha\sum_{j=1}^{p}|\,\beta_j|\right] \quad (1)$$

The first term minimizes the squared error between the predicted and actual values; the second term is the L2 regularization term, which encourages the model to generate sparse solutions; and the third term is the L1 regularization term, which is used to control the complexity of the model. In this paper, the L1/L2 ratio is set to 0.5, the nested loop is used to iterate over each value, and the elastic

network regression model is fitted on 80 samples to calculate the average coefficient at each value. Next, the number of selected features for each value was calculated. After that, the coefficient path plot with the abscissa as the logarithmic scale and the ordinate as the regression coefficient is shown in Figure 6, and the effects of different variables on the regression coefficient were observed. Next, we performed a 10-fold cross-validation, calculated the root mean square error (RMSE) of the cross-validation, and plotted a curve between alpha and RMSE to select the best alpha value. The model is trained with the best alpha value to extract the characteristic coefficients. The number of features corresponding to the optimal alpha value is shown in Figure 7. We then sort according to the absolute value of the characteristic coefficients, and output the feature sorting result. Because there are 20 features corresponding to the optimal alpha value, the first 20 features are selected to draw the importance diagram of the feature coefficients and display them visually, as shown in Figure 8.

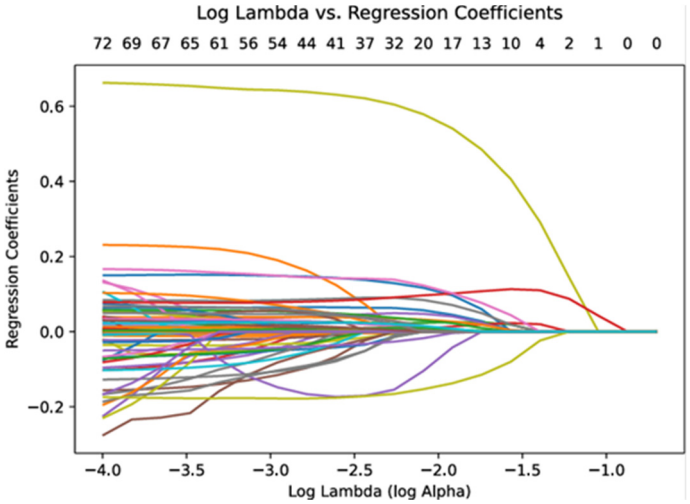

Figure 6. Coefficient path diagram. Lines of different colors represent the coefficient trajectories of individual features as the regularization penalty (λ) increases. Selected feature indices are annotated on the right.

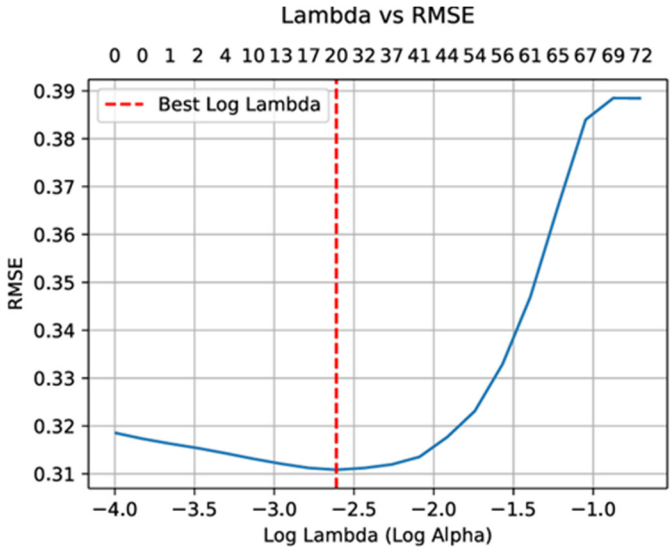

Figure 7. The number of features corresponding to the optimal alpha value.

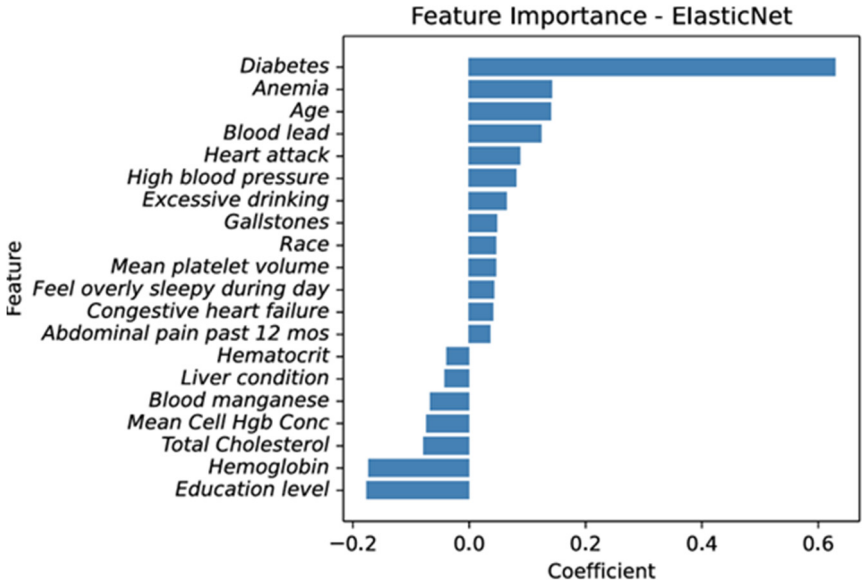

Figure 8. Feature importance plot of elastic net regression.

Recursive feature elimination with cross validation (RFECV), a wrapper method, combines the feature selection capability of recursive feature elimination (RFE) and the evaluation mechanism of cross validation (CV). RFECV automatically identifies the optimal feature subset through cross-validation, thereby balancing model performance with reduced complexity and overfitting risk. Compared to RFE alone, RFECV accounts for the model's generalization ability, leading to more robust feature selection results. In this study, a random forest (RF) model is set as the base model, with a maximum of 1000 iterations, and all features are included for RFE. The RFECV selector is then instantiated for feature selection. Consistent with the above approach, the optimal number of features is determined by plotting the relationship between the number of features and the cross-validation score, as shown in Figure 9. We achieve the maximum cross-validation score with 9 traits. The feature visualization is shown in Figure 10.

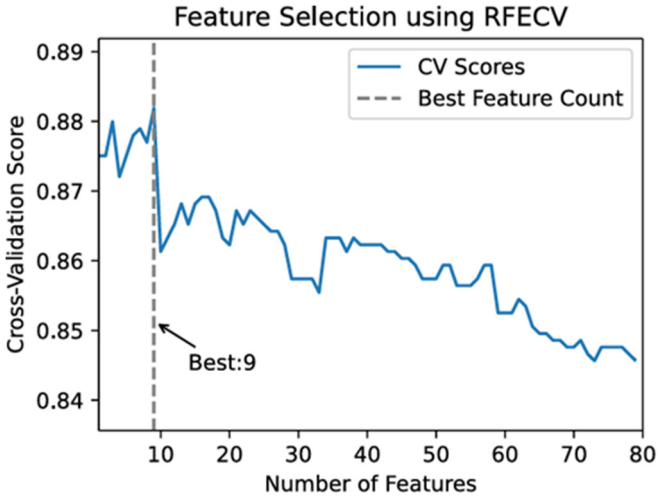

Figure 9. A plot of the relationship between the number of features and the cross-validation score.

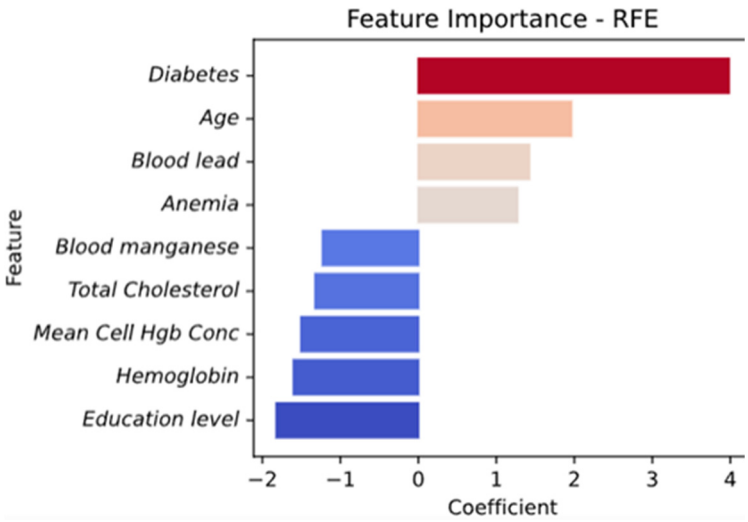

Figure 10. The number of features corresponding to the optimal alpha value.

By using different feature selection methods, we consider different angles and principles to screen out the important features respectively, as shown in Table 2.

Table 2. The number and name of features corresponding to the three methods of feature filtering.

Method	Number	Filtered features
LR	16	Diabetes, Age, Education level, Blood lead, Blood manganese, Anemia, Mean Cell Hgb Conc, Hemoglobin, Total Cholesterol, High blood pressure, Liver condition, Hematocrit, Mean platelet volume, Blood mercury, Race, Feel overly sleepy during day
Elastic Net Regression	20	Diabetes, Education level, Hemoglobin, Anemia, Age, Blood lead, Heart attack, High blood pressure, Total Cholesterol, Mean Cell Hgb Conc, Blood manganese, Excessive drinking, Gallstones, Race, Mean platelet volume, Feel overly sleepy during day, Congestive heart failure, Liver condition, Hematocrit, Abdominal pain past 12 mos
RFECV	9	Diabetes, Age, Education level, Hemoglobin, Mean Cell Hgb Conc, Blood lead, Total Cholesterol, Anemia, Blood manganese

Comparing these screening results, we found that some features were selected among multiple methods, indicating that they were of high importance and had good stability. Therefore, we use features that appear twice or more times as the final selection features to ensure that the selected features are more representative and discriminatory. In the end, we obtained 15 characteristics with importance and stability: Diabetes, Age, Education level, Blood lead, Blood manganese, Anemia, mean cell corpuscular hemoglobin concentration (Mean Cell Hgb Conc), Hemoglobin, Total Cholesterol, High blood pressure, Liver condition, Hematocrit, Mean platelet volume, Race, Feel overly sleepy during day. The final characteristics of the screened are shown in Table 3.

In order to select features, a correlation analysis method is used in this paper. The Spearman correlation coefficient was used to calculate the correlation between features. The absolute values of Spearman's correlation coefficients are weakly correlated at [0,0.3], moderately correlated, and strongly correlated at [0.3,0.7], and [0.7,1]. First, according to the range of correlation coefficients, the correlation coefficient threshold was set to 0.7, and the combination of features with correlation

coefficients greater than 0.7 with other features was found. Next, the features with the higher correlation coefficient with the label were retained, and the lower ones were eliminated. The correlation plot is shown in Figure 11. As can be seen in Figure 11, the correlation coefficient between Hemoglobin and Hematocrit is 0.98. In accordance with the above rules, hematocrit is a feature that is excluded in this paper. In the end, we retained the 14-dimensional features and used them for the construction of the model, and the final subset of variables is shown in Table 4.

Table 3. Final features after screening.

Number of repetitions	Features
3	Diabetes, Age, Education level, Blood lead, Blood manganese, Anemia, Mean Cell Hgb Conc, Hemoglobin, Total Cholesterol
2	High blood pressure, Liver condition, Hematocrit, Mean platelet volume, Race, Feel overly sleepy during day
1	Heart attack, Excessive drinking, Gallstones, Congestive heart failure, Abdominal pain past 12 mos, Blood mercury

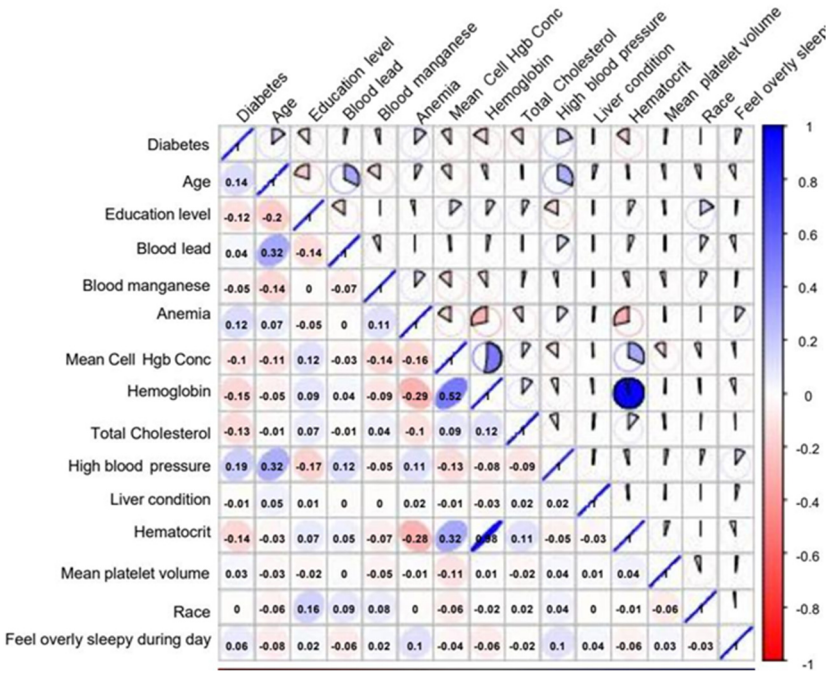

Figure 11. Feature correlation map.

Table 4. Number and type of features corresponding to the three methods of feature filtering.

Feature	Type
Diabetes	Categorial
Age	Numeric
Education level	Categorical
Blood lead	Numeric

Continued on next page

Feature	Type
Blood manganese	Numeric
Anemia	Categorical
Mean Cell Hgb Conc	Numeric
Hemoglobin	Numeric
Total Cholesterol	Numeric
High blood pressure	Categorical
Liver condition	Categorical
Mean platelet volume	Numeric
Race	Categorical
Feel overly sleepy during day	Categorical

2.2.3. Class imbalance data processing

Class imbalance is one of the common problems in medical research. In datasets with unbalanced categories, classifiers tend to focus on the majority and ignore or treat instances of the minority as noisy data. This increases the rate of misclassification of minorities, which are often more critical in medical informatics applications. To solve this problem, this paper uses oversampling and undersampling methods to balance different classes of samples and combines them with classifier models for performance evaluation, and compares and analyzes different sampling techniques.

Oversampling and undersampling are two commonly used data- driven sampling methods to deal with class imbalance. In this paper, two synthetic oversampling methods are used: synthetic minority oversampling technique (SMOTE) and adaptive synthetic sampling (ADASYN), as well as two commonly used undersampling techniques, NearMiss and ClusterCentroids. An algorithm-driven approach is one that deals with class imbalances. This method does not change the distribution of input data, and the classification algorithm is adjusted to specifically deal with minority classes, mainly including costsensitive learning, threshold setting, and mixed methods (such as ensemble learners). Cost-sensitive learning is used in this study. An important concept in cost-sensitive learning is the cost matrix, which is represented by FP (false positive), FN (false negative), TP (true positive) and TN (true negative). Taking binary classification as an example, a two-dimensional matrix can be used to describe the prediction results of the classification algorithm as shown in Table 5, where $\mathbf{C}(i,j)$ represents the cost of misclassification to classify the instance, i is the prediction class, and j is the actual class.

Table 5. Example of a binary classification cost matrix.

	Actual negative	Actual positive
Predict negative	C(0,0) or TN	C(0,1) or FN
Predict positive	C(1,0) or FP	C(1,1) or TP

The dataset was divided into a training set and a test set to evaluate the performance and generalization ability of the model, with the test set accounting for 30% of the total dataset. The five imbalanced data sampling methods introduced above were applied to the random forest classifier for evaluation. To evaluate the performance of each method, a hierarchical 10-fold cross-validation was used to train and evaluate the model multiple times on different training and validation sets. In order

to compare the effects of different methods, a classification report of each method was output, as shown in Table 6. From the results of the four evaluation indicators, it can be seen that cost-sensitive learning has the strongest performance in all indicators, especially in the accuracy and recall rate of important indicators. High recall rates are often important in stroke medical settings, as the cost of missing a positive case (not diagnosing an actual patient) is significant. At the same time, this method maintains a high accuracy rate, reducing the possibility of false positives.

Table 6. Performance metrics for different sampling methods.

Method	Accuracy	Precision	Recall	F1 Score
SMOTE	0.95	0.95	0.94	0.95
ADASYN	0.95	0.93	0.97	0.95
Near Miss	0.87	0.91	0.81	0.86
Cluster Centroids	0.77	0.77	0.78	0.77
Cost-sensitive learning	0.98	0.96	0.99	0.98

In addition, by plotting the receiver operating characteristic (ROC) curves corresponding to different sampling techniques in the random forest classifier, it is possible to visually compare their performance differences. The ROC curves for the five sampling techniques are shown in Figure 12. As can be seen from the ROC curve, the area under the curve (AUC) value of the cost-sensitive method is the highest among the five methods. Overall, cost-sensitive learning provides the best performance.

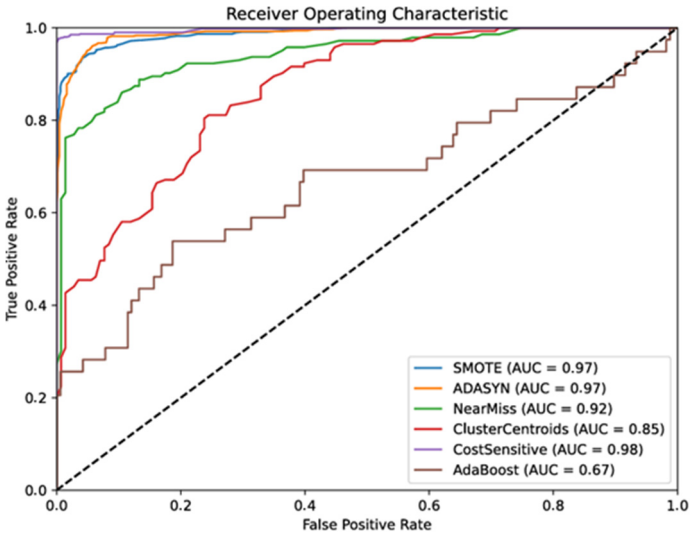

Figure 12. ROC curves for five sampling techniques.

2.3. Machine learning classification models

To comprehensively evaluate the predictive power of the selected features for stroke risk, we employed and compared a suite of seven machine learning algorithms. This selection encompasses both well-established traditional models and powerful ensemble methods, enabling a robust assessment of linear, non-linear, and complex interaction effects within the data. The models were implemented using scikit-learn and XGBoost libraries in Python with hyperparameters optimized via grid search.

2.3.1. Traditional machine learning models

Support vector machine (SVM): a maximum-margin classifier effective in high-dimensional spaces. It was chosen to investigate the potential existence of a clear nonlinear boundary separating stroke and non-stroke cases in our feature space.

Naive Bayes (NB): a probabilistic classifier based on Bayes' theorem with strong assumptions of feature independence. Despite its simplicity, it serves as a high-baseline model to gauge the inherent predictability of the data.

Logistic regression (LR): a linear model that provides well calibrated probability estimates. Its high interpretability allows for direct assessment of feature effect sizes and directions, serving as a critical benchmark.

Multilayer perceptron (MLP): a foundational class of artificial neural networks capable of learning complex, non-linear relationships. It was included to represent the performance of basic deep learning architectures on this tabular dataset.

2.3.2. Integrate machine learning models

Ensemble methods, which combine multiple base learners to improve generalization, were a primary focus due to their proven efficacy in biomedical prediction tasks.

Random forest (RF): a bagging-based ensemble of decision trees that introduces randomness in both sample and feature selection. It is renowned for its robustness to overfitting, ability to handle mixed data types, and intrinsic feature importance measures, making it a strong candidate for our clinical dataset.

Gradient boosting decision tree (GBDT): a boosting-based model that sequentially builds trees to correct the errors of previous ones. Its iterative nature often yields high predictive accuracy.

Xtreme gradient boosting (XGBoost): an advanced and optimized implementation of gradient boosting that incorporates regularization to prevent overfitting. It is frequently a top performer in structured data competitions and was included as a state-of-the-art benchmark.

All models were trained on the same processed feature set and evaluated using a consistent framework, as detailed in Section 2.4. This comparative approach allows us to identify not only the best-performing algorithm but also to understand the underlying data patterns captured by different modeling philosophies.

2.4. Model evaluation criteria

This paper comprehensively evaluates the performance of seven machine learning models, including Accuracy, Precision, Recall, F1 Score, ROC, and AUC, as well as multiple evaluation indicators, including the calibration curve and brier score. These metrics look at the performance of the model across multiple dimensions, ensuring that we can comprehensively analyze and compare the strengths and weaknesses of each model, as defined below:

$$Accuracy = \frac{TP+TN}{TP+FP+TN+FN} \tag{2}$$

$$Precision = \frac{TP}{TP+FP} \tag{3}$$

$$Recall = \frac{TP}{TP+FN} \tag{4}$$

$$F1 = 2 \cdot \frac{Precision \cdot Recall}{Precision+Recall} \tag{5}$$

$$Brierscore = \frac{1}{N}\sum_{t=1}^{N}(f_t - o_t)^2 \tag{6}$$

Among them, TP is the number of true positives, TN is the number of true negatives, FP is the number of false positives, FN is the number of false negatives, f_i is the predicted probability, and o_i is the observation probability.

Accuracy reflects the likelihood that a prediction will be correct, while precision measures the proportion of a prediction judged to be positive that is truly positive. The recall rate measures the rate of all actual positive samples that were correctly predicted to be positive. The F1 score provides a balanced metric for the blended average of accuracy and recall. The ROC curve, together with the AUC value, reveals the generalization efficiency of the classifier. The calibration curve is a graphical representation of the Hosmer-Lemeshow goodness-of-fit test, which shows the degree of matching between the predicted probability and the actual probability, and the calibration curve close to $y = x$ and a lower Brier score together indicate the higher prediction accuracy of the model.

3. Early prediction models based on machine learning

3.1. Comparative analysis of model performance

Based on the 14-dimensional features screened in 2.2.3, this paper constructs an early prediction model for stroke patients on 70% of the training set and uses grid search and 10-fold cross-validation for hyperparameter tuning. Subsequently, the performance of the model was evaluated on the remaining 30% of the test set, as shown in Table 7. As can be observed from the table, the random forest model outperforms other machine learning models in terms of accuracy, precision, recall, and F1 score. Furthermore, among the ROC curves and AUC values of the seven prediction models for stroke patients shown in Figure 13, it can be clearly seen that the AUC value of random forest is still the highest. This indicates that compared with other models, random forests have the strongest generalization ability and can better distinguish between positive and negative patients.

Table 7. Performance metrics for different models.

Model	Accuracy	Precision	Recall	F1 Score
RF	0.96	0.93	0.98	0.95
XGB	0.92	0.87	0.98	0.92
GBDT	0.90	0.84	0.96	0.90
NB	0.71	0.96	0.41	0.57
LR	0.79	0.78	0.77	0.77
SVM	0.80	0.79	0.77	0.78
MLP	0.80	0.78	0.80	0.79

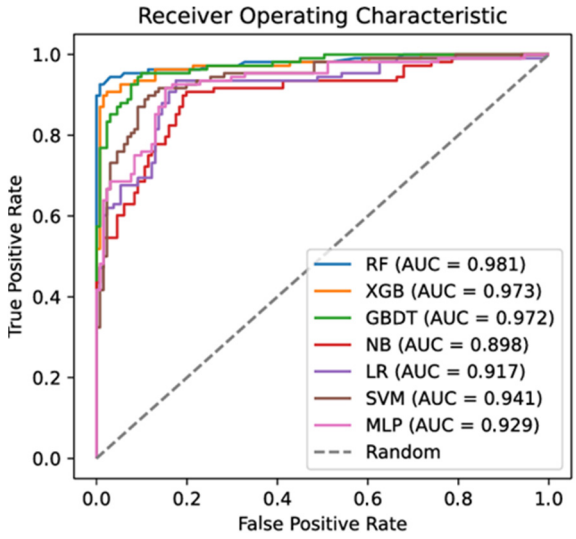

Figure 13. Comparison of ROC curves of the model.

Although the model has good discrimination ability and shows high prediction accuracy on specific samples, its performance may not be consistent across different sample sets. Therefore, it is particularly important to evaluate the consistency of the model across multiple sample sets, that is, to measure the calibration of the difference between the predicted value and the true value. To quantify calibration, the Brill score becomes a key metric, and visual analysis of calibration curves helps to see how well the predictions match the actual observed probabilities. Figure 14 illustrates the calibration curve plotted in this paper, constructed by arranging and dividing the predicted probabilities into ten equal parts for the samples in the test set. The abscissa of each point in the graph represents the mean of the predicted probability of patients in the corresponding interval, and the ordinate represents the actual probability of all patients in the interval. We observed that the calibration curve of the random forest was close to ideal and relatively stable. At the same time, it has a Brier score of 0.06, the lowest among the 7 models.

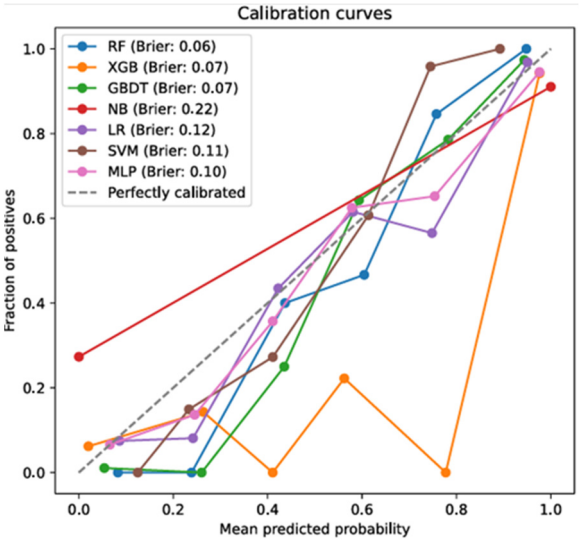

Figure 14. Comparison of calibration curves for the model.

3.2. Interpretability analysis based on Shapley additive explanations (SHAP) theory

3.2.1. SHAP

Many machine learning models function as "black boxes", rendering their predictions difficult to interpret. In such models, inputs yield predictions without revealing the underlying rationale. SHAP, introduced by Scott et al. in 2017, offers insights into machine learning model outputs. SHAP merges Shapley value concepts from game theory with local interpretability measures to assess feature contributions to predictions. By assigning Shapley values to individual features, SHAP clarifies the influence of each feature on the prediction outcome, making the decision-making process transparent.

SHAP is an additive feature attribution method that represents the interpretation of Shapley values in a linear model. For the Shapley value calculation of features, x_j is the weighted sum of all feature combinations, and the calculation formula is as follows:

$$\phi_j = \sum_{s \subseteq [x_1,\ldots,x_p] \setminus \{x_j\}} \frac{|\mathbb{S}|!(p-|S|-1)!}{p!} (f(S \cup \{x_j\}) - f(S)), \tag{7}$$

where $x_1, x_2, \ldots, x_p$ is the set of all features, p is the number of features, S is the subset of all possible features, x_j is the feature vector to be explained, S is the predicted value of the model under subset $f(S)$, and $|S|$ is the size of the subset. For the predicted value of the model $f(x)$, it is expressed as

$$f(x) = F(z') = \phi_0 + \sum_{j=1}^{p} \phi_j z_j', \tag{8}$$

where F is the explanatory model, $z' \in \{0,1\}^p$ is the eigen-alliance vector, and Φ_j is the Shapley value of feature j.

In this paper, random forest is applied as a multi-tree-based ensemble method, and although it exhibits high prediction accuracy, its model structure complexity limits the interpretability. In order to enhance the explanatory power of the model, we used SHAP analysis technology to evaluate the random forest model we constructed from both macro and micro levels. This approach can help explain not only the behavior of the model as a whole but also the attribute contributions behind individual predictions.

3.2.2. An overall interpretation of the model

Figure 15 illustrates the overall interpretation of the random forest model using SHAP theory. Each row represents a feature, each dot represents a sample, the shade of color represents the size of the feature value, the left side of the figure shows the feature name, and the abscissa represents the size of the Shapley value. The Shapley value of a feature of a sample is greater than zero, which means that the feature has a positive effect on the prediction results of the sample, and vice versa. The figure shows the ranking of the importance of the 14 selected features, decreasing in importance from top to bottom. The ranking of feature importance is obtained by sorting by the absolute mean of Shapley values for all samples. Diabetes, High blood pressure, Age, Hemoglobin, Total Cholesterol, Education level, and Blood lead all played an important role in the prediction model. The graph also shows that lower Total Cholesterol, Education level, and Mean Cell Hgb Conc lead to higher stroke rates. For Hemoglobin, Blood lead, Mean platelet volume, Blood manganese, and Race, it was observed that the

color display did not follow a trend from light to dark with Shapley values, suggesting a lack of a clear linear correlation between eigenvalues and Shapley values. This phenomenon indirectly reflects the ability of the random forest model to capture the nonlinear relationship between the target and the feature, so as to achieve excellent prediction performance.

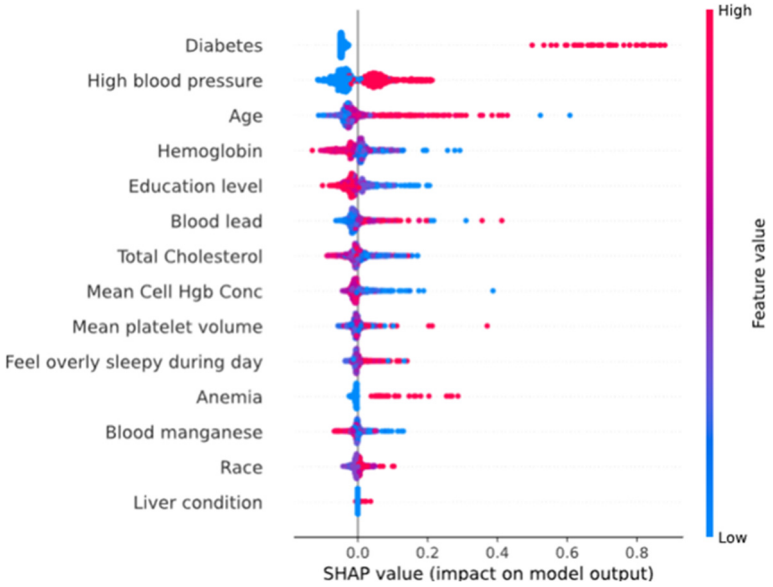

Figure 15. Overall explanation of each feature in the SHAP framework.

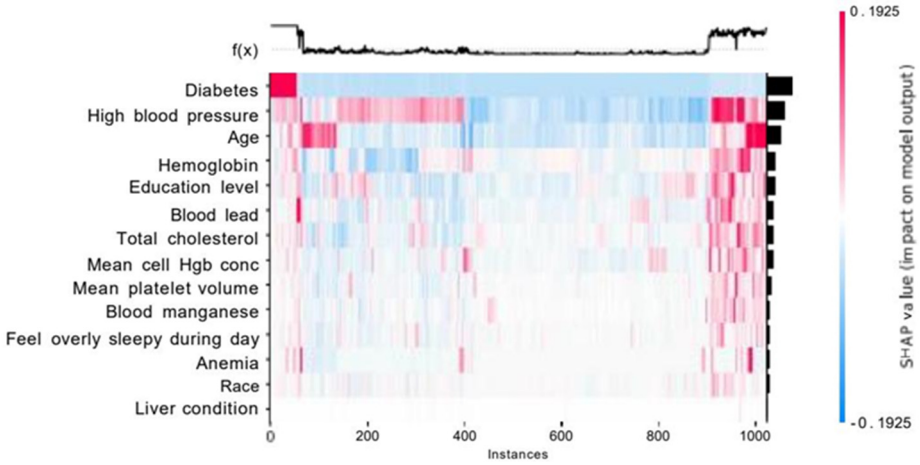

Figure 16. Heatmap under the SHAP framework.

Figure 16 shows the heatmap of a random forest model using the SHAP theory. Heatmaps are designed to use supervised clustering to show the overall substructure of a dataset and the correlation of features. Compared to simple clustering, supervised clustering does not use the original eigenvalues of data points but uses their Shapley values for hierarchical clustering and returns the order after clustering. Putting data points with similar Shapley values together to form a sort of heatmap can help show the relationship between model predictions and features more clearly, revealing the structure and patterns of the dataset. In the figure, the horizontal and vertical coordinates represent the instance and

the model input, respectively. They are color-coded to indicate the impact of a feature on the model output; red to indicate the contribution of the feature to increasing the model output (i.e., the current feature increases the risk of stroke) with a positive Shapley value; and blue to indicate the contribution of the feature to reducing the output of the model (i.e., the current feature reduces the risk of stroke), in which case the Shapley value is negative. As you can see from the graph, the heatmap groups together samples with the same etiology and the same model output, that is, patients with a similar probability of having a stroke.

3.2.3. Individual interpretation of the model

The interpretability analysis of a single sample is achieved by calculating the SHAP values of each feature of the sample to be explained, and this process reveals the degree to which they contribute to the prediction results of the sample by comparing the size of the SHAP values of each feature and their positive and negative values. The predicted value of a single sample is calculated by adding up the SHAP values of all its features and adding the average predicted value of the model.

Figure 17 shows a waterfall chart of the scores of a single stroke patient. The horizontal axis is the SHAP value, and the vertical axis is the value of each feature of the sample. Blue indicates that the feature has a negative effect on the prediction (arrow to the left, the SHAP value decreases), and red indicates that the feature has a positive effect on the prediction (arrow to the right, the SHAP value increases). $E[f(x)]$ represents the SHAP benchmark value, which is the mean value predicted by the model. As can be seen from Figure 17, the SHAP values of Total Cholesterol, Hemoglobin, Mean platelet volume, and Mean Cell Hgb Conc are positive and located in the red area, indicating that these characteristics have a positive effect on the occurrence of stroke in this patient. Diabetes, Age, Hypertension, Blood lead, and Ethnicity have negative SHAP values in the blue area, indicating that these features have a negative effect on the development of stroke in this patient. The negative effects of all features are significantly greater than the positive effects, reflecting the extremely low probability of stroke in this patient.

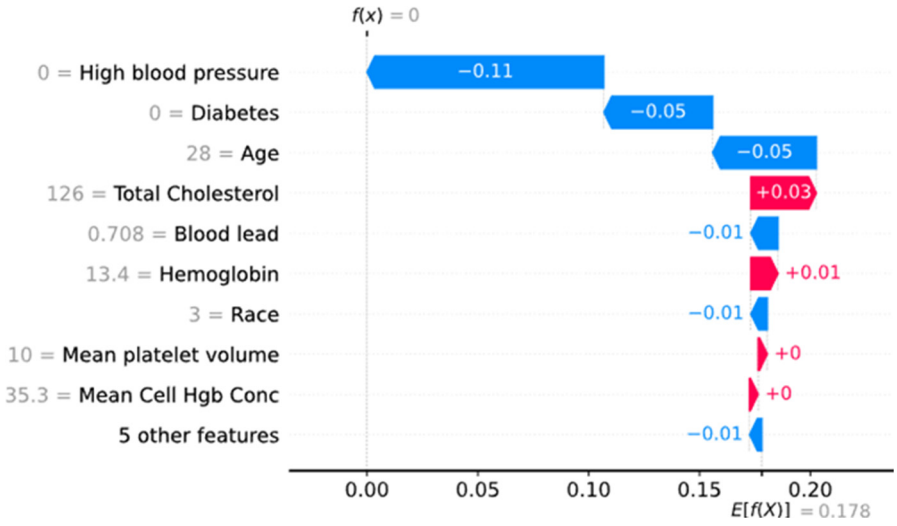

Figure 17. Waterfall chart of a single stroke patient.

The explanatory force map provides a direct view of the specific impact of each feature on the prediction outcome. In this paper, we selected two cases with stroke from the test set for an in-depth interpretability analysis, and the explanatory power plots for these two cases are shown in Figure 18(a),(b). The bar lengths of the features in the graph represent the relative magnitude of the SHAP values, the base value represents the average predicted value of the model for all samples, and $f(x)$ represents the output value of sample x to be interpreted. As can be seen from Figure 18, the negative effect of all characteristics of both patients is greater than the positive effect, and the likelihood of stroke is small.

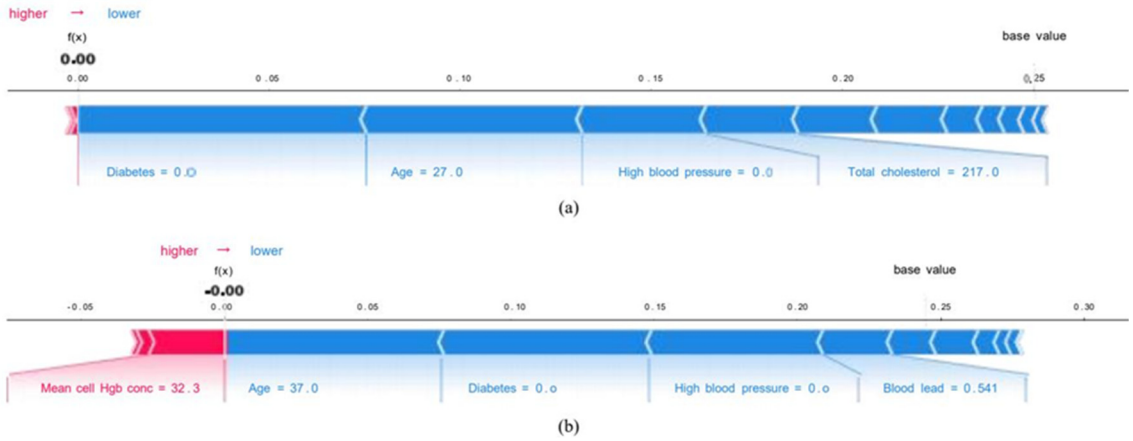

Figure 18. Explanatory power in stroke patients.

3.3. Association of heavy metals with stroke

The preliminary results of this study show that blood lead and blood manganese levels have a significant correlation with the occurrence of stroke. This finding points to potential biomarkers that may be critical for the identification of risk factors and the early prediction of stroke. Although we have noted an association between blood lead and manganese levels and stroke, further research is needed to determine whether there is a causal relationship. Multivariate regression was performed to examine the association between blood lead and blood manganese levels and stroke, including covariates as confounders. After adjusting for multiple confounding factors in the model, a generalized additive model (GAM) was used to fit a smooth curve with the dependent variable as a continuous variable to analyze the dose- response correlation.

In this study, a logistic regression model was used to further analyze the relationship between blood lead and stroke, and maximum likelihood estimation (MLE) was used to estimate the parameters. We obtained a Z-score of 4.987 for the coefficient of blood lead divided by its standard error, corresponding to a p-value of 0.000, indicating that the association between blood lead levels and stroke is statistically significant. In addition, the 95% confidence interval for the effect of blood lead levels on stroke risk ranged from 0.227 to 0.522, excluding 0, supporting a significant impact of these variables on stroke risk.

Nonlinear relationships are common in many studies of hormone levels, drug doses, and diseases, as they are often not simple linear relationships but rather quantitative to qualitative changes. This phenomenon of quantitative change to qualitative change is often manifested as a threshold effect; that is, with the increase of factor *X*, the magnitude of the effect on *Y* will change.

Therefore, the study of threshold effect has become the highlight of some studies. By plotting the dose-response curve, we can see a positive correlation between blood lead concentration and stroke risk, as shown in Figure 19.

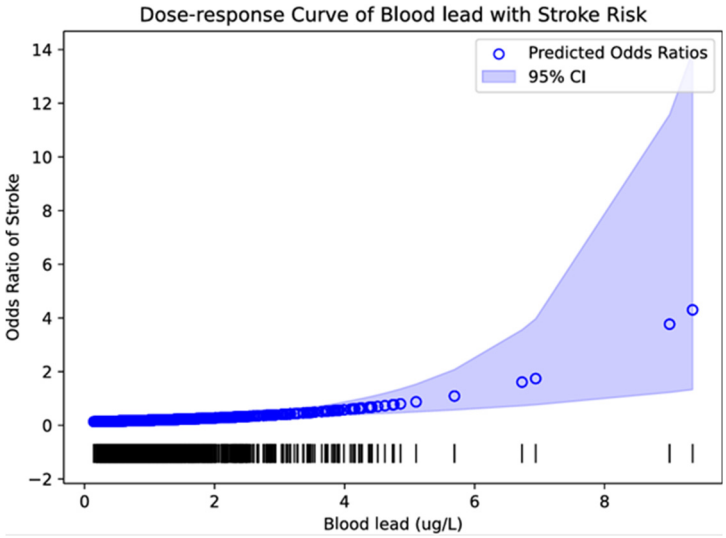

Figure 19. Dose-response curve of blood lead concentration versus stroke risk.

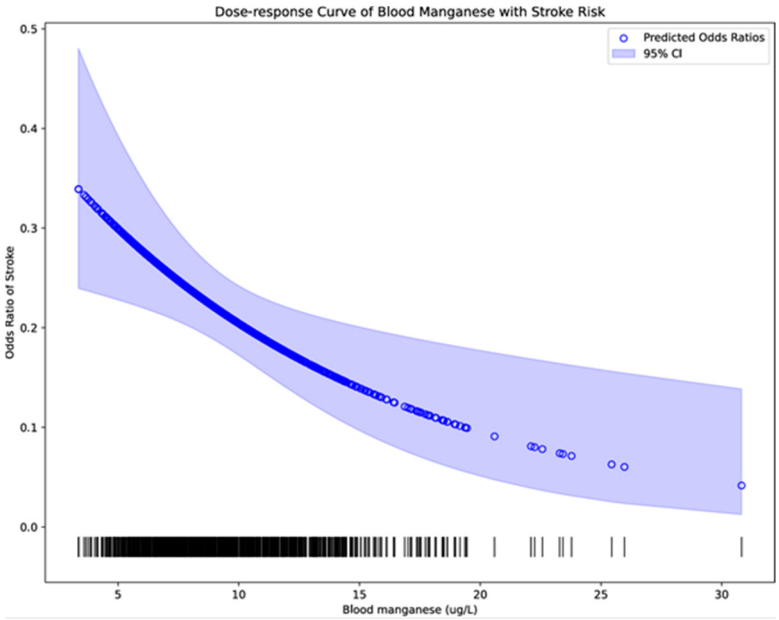

Figure 20. Dose-response curve of blood manganese concentration versus stroke risk.

The logistic regression model was also used to further analyze the relationship between blood manganese and stroke, and the MLE was used to estimate the parameters. We arrived at an estimated coefficient of 0.0764 for blood manganese, a standard error of 0.028, a Z-value of 2.741, and a corresponding p-value of 0.006, indicating that the effect of blood manganese is statistically significant. In addition, the 95% confidence interval ranged from 0.131 to 0.022, excluding 0, supporting the conclusion that blood manganese reduces the risk of stroke. Taken together, this logistic regression

model suggests that the risk of stroke actually decreases as blood manganese levels increase. This conclusion is also supported by the dose-response curve plotted in Figure 20. The dose-response curve shows the change in stroke incidence at different blood manganese levels, providing visual evidence to support the possible impact of blood manganese levels on stroke risk. Such an analysis can help researchers and physicians understand the specific patterns of stroke risk within the range of changes in blood manganese concentrations, which can help determine target blood manganese levels to minimize the probability of stroke.

4. Discussion

This study developed and evaluated machine learning models for early stroke risk prediction by innovatively integrating blood heavy metal concentrations as key predictive features. Through rigorous feature engineering applied to NHANES (2017–2020) data, we identified a 14-dimensional feature subset, including blood lead and blood manganese, which demonstrated stable and significant importance. Among seven machine learning models compared, the random forest (RF) model achieved the best overall performance (Accuracy: 0.96, Recall: 0.98, F1-score: 0.95). More importantly, interpretability analysis using SHAP revealed that elevated blood lead levels were associated with increased stroke risk, whereas higher blood manganese levels were associated with a decreased risk. This opposing association was further statistically confirmed, with blood lead showing a significant positive correlation (95% CI: 0.227–0.522, p-value < 0.001) and blood manganese a significant negative correlation (95% CI: –0.131– –0.022, p-value = 0.006) with stroke outcome.

Our findings regarding the divergent roles of lead and manganese in stroke risk resonate with, yet extend, existing epidemiological evidence. The positive association between blood lead and stroke risk aligns consistently with prior research. Menke et al. demonstrated that even low-level blood lead (below 10 μg/dL) remained a significant predictor of cardiovascular mortality, highlighting its persistent public health threat [18]. Similarly, Cao et al. found that higher concentrations of lead and cadmium were significantly associated with increased stroke mortality [17]. Our results, derived from a machine learning framework that controls for numerous confounders, provide robust supportive evidence for the role of lead as a risk factor.

Conversely, the protective association observed for blood manganese is particularly noteworthy and finds support in nutritional epidemiology. A large prospective cohort study in Japan reported that higher dietary manganese intake was associated with a lower risk of total stroke, ischemic stroke, and cardiovascular disease mortality, independent of various confounding factors [19]. Our study translates this dietary-level association to the biomarker level, directly linking blood manganese concentration with reduced stroke risk in a U.S. population. This convergence of evidence from different populations and methodologies strengthens the hypothesis that manganese may play a beneficial role in cardiovascular health.

This study contributes to the field through several methodological advancements. First, we moved beyond merely establishing statistical associations to operationalizing heavy metal exposure within a predictive clinical model. By embedding blood lead and manganese into a high-performance RF classifier, we demonstrate their practical utility in risk stratification. Second, our rigorous multi-stage feature selection process—employing filter (logistic regression), embedded (elastic net), and wrapper (RFECV) methods—ensured the robustness and biological plausibility of the selected features. This approach mitigated the risk of overfitting and enhanced model interpretability compared to studies

using black-box models without feature refinement. Third, we directly addressed the critical challenge of class imbalance common in medical datasets. By systematically comparing various sampling and algorithm-driven techniques, we demonstrated that cost-sensitive learning was most effective for our data, optimizing both accuracy and the clinically crucial recall metric. Finally, the application of SHAP interpretability framework transformed the RF model from a "black box" into an interpretable tool. It allowed us to globally rank feature importance (e.g., confirming diabetes and hypertension as top predictors) and locally explain individual predictions, thereby bridging the gap between model performance and clinical understanding. This dual-level interpretation underscores the specific, non-linear contribution of heavy metals to the model's decisions.

Several limitations of this study should be acknowledged. First, its cross-sectional design using NHANES data precludes the establishment of causal relationships between heavy metals and stroke. The observed associations, while adjusted for multiple confounders, may be influenced by residual confounding or reverse causality. Second, the scope of environmental exposure was limited to five heavy metals measured in blood. Other potentially neurotoxic metals (e.g., arsenic, chromium) or broader environmental factors (e.g., air pollution PM2.5, organic pollutants) were not included and warrant investigation in future, more comprehensive exposure-wide association studies (ExWAS). Third, the model's generalizability requires external validation in independent, prospective cohorts and diverse ethnic populations, as our data was sourced from the U.S. NHANES participants.

Future research should focus on the following: (1) Prospective validation in longitudinal cohorts to assess the temporal relationship and predictive power of the heavy metal signature; (2) Integration of multi-omics data (e.g., genomics, metabolomics) with environmental exposures to uncover underlying biological pathways; (3) Development of real-time clinical decision support tools incorporating these models to enable personalized risk assessment and preventive interventions.

5. Conclusions

Through the combined use of data analysis and machine learning techniques, this paper aims to improve the accuracy of prediction of stroke risk, especially in terms of the inclusion of heavy metal content in environmental exposures as a new risk factor. Based on the in-depth analysis of the NHANES database, this paper not only uses a variety of data processing and feature selection techniques to optimize the performance of the model but also pays special attention to the association between blood lead and blood manganese, two heavy metals, and stroke risk, and deeply explains the driving factors of model prediction through SHAP theory. The results showe that the random forest model performs well in predicting the early risk of stroke, and that blood lead and blood manganese concentrations are associated with stroke risk.

Overall, this paper combines environmental factors and data analysis techniques to provide new perspectives and methods for stroke risk prediction and arrhythmia diagnosis. By focusing on heavy metal exposure, this paper expands the understanding of stroke risk factors, and emphasizes the importance of environmental health in disease prevention and management. At the same time, by applying and improving deep learning models, this paper shows how medical research and clinical practice can benefit from the advancement of artificial intelligence in the era of big data.

Authorship contribution statement

FenQi Liu: Investigation, KeXin Li: Writing-original draft, Software, Formal analysis, Data curation, EnXiao Zhu: Writing-review & editing.

Data availability

Data will be made available on request.

Use of AI tools declaration

The authors declare they have not used Artificial Intelligence (AI) tools in the creation of this article.

Acknowledgments

This work was supported by the National Key R&D Program of China [No. 2024YFB3411500] and the National Natural Science Foundation of China [No. U23A2065].

Conflict of interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

1. Powers WJ, (2020) Acute ischemic stroke. *N Engl J Med* 383: 252–260. https://doi.org/10.1056/NEJMcp1917030
2. A Montaño, DF Hanley, JC Hemphill III, (2021) Hemorrhagic stroke. *Handb Clin Neurol* 176: 229–248. https://doi.org/10.1016/B978-0-444-64034-5.00019-5
3. Johnston SC, (2002) Transient ischemic attack. *N Engl J Med* 347: 1687–1692. https://doi.org/10.1056/NEJMcp020891
4. RL Sacco, SE Kasner, JP Broderick, LR Caplan, JJ Connors, A Culebras, et al. (2013) An updated definition of stroke for the 21st century: A statement for healthcare professionals from the American Heart Association/American Stroke Association. *Stroke* 44: 2064–2089. https://doi.org/10.1161/STR.0b013e318296aeca
5. GBD 2016 Stroke Collaborators，(2019) Global, regional, and national burden of stroke, 1990–2016: A systematic analysis for the Global Burden of Disease Study 2016. *Lancet Neurol* 18: 439–458. https://doi.org/10.1016/S1474-4422(19)30034-1
6. Feigin VL, Brainin M, Norrving B, Martins S, Sacco RL, Hacke W, et al. (2022) World Stroke Organization (WSO): Global stroke fact sheet 2022. *Int J Stroke* 17: 18–29. https://doi.org/10.1177/17474930211065917
7. Zuo W, Yang X, (2024) A machine learning model predicts stroke associated with blood cadmium levels. *Sci Rep* 14: 14739. https://doi.org/10.1038/s41598-024-65633-w

8. Rehman K, Fatima F, Waheed I, Akash MSH, Prevalence of exposure of heavy metals and their impact on health consequences. *J Cell Biochem* 119: 157–184. https://doi.org/10.1002/jcb.26234
9. Jarup L, (2003) Hazards of heavy metal contamination. *Br Med Bull* 68: 167–182. https://doi.org/10.1093/bmb/ldg032
10. M Jaishankar, T Tseten, N Anbalagan, Mathew BB, Beeregowda KN, (2014) Toxicity, mechanism and health effects of some heavy metals. *Interdiscip Toxicol* 7: 60–72. https://doi.org/10.2478/intox-2014-0009
11. Amini L, Azarpazhouh R, Farzadfar MT, Mousavi SA, Jazaieri F, Khorvash F, et al. Prediction and control of stroke by data mining. *Int J Prev Med* 4: S245.
12. Sung SF, Hsieh CY, Yang YHK, Lin HJ, Chen CH, Chen YW, et al. (2015) Developing a stroke severity index based on administrative data was feasible using data mining techniques. *J Clin Epidemiol* 68: 1292–1300. https://doi.org/10.1016/j.jclinepi.2015.01.009
13. Adam SY, Yousif A, Bashir MB, (2016) Classification of ischemic stroke using machine learning algorithms. *Int J Comput Appl* 149: 26–31. https://doi.org/10.5120/ijca2016911607
14. Sailasya G, Kumari GLA, (2021) Analyzing the performance of stroke prediction using ML classification algorithms. *Int J Adv Comput Sci Appl* 12. https://doi.org/10.14569/IJACSA.2021.0120662
15. Dev S, Wang H, Nwosu CS, Jain N, Veeravalli B, John D, (2022) A predictive analytics approach for stroke prediction using machine learning and neural networks. *Healthcare Anal* 2: 100032. https://doi.org/10.1016/j.health.2022.100032
16. Cao Z, Bakulski KM, Paulson HL, Wang X, (2023) Exposure to heavy metals, obesity, and stroke mortality in the United States, preprint, MedRxiv, 2023 Sep 18:2023.09.18.23295722. https://doi.org/10.1101/2023.09.18.23295722
17. Menke A, Muntner P, Batuman V, Silbergeld EK, Guallar E, (2006) Blood lead below 0.48 μmol/L (10 μg/dL) and mortality among US adults. *Circulation* 114: 1388–1394. https://doi.org/10.1161/CIRCULATIONAHA.106.628321
18. Meishuo O, Eshak ES, Muraki I, Cui R, Shirai K, Iso H, et al. (2022) Association between dietary manganese intake and mortality from cardiovascular disease in Japanese population: The Japan collaborative cohort study. *J Atheroscler Thromb* 29: 1432–1447. https://doi.org/10.5551/jat.63195
19. Curtin LR, Mohadjer LK, Dohrmann SM, Montaquila JM, Kruszan-Moran D, Mirel LB, et al. The national health and nutrition examination survey: Sample design, 1999–2006. *Vital Health Stat Ser 2 Data Eval Methods Res* 155: 1–39.
20. Johnson CL, Dohrmann SM, Burt VL, Mohadjer LK, (2014) National health and nutrition examination survey: Sample design, 2011–2014. *Vital Health Stat Ser 2 Data Eval Methods Res* 2014: 1–33.
21. Chen TC, Clark J, Riddles MK, Mohadjer LK, Fakhouri THI, (2020) National Health and Nutrition Examination Survey, 2015–2018: Sample design and estimation procedures. *Vital Health Stat Ser 2 Data Eval Methods Res* 2020: 1–35.
22. Bengio Y, Courville A, Vincent P, (2013) Representation learning: A review and new perspectives. *IEEE Trans Pattern Anal Mach Intell* 35: 1798–1828. https://doi.org/10.1109/TPAMI.2013.50
23. Hong S, Lynn HS, (2020) Accuracy of random-forest-based imputation of missing data in the presence of non-normality, nonlinearity, and interaction. *BMC Med Res Methodol* 20: 1–12. https://doi.org/10.1186/s12874-020-01080-1

24. Yu L, Liu H, (2004) Efficient feature selection via analysis of relevance and redundancy. *J Mach Learn Res* 5: 1205–1224.
25. Oh IS, Lee JS, Moon BR, (2004) Hybrid genetic algorithms for feature selection. *IEEE Trans Pattern Anal Mach Intell* 26: 1424–1437. https://doi.org/10.1109/TPAMI.2004.105

Appendix

Table A.1. Performance metrics for different models.

Label	Category name	Description
0		Not an actual annotation
1	N	Normal beat
2	L	Left bundle branch block beat
3	R	Right bundle branch block beat
4	a	Aberrated atrial premature beat
5	V	Premature ventricular contraction
6	F	Fusion of ventricular and normal beat
7	J	Nodal (junctional) premature beat
8	A	Atrial premature contraction
9	S	Premature or ectopic supraventricular beat
10	E	Ventricular escape beat
11	j	Nodal (junctional) escape beat
12	/	Paced beat
13	Q	Unclassifiable beat
14	~	Signal quality change
16	—	Isolated QRSlike artifact
18	S	ST change
19	T	Twave change
20	*	Systole
21	D	Diastole
22	”	Comment annotation
23	=	Measurement annotation
24	p	Pwave peak
25	B	Left or right bundle branch block
26	^	Nonconducted pacer spike
27	t	Twave peak
28	+	Rhythm change
29	u	Uwave peak
30	?	Learning
31	!	Ventricular flutter wave
32	[]	End of ventricular flutter/fibrillation
34	e	Atrial escape beat
35	n	Supraventricular escape beat
36	@	Link to external data (aux note contains URL)

Continued on next page

Label	Category name	Description
37	x	Nonconducted Pwave (blocked APB)
38	f	Fusion of paced and normal beat
39	(	Waveform onset PQ junction (begin of QRS)
40	)	Waveform end IPT junction (J point, end of QRS)