BMC Medical Informatics and Decision Making

RESEARCH

Open Access

Hematoma expansion prediction based on SMOTE and XGBoost algorithm

Yan Li[1], Chaonan Du[2], Sikai Ge[1], Ruonan Zhang[1], Yiming Shao[1], Keyu Chen[1], Zhepeng Li[1] and Fei Ma[1*]

Abstract

Hematoma expansion (HE) is a high risky symptom with high rate of occurrence for patients who have undergone spontaneous intracerebral hemorrhage (ICH) after a major accident or illness. Correct prediction of the occurrence of HE in advance is critical to help the doctors to determine the next step medical treatment. Most existing studies focus only on the occurrence of HE within 6 h after the occurrence of ICH, while in reality a considerable number of patients have HE after the first 6 h but within 24 h. In this study, based on the medical doctors recommendation, we focus on prediction of the occurrence of HE within 24 h, as well as the occurrence of HE every 6 h within 24 h. Based on the demographics and computer tomography (CT) image extraction information, we used the XGBoost method to predict the occurrence of HE within 24 h. In this study, to solve the issue of highly imbalanced data set, which is a frequent case in medical data analysis, we used the SMOTE algorithm for data augmentation. To evaluate our method, we used a data set consisting of 582 patients records, and compared the results of proposed method as well as few machine learning methods. Our experiments show that XGBoost achieved the best prediction performance on the balanced dataset processed by the SMOTE algorithm with an accuracy of 0.82 and F1-score of 0.82. Moreover, our proposed method predicts the occurrence of HE within 6, 12, 18 and 24 h at the accuracy of 0.89, 0.82, 0.87 and 0.94, indicating that the HE occurrence within 24 h can be predicted accurately by the proposed method.

Keywords Hematoma expansion, XGBoost, SMOTE, Machine learning prediction, Unbalanced dataset

Introduction

Spontaneous intracerebral hemorrhage (ICH) is defined as sudden bleeding from the brain parenchyma that may extend to the ventricles or subarachnoid space [1]. It has been recognized as an important health issue, contributing to 7.1 million cases and 3.1 million deaths in 2021 [2]. Besides the high mortality rate, ICH may lead to various disabilities, such as epilepsy, psychosis, mood disorders, hemiplegia, and more than 60% of patients can't regain functional independence [3]. In China, there are 0.6–0.8% of people suffering from ICH every year, and the mortality in the acute stage is between 30 and 40% [3]. Moreover, it is believed that the gravity of the situation will worsen in the future [4] due to the growth of aging population in China. Although medical technology has achieved significant advancements, an effective and safe treatment for ICH has not been performed. Hematoma expansion (HE) is one of the common and significant phenomena of ICH, which is closely associated with the deterioration of early neurological function, and is also an independent predictive factor for poor prognosis and increased mortality.

*Correspondence:
Fei Ma
Fei.Ma@xjtlu.edu.cn
[1]Department of Mathematics and Physics, Xi'an Jiaotong-Liverpool University, Suzhou, China
[2]Department of Neurosurgery, Affiliated Jinling Hospital, Medical School of Nanjing University, Nanjing, China

Sato et al. [5] believe that HE has important prognostic value for adverse outcomes and mortality in patients. In clinical practice, it is crucial to predict whether there is expansion of a cerebral hematoma, and then select the appropriate clinical treatment plan. The hematoma volume is recognized as a focus for the 30-day mortality, and preventing the hematoma from expanding is essential for ICH treatments, such as INTERACT, ATACHII and STOP-AUST [1]. In addition, it is known that early treatments on preventing HE can decrease the death rate of ICH. For example, in 2009, Anderson et al. found that early intensive blood pressure-lowing treatment can limit hematoma growth over 72 h in ICH [6]. However, the optimal time for treatment may slip away because the diagnosis of coagulopathy, such as Prothrombin Time (PT), International Normalized Ratio (INR), Activated Partial Thromboplastin Time (APTT), requires 1 to 2 h to be confirmed after blood sample collection and therefore, it is necessary to develop efficient methods to predict HE.

It was considered that HE tends to occur within 6 h after ICH [7], but in clinical, a considerable number of patients experience HE within the first 24 h after onset. Non-contrast computer tomography (CT) is the preferred initial examination in emergency after onset due to its speed and convenience. Several studies have found that some CT image markers, such as shape and heterogeneity of hematoma, island sign, satellite sign, blend sign, black hole sign and swirl sign, have a significant impact on the prediction of HE within 6 h after onset of ICH [8–10]. These CT image markers were generally considered to be associated with HE within 6 h. The value of these markers in the first 24 h after onset is still uncertain. Therefore, it is necessary to explore whether these markers remain predictive within 24 h of ICH onset.

In recent years, the volume of medical data has exploded, and the development of artificial intelligence has made it possible to analyze and interpret different types of medical information [11]. The use of large models to analyze medical images and electronic medical records to assist doctors in diagnosis and decision-making is becoming increasingly popular. In the healthcare domain, machine learning (ML) algorithms play a major role in accurately classifying and predicting various diseases. With ML, experts can analyze and evaluate datasets containing diagnostic information, electronic medical records, and image information to help them de- velop effective treatment strategies [12, 13]. In this study, we aim to predict HE occurrence within 24 h, establish an effective ML-based prediction method, and verify if the factors used for predicting HE occurrence within 6 h are still applicable for predicting HE occurrence within 24 h. In the following, we will list the main contributions of our work:

- We use the Extreme Gradient Boosting (XGBoost) algorithm to predict the HE occurrence within 24 h, as well within 6, 12, 18 and 24 h.
- We use the Synthetic Minority Oversampling Technique (SMOTE) to process and to cope with the imbalanced dataset.
- We identify the key indicators that contribute to the occurrence of HE based on SHapley Additive exPlanations (SHAP) values.
- Through comparison of results with few state of the art methods, including Support Vector Machine (SVM) [1], Random Forest (RF) [2], Logistic Regression (LR) [14] and k-nearest neighbors (KNN) [15], XGBoost showed better predictive performance, verified that XGBoost can be used for HE occurrence prediction.
- We prove that HE can be accurately predicted within 24 h based on indica-tors.

The rest of the paper is designed as follows: In Sect. 2, we present a literature review of the different methods and results of HE prediction. Section 3 includes dataset description and principle of the methodology. Section 4 and Sect. 5 provide results and discussions respectively, while the last section gives the conclusion.

Related work

Extensive prognostic scoring systems have been proposed in multiple literature for prediction of HE in ICH, such as A 9-point prediction score which selects warfarin medication history, point signs, time from symptom onset to first head CT, and baseline hematoma volume as evaluation criteria [16], 24-point BRAIN score [17], Hematoma Expansion Prediction (HEP) score [18], 7-point prediction score choosing baseline hematoma volume, mixed sign, island sign, whirlpool sign, anticoagulant therapy, ICH, time from symptom onset to first head CT, and baseline hematoma volume as evaluation indicators [19], HEAVN score [20], NAG scale [21]. In terms of indicator selection, Nawabi [22] employed Cohen's kappa coefficient for confirming the reliability of CT features on CTA in patients with ICH.

The logistic regression model, as a basic model for predicting dichotomies in statistics, has been widely used in the medical field especially for predicting HE in ICH. Chan et al. [14] used univariate feature selection methods for feature selection, Fisher's exact test and the Kruskal-Wallis test for each feature to determine the optimal subset of features and multivariate logistic regression to establish an automatic prediction model for HE. In studying 118 patients with ICH, Sakuta et al. [21] utilized univariate feature selection methods such as cardinality test, Fisher exact test, T-test and Mann-Whitney U-test for feature selection. After determining the optimal subset

of features, a prediction model was developed using multivariate logistic regression and a scale was created. Besides, Yang et al. [23] performed univariate and binary logistic regression analysis, screened out independent pre- dictors significantly related to HE, and established a new SICH-HE model. This model offers a theoretical foundation for clinicians to promptly identify high-risk HE patients and validate the early surgical decision-making process.

ML has been widely applied in medicine [24] and especially, SVM, RF, DT, KNN and Adaboost all present good performance in predicting HE using routinely available variables [25]. Liu et al. [1] established a SVM model to predict HE, but in comparison to SVM, the model based on the RF algorithm demonstrated higher accuracy [2]. Furthermore, A multi-task deep learning approach that allows simultaneous tumor segmentation and response prediction has two Siamese sub-networks joined at multiple layers, which enables integration of multi-scale feature representations and in-depth comparison of pretreatment and post-treatment images [26]. Ma et al. [27] compared the prediction effects of ResNet-18, ResNet-34 and VGG- 16 neural networks. ResNet-34 achieves the most robust generalization capability in HE prediction and is superior to other mainstream models, which will facilitate accurate, efficient, and automated HE prediction. It addresses the limitations of neural networks in predicting HE through quantitative volume and texture analysis (CTTA) of CT images [28]. A fuzzy C-means (FCM) intelligent segmentation algorithm was established by Xu et al. [29] for intelligent segmentation of patients' brain CT images, which holds high clinical value for the early prediction of HE in patients with ICH.

XGBoost, a gradient boosting learning model, has been widely used for analyzing medical data for classification and prediction in healthcare. It has achieved accurate prediction in hypertension outcomes [30], diabetes [31], cardiovascular [32] and coronary heart diseases [33]. Compared with deep learning models, the biggest advantage is that it has faster speed and stronger robustness when processing large-scale datasets [34]. However, the literature on the prediction of HE by XGBoost is scarce. One possible explanation is that the prediction of HE is usually typically reliant CT images. Unless the corresponding indicators are extracted from the images, deep learning models are often more applicable for image data based prediction. Once the image information is extracted and converted into tabular data, XGBoost could also achieve excellent prediction results in HE prediction [35].

SMOTE, a Synthetic Minority Oversampling Technique, is an enhanced method based on the random oversampling, which simply duplicates samples to increase minority samples. SMOTE addresses the issue of model overfitting, this occurs when the model learns overly specific information that lacks generalization [36]. In clinical data analysis, there is often a bias in the data obtained, which means that the ratio of positive data to negative data is not balanced [37]. Therefore, SMOTE has a wide range of applications in the medical field. For instance, Alghamdi et al. [38] used SMOTE to address the negative impact of imbalanced categories in the constructed model when they carried out the project of predicting diabetes. Pandey and Janghel [39] used SMOTE technique to address the issue of class imbalance in the MIT-BIH database for their study on arrhythmia detection. Besides, Wang et al. [40], Francis, Prasad and Zahoor-Ul-Huq [41] and Xu et al. [42] all used SMOTE to solve the problem of uneven data distribution and compared it with other methods, further showcasing the viability and effectiveness of this approach in medical applications.

Materials and methods

Population

In this study, we investigated a database of brain hemorrhage cases collected by the emergency department and neurosurgery department of a local hospital in Xuzhou, China. A total of 892 patients diagnosed with ICH from 2014 to 2019 were extracted from the database. All personal information about the patients was erased.

The studied population should met revised diagnostic criteria raised in the 4th National Conference on Cerebrovascular Disease: (1) ICH diagnosed by CT images; (2) Previous history of hypertension; (3) Age no less than 18 years old; (4) First cranial CT within 6 h of onset, and follow-up cranial CT within 24 h of first cranial CT. After applying these criteria, a total of 582 records were retained and used in the study.

Data extraction

The patient's cranial imaging findings were interpreted by the imaging physician to determine the site of the cerebral hemorrhage, the volume of the hematoma, whether the hematoma was regular, and whether the hematoma had broken into the ventricles. If the absolute value of the hematoma volume increased by 6 ml or the percentage of increase in the hematoma volume was above 33% between the first and second CT examinations, the hematoma was considered enlarged. The hematoma was divided into two groups: the HE group (case group) and the non-HE group (control group).

The demographics and CT image extraction information we used as indicators in this study included age, gender, diabetes mellitus, alcohol use, ICH position, adimission GCS score, baseline hematoma volume and hematoma expansion, as well as admission SBP, and admission DBP, left or right, hematoma shape score,

Table 1 List of abbreviations

Abbreviations	Full Name
SBP	Systolic blood pressure
DBP	Diastolic blood pressure
IVH	Intraventricular hemorrhage
SAH	Subarachnoid haemorrhage
MLS	Medical laboratory science
GCS	Glasgow coma scale
Left/Right	Site of ICH

Table 2 Variable description

Characteristics	Case group	Control group	*P* value
Age	59.28±13.51	60.12±12.41	0.03
Gender (Male)	53(46.5%)	221(49.5%)	0.58
Diabetes mellitus	9(7.9%)	64(15.1%)	0.19
Alcohol use	12(10.5%)	43(9.4%)	0.19
Admission SBP	173.63±20.57	168.63±19.24	0.02
Admission DBP	101.25±14.39	98.56±12.31	0.04
ICH position	79(46.8%)	196(41.8%)	0.56
Left/Right	62(47.7%)	182(38.9%)	0.45
Shape score	2.81±1.63	2.80±1.61	0.34
Heterogeneity	3.72±1.68	3.64±1.59	0.43
Island sign	45(39.5%)	112(23.9%)	0.28
Satellite sign	10(8.8%)	45(9.6%)	0.24
Black hole sign	15(13.2%)	47(15.2%)	0.17
Blend sign	60(52.6%)	192(46.8%)	0.12
Swirl sign	13(11.4%)	52(12.8%)	0.61
IVH	39(34.2%)	107(39.7%)	0.54
SAH	7(6.1%)	22(7.2%)	0.77
MLS	34(29.8%)	102(28.8%)	0.89
Admission GCS score	10.49±2.99	10.27±2.21	0.19
Hematoma volume	15.73±15.26	14.78±12.37	0.01

The first column corresponds to the name of the indicators used, and the second and third columns correspond to the statistics of each indicator in the case group and the control group, respectively. The statistic of numerical indicators such as age is expressed by means variance, and the statistic of binary indicators such as alcohol use is expressed by number and ratio. P values in the last column are used to determine whether various indicators are significant

hematoma heterogeneity, island sign, satellite sign, black hole sign, blend sign, swirl sign, IVH, SAH and MLS. Among those indicators, hematoma expansion is our target predictive variable. Table 1 lists and describes these indicators.

Statistics

Of these 582 patients studied, 114 (19.6%) of them were diagnosed with an HE. The composition for the different variables are shown in Table 2. Univariate analysis shows the differences between the two groups are statistically significant in terms of admission SBP, admission DBP admission GCS score and baseline hematoma volume ($p<0.05$). While the other indicators are not statistically significant ($p>0.05$).

Data pre-processing

Classification variables, such as gender and various signs, were transformed into binary variables by label encoding. In terms of gender, male was encoded as 1 and female was encoded as 0. In terms of various signs, those that appear were encoded as 1, while those that did not appear are coded as 0. A special variable is the ICH position. Given that almost 50% of ICH occurs in the basal ganglia, we group the basal ganglia into one category, and uniformly assign the remaining positions to another category for coding. After the feature standardization is complete, we fill in the missing values with the average value of the column. This step is specific to machine learning models except XGBoost, which has its own processing for missing values. For all continuous variables, we kept the values and applied Z-score normalization to the data.

Implementation process

After obtaining the preprocessed data, we first randomly split the dataset into a training set and a test set in the ratio of 8:2. We then applied a variety of machine learning algorithms to train on the training set and validate on the test set, and compared the accuracy, precision, recall and F1-score obtained by different algorithms. Meanwhile, the parameters of the algorithms were determined by the Grid Search in Sklearn. Afterward, we used SMOTE to augment the data for the case group, ensuring that its sample size was the same as the control group. The same process was used to train the augmented data, and the predicted performance was combined and discussed with the results previously obtained. After training the models, we analyzed the indicator importance using SHAP value.

In addition, we divided the augmented data into four groups according to the time interval between the first and subsequent CT examinations. We demonstrated that HE can be accurately predicted within 24 h based on indicators by com- paring the predictive performance among the four groups. The complete steps of our proposed model are presented in Fig. 1.

XGBoost prediction

Extreme Gradient Boosting (XGBoost) is a reliable and open-source gradient tree boosting model. It started as a research project by Tianqi Chen in 2014 [43]. As a supervised learning algorithm, it combines an ensemble of estimates from a set of trees. Compared to traditional gradient boosting decision trees, XGBoost has the advantage of column sampling and can also continue tree construction with missing values by transforming the missing values into a sparse matrix, which can effectively helps avoid some overfitting problems.

Given a dataset of form:

$\mathcal{D} = \{(x_i, y_i) : i = 1...n, x_i \in \mathbf{R}^m, y_i \in \mathbf{R}\}$,

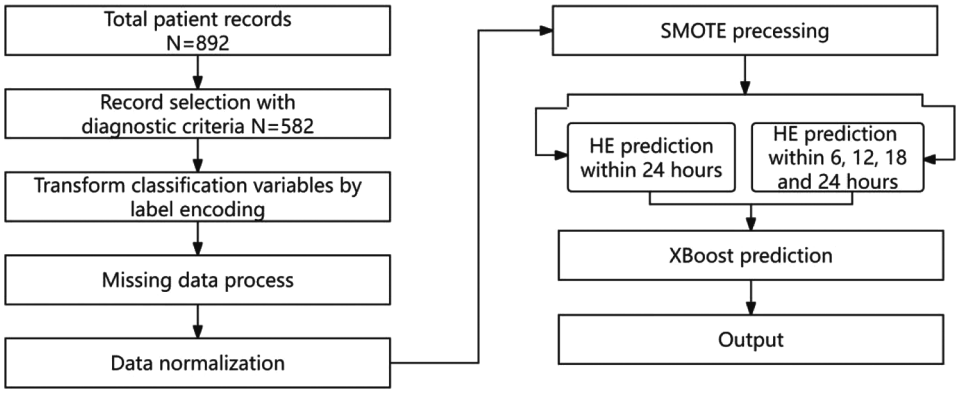

Fig. 1 Steps of the proposed model

Table 3 Samples before and after using SMOTE

Class	1–6 h		7–12 h		13–18 h		19–24 h	
	Before	After	Before	After	Before	After	Before	After
Class 0 Non Hematoma Expansion	117	117	103	103	117	117	155	155
Class 1 Hematoma Expansion	29	117	27	103	28	117	28	155

we get n observations with m features each and with a corresponding variable y. Let $\hat{y}_i$ be defined as a result given by an ensemble represented by the generalised model as follows:

$$\hat{Y}_i = \sum_{k=1}^{K} f_k(x_i) \tag{1}$$

In the above formula, f_k is a regression tree, and $f_k(x_i)$ represents the score given by the k-th to the i-th observation in data.

Then the objective function to be minimized in step t is expressed as:

$$L^{(t)} = \sum_{i=1}^{n} l(y_i,\ \hat{y}_i^{(t-1)} + f_t(x_i)) + \omega(f_t) \tag{2}$$

where $\hat{y}_i^{(t-1)}$ denotes the prediction result of the previous $t-1$ trees for sample x_i, f_t stands for the t tree, l is loss function and ω is the canonical term used for the t-th tree [44].

Smote

An imbalanced dataset is one in which the number of examples in one class is significantly different from the number of examples in other classes. To deal with the over-fitting problems that often occur when facing an imbalanced datasets, a Synthetic Minority Oversampling Technique called "SMOTE" was proposed by Chawla et al. [36]. The fundamental concept of this method is to generate new samples for the minority class in the data set by means of a linear interpolation algorithm. Compared with random over-sampling techniques, this algorithm can increase the variety of training samples instead of repeating the original training samples, thus effectively solving the over-fitting problem. The steps for this technique are described as follows [45]:

a) For each sample point x_i in the minority class set A, calculate its Euclidean distance with every other points in set A, and obtain the k-nearest neighbours of x_i.

b) For k-nearest neighbours of x_i, arbitrarily choose the appropriate number of samples N (i.e. x_1,...,x_N) to form a new sample set A_1. Here the sampling multiplier N is based on the proportion of sample imbalance.

c) For every $x_j \in A_1$ $(j = 1, 2, ..., N)$, A new sample point x_i is synthesized by the following linear interpolation formula:

$$x_i' = x_i + rand(0,1) \cdot |x_i - x_j| \tag{3}$$

d) The newly generated minority samples are combined with the original sample A to form a new sample A′ for training.

In this study, the number of patients with HE was much lower than the number of patients without HE. The imbalance in the data samples significantly affects the performance of the prediction model, therefore, before constructing the prediction model [46], we propose balancing the dataset using the SMOTE algorithm, which can generate new minority class samples to achieve a balanced data sample between classes. The distribution of the data before and after SMOTE can be seen in Table 3.

SHAP values

One of the major problems with machine learning models is that the models themselves are not interpretable, and SHAP (SHapley Additive exPlanations) is one approach to tackle this problem. SHAP is based on the Shapley value, a game-theoretic concept introduced by economist Lloyd Shapley, which is interpreted by SHAP as an additive feature attribution method that determines the importance of an individual by calculating the contribution of that individual in cooperation. The model explains the predicted values through a linear function of binary variables [47]:

$$g(z) = \varphi_0 + \sum_{i=1}^{M} \varphi_{z_i} \tag{4}$$

Here φ_0 stands for the the typical prediction, M is the number of features of the simplified input and the SHAP value φ_{z_i} represents its direct effect on the model prediction.

We calculate the SHAP value for each of the covariates on the test set. After that, a summary plot was drawn to present the SHAP values of each feature, and by colour we can see the relationship between the size of the feature and the predicted impact, as well as showing its eigenvalue distribution. Meanwhile, the dependence plot clearly shows how individual features affect the prediction results of the model.

Results

Predicted performance

Overall, the five models produced better results in terms of precision, recall and F1- score with the balanced datasets (see Table 4). However, the accuracy of the models with the balanced datasets was lower than with the imbalanced dataset, which indicates that the classification of HE was towards the majority samples of non-HE. These results clearly show that designing models using imbalanced datasets will lead to significant inaccuracies, which cannot identify HE and non-HE precisely and this verifies the necessity of using a balancing algorithm to balance datasets in the first step of the classification process. In contrast, the F1-score is more convincing when evaluating a model's predictive performance on unbalanced data.

Table 4 Predictive performance of different models on balanced and imbalanced datasets

Dataset	Methods	Accuracy	Precision	Recall	F1-score
IBT	XGBoost	0.90	0.51	0.62	0.60
IBT	SVM [1]	0.87	0.44	0.50	0.47
IBT	RF [2]	0.85	0.43	0.46	0.49
IBT	LR [14]	0.86	0.44	0.50	0.46
IBT	KNN [15]	0.87	0.44	0.51	0.48
BT	XGBoost	0.82	0.82	0.82	0.82
BT	SVM [1]	0.78	0.81	0.77	0.77
BT	RF [2]	0.80	0.79	0.77	0.80
BT	LR [14]	0.69	0.69	0.69	0.69
BT	KNN [15]	0.80	0.78	0.77	0.78

In addition, the ensemble models outperformed the single classifiers, as determined by the performance indicators, among which, the Area Under Curve and precision values of XGBoost with SMOTE exceeded those of SVM, RF and LR with SMOTE algorithm. Particularly, XGBoost with SMOTE produced the highest results among all classifier models with an accuracy of 0.82 and a F1-score of 0.82 on a balanced dataset. Especially, the F1-score value indicates that the XGBoost model can distinguish between HE and non-HE precisely.

The first five rows correspond to the predictions of the five machine learning algorithms on the imbalanced dataset, and the last five rows correspond to their respective predictions on the balanced dataset.

Feature importance

Figure 2 shows the list of the top 10 features among all that are used in the XGBoost model, following the order of contribution for each evaluation metric. Of all, the most influential covariate for predicting whether a hematoma will enlarge is the initial hematoma volume. Admission SBP, age and admission DBP also play an important role in forecasting.

SHAP value analysis

We analyzed the relative effect of the top 10 features on the model at each data point in the test set according to the mean absolute SHAP value (Fig. 3). The summary plot was applied to identify influential covariates. Each point in the summary plot indicated the Shapley value and observation value for the characteristic, with the color indicating the value of the characteristic. According to these results, baseline hematoma volume, admission DBP, age, admission SBP and GCS carried model's forecasting power.

To further investigate the impact of each variable, we analyzed the SHAP values of the selected 4 important covariates separately in Fig. 4. The admission DBP, except for the range of 85 to 100, positively contributes to the HE (Fig. 4(a)). For age, we saw a relationship with HE before and after 65 (Fig. 4(b)). In terms of admission SBP, the relationship with HE is "W" shaped and the peak of SHAP value occurring when the admission SBP is 180 (Fig. 4(c)). As for the baseline hematoma volume, the overall trend is relatively stable, with the lowest point occurring when the baseline hematoma volume equals 18 (Fig. 4(d)).

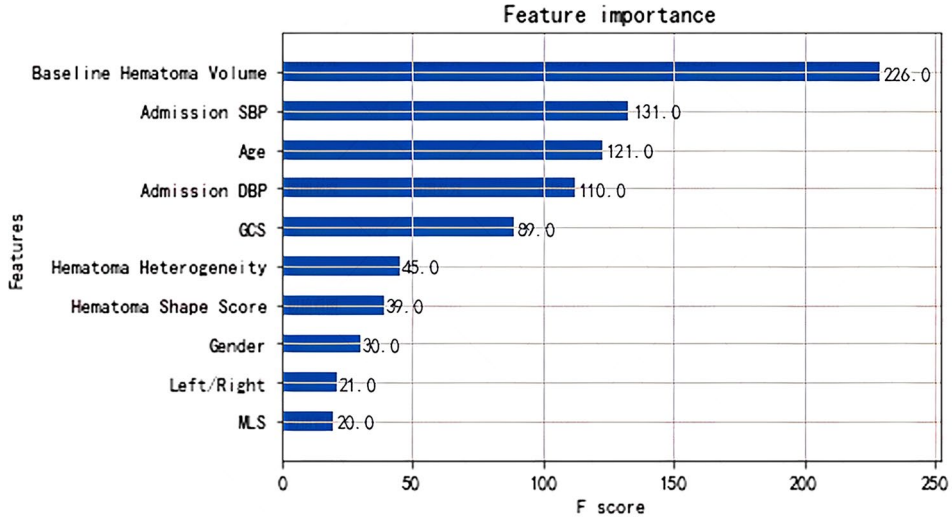

Fig. 2 Contributions of features from XGBoost for the whole dataset

Predicted performance of different time groups

In previous studies [8, 48–50], CT imaging markers such as shape and heterogeneity of hematoma, island sign, satellite sign, blend sign, black hole sign, and swirl sign had good effects on the prediction of HE within 6 h after the onset of ICH. Studies have found that more than one third of ICH patients who underwent CT scanning within a few hours after onset had HE [43]. But there were still a number of patients who could have HE within the first 24 h after 6 h. Thus, we further investigated the predictive ability of these markers to HE within 24 h after ICH. We have evenly divided 24 h into four parts, that is, $T_1 = (0, 6]$, $T_2 = (6, 12]$, $T_3 = (12, 18]$ and $T_4 = (18, 24]$.

Table 5 summarizes the performances of the algorithms in particular time periods T_1, T_2, T_3 and T_4 and it is obvious that the disaggregated results are better than that presented in Table 4. At the same time, XGBoost has also achieved the best prediction performance on different time groups among all machine learning algorithms, whereas LR presented the worst predictive ability in the dataset.

By comparing the predictions between groups, the highest accuracy and F-score was achieved in time group T_4 and the two values were 0.94 and 0.93, respectively. It is worth noting that, except for the first time group, the predictive performances of the models all improve long with time implying that although the area under the curve of ROC curves decreased with time, it still maintained a high accuracy. Therefore, these CT image markers have high predictive power and could be regarded as reliable indicators for predicting HE in the first 24 h after ICH. Besides, our model also validates that HE occurring within 24 h can be predicted with the help of machine learning models.

Discussion

This study has developed classification models to forecast HE based on different machine learning models combining the SMOTE algorithm. By analysing real cases of cerebral hemorrhage in hypertensive patients over the past six years, we have confirmed the feasibility of such hematoma prediction and summarized the main features for prediction results.

Risk factor

Many risk factors for HE have been clinically proven. As did previous studies, we found that elevated SBP is a risk factor. Also, age, initial hematoma volume and DBP differed between expanders and non-expanders. In addition, HE can also be induced by an increase in SBP, but in our study, we found that SBP at admission did not show a linear relationship with HE. A higher SBP did not correspond to a higher probability of HE. As this is a retrospective exercise, the data set itself may be subject to selection bias, and therefore, a prospective double-blind study is required. Furthermore, many studies have illustrated that GCS score is the most important risk factor for determining ICH patients. However, in this study, the GCS was not particularly important for the outcome. This argument is also consistent with Rangaraj's [51] findings.

The current treatment for ICH is mainly in the management approach in reducing SBP value since early antihypertensive treatment can effectively reduce the risk of HE [52, 53]. Studies [54, 55] have shown that intensive blood pressure lowering within 24 h of admission can reduce the risk of HE and thus reduce the risk in patients. The specific criteria for lowering blood pressure are to lower SBP to below 180. As we have discussed regarding Fig. 4(c), the peak of the SHAP value occurs when the

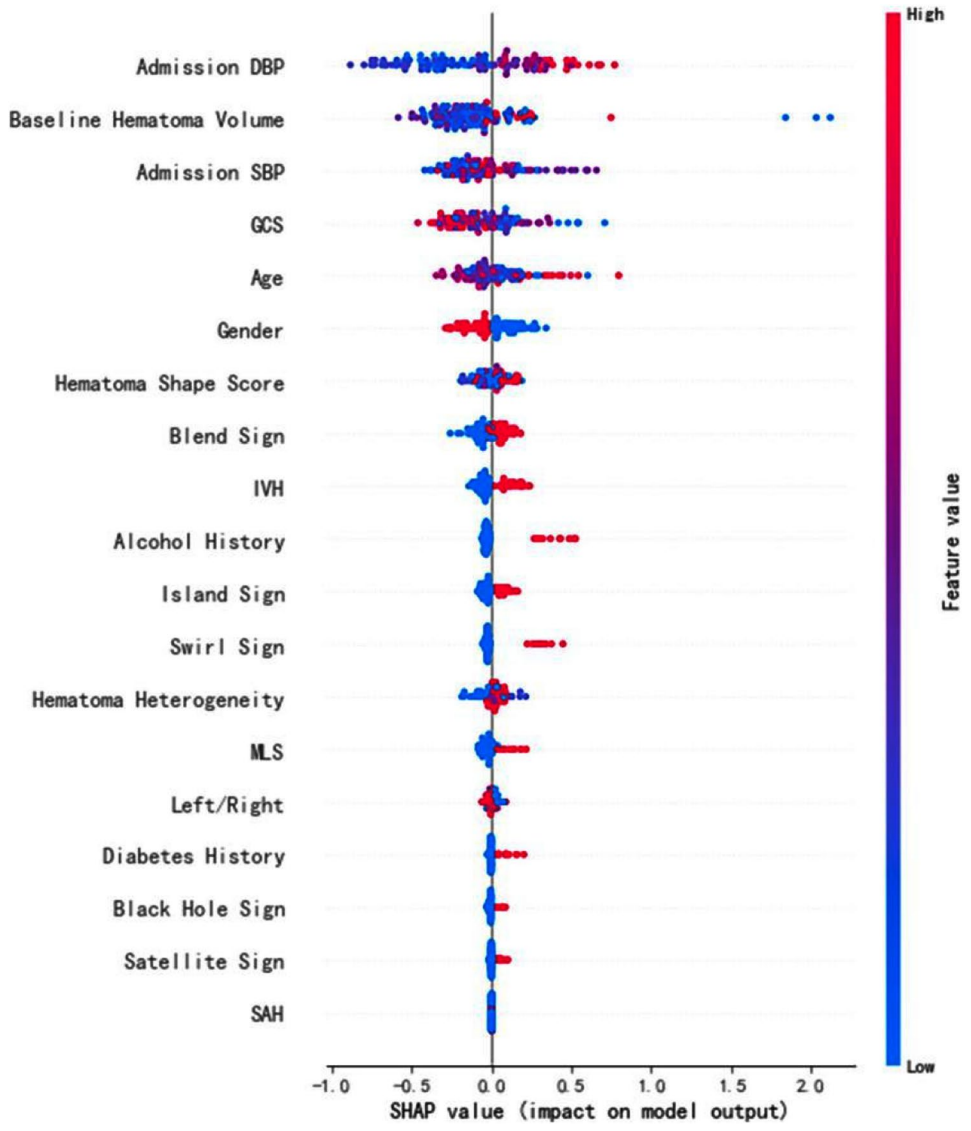

Fig. 3 SHAP value (effect on model output)

admission SBP equals 180. Our findings also validate previous studies.

Imbalanced data sets

As Dong et al. [56], and Liu [57] pointed out, a common issue in current research in the medical field is how to handle imbalanced datasets. The low prevalence of many diseases results in a small proportion of the data set being labeled with this type of disease. Unfortunately, most machine learning algorithms typically make poor predictions for the minority class. Thus, being able to make accurate predictions for these few occurrences of disease is valuable. When facing such a dataset, resampling the training data is a practical method to address the issue of imbalance. Down-sampling majority classes, over-sampling minority classes, or some combinations are commonly applied [44, 47]. For example, the ratio of the sample size of the experimental and control groups was adjusted from 5:1 to 1:1 by the SMOTE algorithm. The original over-accuracy and under-performance of the F1-score were iterated to obtain a more convincing result and provides a reliable basis for clin- ical clinical applications. Another approach is cost-sensitive learning, which reformulates existing learning algorithms by giving more weight to the minority classes [58, 59].

Ensemble boosting learning methods

A comparison among different methods in this study clearly demonstrates that machine learning algorithms can achieve more accurate prediction performance than logistic regression algorithms for such multivariate datasets. In particular, ensemble boosting learning methods

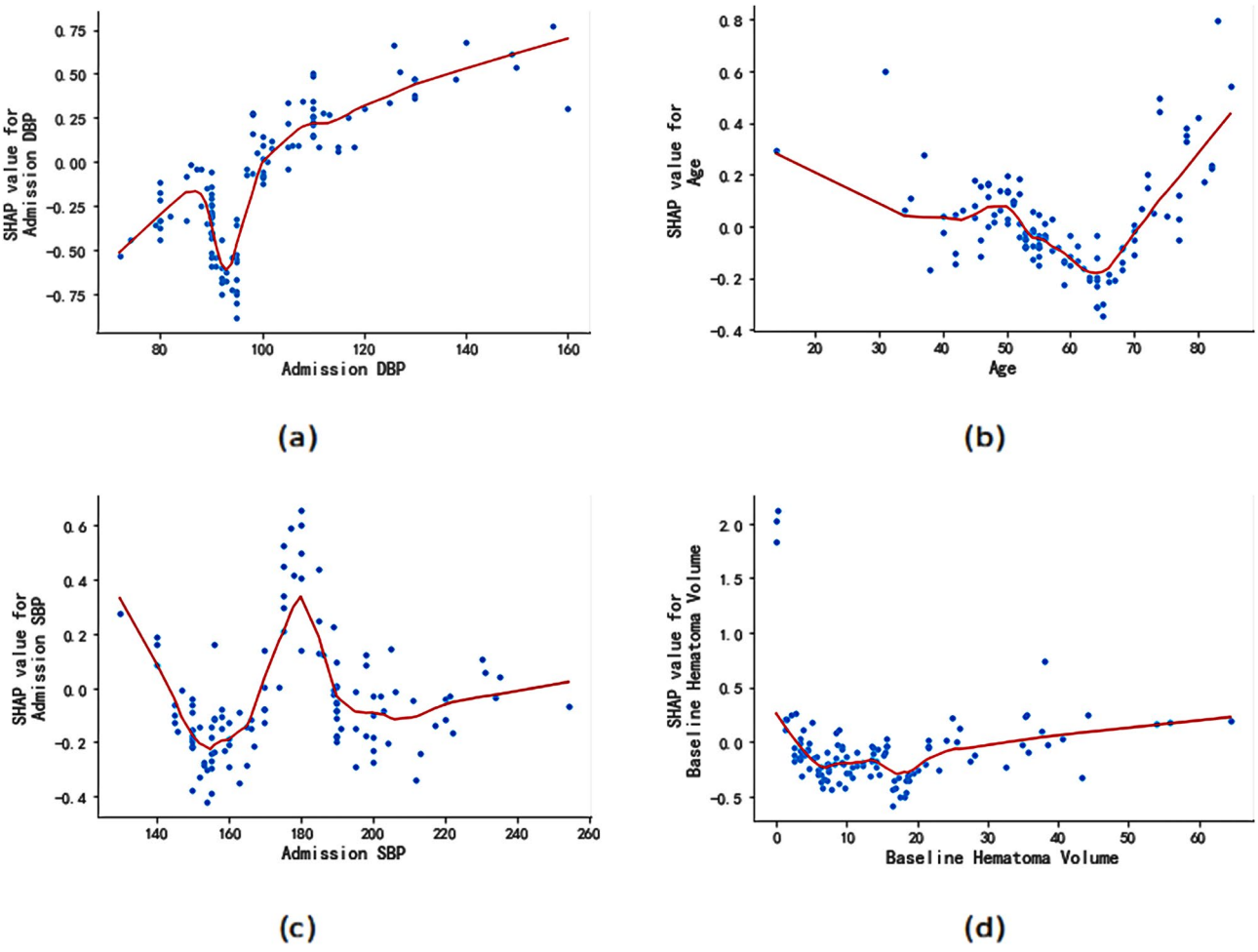

Fig. 4 Four examples of dependence plots showing the effect on the HE with respect to the feature value. Points represents the SHAP values, while lines indicate the LOESS fitted smooth representation of the relationship

such as XGBoost, tend to be favored. More studies are experimenting with various ensemble boosting learning methods to predict ICH patients instead of the traditional extensive prognostic scoring system. Ensemble boosting learning algorithms integrate several weak classifiers to reduce the poten- tial bias of individual model, obtaining significantly superior generalization performance compared to a single learner and avoiding the production of biased and unstable results [60, 61].

With the continuous optimization of deep learning, it has proven to be a powerful predictive tool in the field of healthcare prediction. However, in the classification and prediction of tabular data, XGBoost demonstrates a large advantage over deep learning models in terms of both accuracy and time, as evidenced in many literature [34, 62]. In addition, deep learning models are often challenging to interpret, which can lead to doctors distrusting the predictive performance of these models if their predictions contradict their intuitive judgment. In contrast, the classical ML model can depict the effects of different variables on the prediction results through SHAP value, which provides interpretability of the prediction results of the XGBoost model, making it more acceptable to physicians.

Limitations

First of all, most of our data is derived from CT image findings and lacks clinical testing indicators like routine blood test results. However, these results have also proven to be important in correctly predicting ICH patients [51]. In addition, our training and test sets are unified from a single system and the results are only for this batch of data sets, it would be more ideal to include some external data sets to validate the reliability and robustness of our model.

Conclusion

HE is a high risky symptom happening frequently on patients who have undergone spontaneous ICH. Correct prediction of the occurrence of HE yields great value

Table 5 The predictive results on different time groups

Methods	Accuracy	Precision	Recall	F1-score
(a) Time group T_1				
XGBoost	0.89	0.91	0.90	0.89
SVM [1]	0.77	0.84	0.76	0.75
RF [2]	0.86	0.89	0.89	0.88
LR [14]	0.74	0.78	0.75	0.74
KNN [15]	0.82	0.88	0.88	0.88
(b) Time group T_2				
Methods	Accuracy	Precision	Recall	F1-score
XGBoost	0.82	0.83	0.83	0.80
SVM [1]	0.70	0.70	0.83	0.69
RF [2]	0.80	0.81	0.81	0.78
LR [14]	0.69	0.69	0.67	0.68
KNN [15]	0.78	0.78	0.77	0.78
(c) Time group T_3				
Methods	Accuracy	Precision	Recall	F1-score
XGBoost	0.87	0.87	0.88	0.87
SVM [1]	0.77	0.82	0.74	0.74
RF [2]	0.85	0.85	0.85	0.85
LR [14]	0.74	0.74	0.75	0.74
KNN [15]	0.86	0.86	0.87	0.87
(d) Time group T_4				
Methods	Accuracy	Precision	Recall	F1-score
XGBoost	0.94	0.94	0.93	0.93
SVM [1]	0.79	0.85	0.80	0.78
RF [2]	0.92	0.92	0.92	0.92
LR [14]	0.84	0.84	0.84	0.84
KNN [15]	0.92	0.92	0.92	0.92

towards determination of critical medical treatment. This study developed a prediction model based on XGBoost to forecast the occurrence of HE. In the comparison of the prediction results obtained by our proposed method and few other machine learning methods, our proposed method achieved the best prediction performance with a prediction accuracy of 0.82 on the balanced dataset processed by the SMOTE algorithm. On the predictions of HE occurrence within 6, 12, 18 and 24 h, the accuracy of the predictions with the proposed method all exceeded 0.8. We have confirmed that HE can be accurately predicted within 24 h based on indicators in a retrospective study. Through our study we can conclude that hematoma volume, admission SBP and admission DBP contribute greatly to the occurrence of HE.

It has been presented that machine learning algorithms can effectively integrate diverse medical data to accurately and efficiently predict targets. Future research is directed towards exploring the generalisability of our proposed predictive model and exploring more advanced data generation algorithms. AI techniques, such as generative AI, could be used to create possible training data, without which, some latest AI methods, such as deep learning based methods, cannot be employed due to its need of large training data set. However, accuracy and closeness of the generated data to the real data need to be researched before generative AI generated data can be used for model training.

Acknowledgements
This study was supported in part by XJTLU laboratory for intelligent computation and financial technology through XJTLU Key Programme Special Fund (KSF-E21).

Author contributions
Y.L. provided the experimental work for the analysis of the data and were the major contributors in writing the manuscript. C.D. and S.G. provided data set and data processing work. R.Z. provided guidance on language improvement. Y.S., K.C. and Z.L. participated in the experiment and wrote the Introduction and Related work part. F.M. provided guidance on the experiment.

Funding
Not applicable.

Data availability
The datasets analysed during the current study are not publicly available. Inquiries regarding datasets and codes access can be directed via email to: yan.li14@student.xjtlu.edu.cn.

Declarations

Ethics approval and consent to participate
The experimental protocol was approved by the School Ethics Review Panel of Xi'an Jiaotong-Liverpool University (No.: ER-SMP-12881113720220628122052). All methods were carried out in accordance with relevant guidelines and regulations. Written informed consent and assent from individuals and respective legal guardians was obtained.

Consent for publication
Not applicable.

Competing interests
The authors declare no competing interests.

Received: 11 March 2023 / Accepted: 30 May 2024
Published online: 19 June 2024

References

1. Liu J, Xu H, et al. Prediction of hematoma expansion in spontaneous intracerebral hemorrhage using support vector machine. EBioMedicine. 2019;43:454–9.
2. Zhu F, Pan Z, Tang Y, et al. Machine learning models predict coagulopathy in spontaneous intracerebral hemorrhage patients in er. CNS Neurosci Ther. 2021;27:92–100.
3. Rao M. People 's medical publishing house. In: Guidelines for Prevention and Treatment of Cerebrovascular Diseases in China. (in Chinese), p. 54 (2007).
4. Rao M. People 's medical publishing house. In: Guidelines for Prevention and Treatment of Cerebrovascular Diseases in China. (in Chinese), p. 1 (2007).
5. Sato S, Delcourt C, Zhang S, et al. Determinants and prognostic significance of hematoma sedimentation levels in acute intracerebral hemorrhage. Cerebrovasc Dis. 2015;41(1–2):80.
6. Craig S, et al. Investigators effects of early intensive blood pressure-lowering treatment on the growth of hematoma and perihematomal edema in acute intracerebral hemorrhage. Stroke. 2010;41:307–12.
7. Feigin V. Worldwide stroke incidence and early case fatality reported in 56 population-based studies: a systematic review. Lancet Neurol. 2009;8:355–69.
8. Li Q. Island sign: an imaging predictor for early hematoma expansion and poor outcome in patients with intracerebral hemorrhage. Stroke. 2018;48:3019.
9. Li Q, Zhang G, et al. Black hole sign: novel imaging marker that predicts hematoma growth in patients with intracerebral hemorrhage. Stroke. 2016;471777–1781:1777–81.

10. Selariu E, et al. Swirl sign in intracerebral haemorrhage: definition, prevalence, reliability and prognostic value. BMC Neurol. 2012;12:109.
11. Kumar V, et al. Ensembling classical machine learning and deep learning approaches for morbidity identification from clinical notes. IEEE Access. 2021;9:7107–26.
12. Wu Z et al. Anno-mi: A dataset of expert-annotated counselling dialogues. In: ICASSP 2022–2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Singapore, pp. 6177–6181 (2022).
13. Wu Z, et al. Creation, analysis and evaluation of annomi, a dataset of expert-annotated counselling dialogues. Future Internet. 2023;15(3):110.
14. Chan S, Conell C, et al. Prediction of intracerebral haemorrhage expansion with clinical, laboratory, pharmacologic, and noncontrast radiographic variables. Int J Stroke. 2015;10(7):1057–61.
15. Tang Z, et al. Predicting hematoma expansion in intracerebral hemorrhage from brain ct scans via k-nearest neighbors matting and deep residual network. Biomed Signal Process Control. 2022;76:103656.
16. Brouwers H, Chang Y, Falcone G, et al. Predicting hematoma expansion after primary intracerebral hemorrhage. JAMA Neurol. 2014;71(2):158–64.
17. Wang X, Arima H, et al. Clinical prediction algorithm (brain) to determine risk of hematoma growth in acute intracerebral hemorrhage. Stroke. 2015;46(2):376–81.
18. Yao X, Xu Y, et al. The hep score: a nomogram-derived hematoma expansion prediction scale. Neurocrit Care. 2015;23(2):179–87.
19. Huang Y, Zhang Q, Yang M. A reliable grading system for prediction of hematoma expansion in intracerebral hemorrhage in the basal ganglia. Biosci Trends. 2018;12(2):193–200.
20. Miyahara M, Noda R, Yamaguchi S, et al. New prediction score for hematoma expansion and neurological deterioration after spontaneous intracerebral hemorrhage: a hospital-based retrospective cohort study. J Stroke Cerebrovasc Dis. 2018;27(9):2543–50.
21. Sakuta K, Sato T, et al. The nag scale: Noble predictive scale for hematoma expansion in intracerebral hemorrhage. J Stroke Cerebrovasc Dis. 2018;27(10):2606–12.
22. Nawabi J, Elsayed S, et al. Inter- and intrarater agreement of spot sign and noncontrast ct markers for early intracerebral hemorrhage expansion. J Stroke Cerebrovasc Dis. 2020;9(4):1020.
23. Yang M, Du C, et al. Nomogram model for predicting hematoma expansion in spontaneous intracerebral hemorrhage: Multicenter retrospective study. World Neurosurg. 2020;137:470–8.
24. Rajkomar A, Dean J, Kohane I. Machine learning in medicine. N Engl J Med. 2019;380(14):1347–58.
25. Tang Z, Zhu Y, Lu X, et al. Deep learning-based prediction of hematoma expansion using a single brain computed tomographic slice in patients with spontaneous intracerebral hemorrhages. World Neurosurg. 2022;8750(22):00749–5.
26. Jin C, Yu H, et al. Predicting treatment response from longitudinal images using multi-task deep learning. Nat Commun. 2021;12(1):1851.
27. Ma C, Wang L, Gao C, et al. Automatic and efficient prediction of hematoma expansion in patients with hypertensive intracerebral hemorrhage using deep learning based on ct images. J Pers Med. 2022;12(5):779.
28. Kanazawa T, Takahashi S, et al. Prediction of postoperative recurrence of chronic subdural hematoma using quantitative volumetric analysis in conjunction with computed tomography texture analysis. J Clin Neurosci. 2020;72:270–6.
29. Xu W, Tang W, Wu L et al. Early prediction of cerebral computed tomography under intelligent segmentation algorithm combined with serological indexes for hematoma enlargement after intracerebral hemorrhage. Comput Math Methods Med. 2022, 5863082 (2022).
30. Chang W, et al. A machine-learning based prediction method for hypertension outcomes based on medical data. Diagnostics. 2019;9:178.
31. Hassan M et al. Diabetes prediction in healthcare at early stage using machine learning approach. In: 12th International Conference on Computing Communication and Networking Technologies (ICCCNT), pp. 01–05 (2021).
32. Dinh A, et al. A data-driven approach to predicting diabetes and cardiovascular disease with machine learning. BMC Med Inf Decis Mak. 2019;19:211.
33. Tama B et al. Improving an intelligent detection system for coronary heart disease using a two-tier classifier ensemble. In: In: BioMed Research Inter National [Internet]. Hindawi, p. 9816142 (2020).
34. Dhaliwal S, et al. Effective intrusion detection system using xgboost. Information. 2018;9(7):149.
35. Tanioka S, Yago T, et al. Machine learning prediction of hematoma expansion in acute intracerebral hemorrhage. Sci Rep. 2022;12(1):12452.
36. Chawla N, Bowyer k, et al. Smote: synthetic minority over-sampling technique. J Artif Intell Res (JAIR). 2002;16:321–57.
37. Wang H, Guo X, Jia Z, et al. Multilevel binomial logistic prediction model for malignant pulmonary nodules based on texture features of ct image. Eur J Radiol. 2010;74(1):124–9.
38. Alghamdi M, Al-Mallah M, Keteyian S, et al. Predicting Diabetes mellitus using smote and ensemble machine learning approach: the henry ford exercise testing (fit) project. PLoS ONE. 2017;12(7):0179805.
39. Pandey S, Janghel R. Automatic detection of arrhythmia from imbalanced ecg database using cnn model with smote. Australas Phys Eng Sci Med. 2019;42(4):1129–39.
40. Wang K, Tian J, et al. Improving risk identification of adverse outcomes in chronic heart failure using smote + enn and machine learning. Risk Manag Healthc Policy. 2021;14:2453–63.
41. Francis PPS, adn Prasad, Zahoor-Ul-Huq S. Medical data classification based on smote and recurrent neural network. Int J Eng Adv Technol. 2020;9:2560–5.
42. Xu Z, Shen D, et al. An oversampling algorithm combining smote and k-means for imbalanced medical data. Inf Sci. 2021;572:574–98.
43. Chen T, Guestrin C. Xgboost: A scalable tree boosting system. In: ACM, editor In Proceedings of the 22Nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 785–794 (2016).
44. Pan B. Application of xgboost algorithm in hourly pm2.5 concentration prediction. iop Conf Ser Earth Environ Sci. 2018;113(1):012127.
45. Keller A, Pandey A. Smote and enn based xgboost prediction model for parkinson's disease detection, 2021 2nd international conference on smart electronics and communication. In: 2021 2nd International Conference on Smart Electronics and Communication, pp. 839–846 (2021).
46. Kumar V et al. Data augmentation for reliability and fairness in counselling quality classification. In: In Proceedings of the 1st Workshop on Scarce Data in Artificial Intelligence for Healthcare - SDAIH, pp. 23–28 (2023).
47. Janssen A, Hoogendoorn M, et al. Application of shap values for inferring the optimal functional form of covariates in pharmacokinetic modeling. CPT Pharmacometrics Syst Pharmacol. 2022;11:1100–10.
48. Nawabi J, Elsayed S, Kniep H, et al. Inter-and intrarater agreement of spot sign and noncontrast ct markers for early intracerebral hemorrhage expansion. J Clin Med. 2020;9:1020.
49. Li Q, Zhang G, Xin X, et al. Black hole sign: novel imaging marker that predicts hematoma growth in patients with intracerebral hemorrhage. Stroke. 2016;47:1777–81.
50. Shimoda Y, Ohtomo S, Arai H et al. A poor outcome predictor in intracerebral hemorrhage. Cerebrovasc Dis. (2017).
51. Rangaraj S, Islam M, et al. Identifying risk factors of intracerebral hemorrhage stability using explainable attention model. Med Biol Eng Comput. 2022;60(2):337–48.
52. Anderson C, Heeley E, et al. Rapid blood-pressure lowering in patients with acute intracerebral hemorrhage. N Engl J Med. 2013;368(25):2355–65.
53. Qureshi A. Intensive blood-pressure lowering in patients with acute cerebral hemorrhage. N Engl J Med. 2016;375:1033–43.
54. Rodriguez-Luna D, Rubiera M, et al. Impact of blood pressure changes and course on hematoma growth in acute intracerebral hemorrhage. Eur J Neurol. 2013;20:1277–83.
55. Oh DM, Shkirkova K, et al. Association between hyperacute blood pressure variability and hematoma expansion after intracerebral hemorrhage: secondary analysis of the fast-mag database. Neurocrit Care. (2022).
56. Dong Q, Gong S, Zhu X. Imbalanced deep learning by minority class incremental rectification. ieee Trans Pattern anal Mach Intell. 2019;41(6):1367–81.
57. Liu P, Zheng G. Handling imbalanced data: uncertainty-guided virtual adversarial training with batch nuclear-norm optimization for semi-supervised medical image classification. IEEE J BIOMEDICAL HEALTH Inf. 2022;41(7):2983–94.
58. Zeng H, Yang C, et al. A lightgbm-based eeg analysis method for driver mental states classification. Comput Intell Neurosci. 2019;9:3761203.
59. Wang Y, Wang T. Application of improved lightgbm model in blood glucose prediction. Appl Sci. 2020;10:3227.
60. Pasha A, Anbalagan R, Setlur A, et al. Implementation of ensemble machine learning algorithms on exome datasets for predicting early diagnosis of cancers. BMC Bioinformatics. 2022;23(1):1–24.
61. Kavzoglu T, Teke A. Predictive performances of ensemble machine learning algorithms in landslide susceptibility mapping using random forest, extreme gradient boosting (xgboost) and natural gradient boosting (ngboost). Bus

Media B V). 2022;47(6):7367–85. Arabian Journal for Science & Engineering (Springer Science
62. Shwartz-Ziv R, Armon A. Tabular data: deep learning is not all you need. Inform Fusion. 2022;81:84–90.

Publisher's Note