BMC Medical Informatics and Decision Making

RESEARCH

Open Access

Explainable prediction of daily hospitalizations for cerebrovascular disease using stacked ensemble learning

Xiaoya Lu[1] and Hang Qiu[1,2*]

Abstract

Background With the prevalence of cerebrovascular disease (CD) and the increasing strain on healthcare resources, forecasting the healthcare demands of cerebrovascular patients has significant implications for optimizing medical resources.

Methods In this study, a stacking ensemble model comprised of four base learners (ridge regression, random forest, gradient boosting decision tree, and artificial neural network) and a meta learner (elastic net) was proposed for predicting the daily number of hospital admissions (HAs) for CD using the historical HAs data, air quality data, and meteorological data in Chengdu, China from 2015 to 2018. To solve the label imbalance problem, a re-weighting method based on label distribution smoothing was integrated into the meta learner. We trained the model using the data from 2015 to 2017 and evaluated its predictive ability using the data in 2018 based on four metrics, including mean absolute error (MAE), root mean square error (RMSE), mean absolute percentage error (MAPE), and coefficient of determination (R^2). In addition, the SHapley Additive exPlanations (SHAP) framework was applied to provide explanation for the prediction of our stacking model.

Results Our proposed model outperformed all the base learners and long short-term memory (LSTM) on two datasets. Particularly, compared with the optimal results obtained by individual models, the MAE, RMSE, and MAPE of the stacking model decreased by 13.9%, 12.7%, and 5.8%, respectively, and the R^2 improved by 6.8% on CD dataset. The model explanation demonstrated that environmental features played a role in further improving the model performance and identified that high temperature and high concentrations of gaseous air pollutants might strongly associate with an increased risk of CD.

Conclusions Our stacking model considering environmental exposure is efficient in predicting daily HAs for CD and has practical value in early warning and healthcare resource allocation.

Keywords Stacking ensemble model, Environmental exposure, Hospital admissions, Cerebrovascular disease, SHAP value

*Correspondence:
Hang Qiu
qiuhang@uestc.edu.cn
[1] School of Computer Science and Engineering, University of Electronic Science and Technology of China, No.2006, Xiyuan Ave, West Hi-Tech Zone, 611731 Chengdu, Sichuan, People's Republic of China
[2] Big Data Research Center, University of Electronic Science and Technology of China, Chengdu, China

Background

Cerebrovascular disease (CD) is a leading cause of death and disability worldwide. The World Health Organization has reported that more than 6 million deaths can be attributed to CD each year [1]. In China, about 13 million people suffered from stroke, a subtype of CD [2]. Although hypertension, high-fat diet, smoking, and

alcohol consumption are well-known risk factors for CD, evidence from epidemiological studies indicates that short-term environmental exposure, such as air pollution and extreme weather conditions, also has an important impact on the onset of CD, resulting in an increased risk of morbidity [3–5]. Additionally, toxicological studies have also presented several credible biological mechanistic pathways for the negative health effects associated with air pollution [6–8]. For example, air pollution exposure may provoke platelet activation, leading to enhanced blood coagulation and thrombosis formation [9].

The growing morbidity and high treatment cost of CD have caused a heavy burden on the limited healthcare resources. Forecasting the daily number of hospital admissions (HAs) for CD is of practical significance to optimize medical resources and protect public health by providing an early-warning signal against the impending incidence. Common traditional regression methods for time series prediction, such as the gray model, simple exponential smoothing (SES) model, and autoregressive integrated moving average (ARIMA) model, have been widely applied in predicting healthcare service demand [10–13]. The traditional methods can be easily implemented but have difficulties dealing with multi-factor effects and non-linear mapping, therefore, these studies seldom extract features from factors other than historical demand series.

Machine learning (ML) methods can overcome these disadvantages of traditional regression models [14] and have been applied by a limited number of studies to forecast the demands for healthcare services associated with environmental exposure. For instance, Qiu et al. [15] found that the light gradient boosting machine model outperformed the other five ML models they tested in predicting the peak demand days for cardiovascular disease (CVD) admissions. The researchers also identified that meteorological conditions and air pollutants substantially contributed to prediction accuracy. Bibi et al. [16] used a backpropagation neural network model to predict emergency department visits and found that the model performance was remarkably improved after considering temperature, humidity, and air pollution. Kassomenos et al. [17] discovered that the use of Artificial Neural Network (ANN) resulted in a 15% increase in the coefficient of determination (R2) compared to the Generalized Linear Model (GLM) for forecasting HAs for CVDs.

In recent years, as an advanced part of artificial intelligence, deep learning (DL) have attracted much attention in related fields owing to their strong abilities in capturing potential complex relationships among variables [18]. Khatibi et al. [19] and Wang et al. [20] proposed novel predictive models based on convolutional neural network and long short-term memory (LSTM) to predict HAs due to mental disorders and cardiopulmonary diseases, respectively.

Despite the widespread use of ML and DL models in predicting healthcare demand, these models have their disadvantages. Lightweight ML models have limited prediction capabilities, and each of them has specific predefined structures and assumptions. No single model can always be optimal in various application scenarios. DL models generally rely on massive amounts of training data and a relatively long time window for the input sequence. In this context, the stacking ensemble technique [21–23] can provide an effective solution to strike a compromise by combining the strengths of multiple lightweight ML models to achieve superior performance using limited amounts of samples. Additionally, most existing studies treated these models as "black boxes" and rarely provided explanations for them, which might reduce their acceptance by the medical community [24]. Thus, it is essential to increase the transparency of ML models in the medical domain.

In this study, we applied stacking ensemble learning based on heterogeneous lightweight ML models to forecast medical demands caused by CD considering short-term environmental exposure and explained the predictions by the SHapley Additive exPlanations (SHAP) method. The main contributions of this study can be summarized as follows:

(1) A stacking ensemble model was proposed to predict daily HAs for CD using the HAs data, air quality data, and meteorological data of the previous 6 days.
(2) A re-weighting method based on label distribution smoothing was integrated into the proposed model to address the label imbalance problem that broadly existed in healthcare data.
(3) A post-hoc interpretation for the prediction mechanism of our proposed model was provided from global and local perspectives, which is conducive to understanding the model and exploring the factors affecting HAs for CD.

Methods

Data collection and preprocessing

The daily counts of HAs due to CD and stroke from January 1, 2015 to December 31, 2018 were collected from the electronic hospitalization summary reports of all the tertiary and secondary hospitals in the urban areas of Chengdu, China (a total of 1461 observations). Patients under the age of 35 or with residential addresses outside of the urban districts of Chengdu were not included in the count of daily

HAs. All the causes of HAs were coded using the International Classification of Disease, Revision 10 (ICD-10), including the HAs for CD (I60-I69) and stroke (I60-I64).

Hourly air pollutant concentrations measured at six monitoring stations in the urban areas of Chengdu were obtained from the China National Environmental Monitoring Center (http://www.cnemc.cn/). The 24 h average concentrations of the particular matter with a diameter less than 2.5 μm ($PM_{2.5}$), the particular matter with a diameter less than 10 μm (PM_{10}), the coarse particular matter with a diameter between 2.5 μm and 10 μm (PM_C), sulfur dioxide (SO_2), nitrogen dioxide (NO_2), carbon monoxide (CO) and the 8 h moving average concentration of ozone (O_3) were calculated as their daily concentrations [25]. The air quality index (AQI) was also obtained, assessed using the air pollutants mentioned above. Because the missing rate of ambient air quality data was 2.73% (40/1461), we used linear interpolation, which has been reported as an effective data filling method when the missing rate is low (e.g., < 5%) [26] to fill in missing values.

Meteorological observations, including the daily average temperature (TEM) and relative humidity (RH), were derived from the Chengdu Meteorological Monitoring Database (http://data.cma.cn/).

Feature extraction and normalization

Based on our collected data, the HAs features, environmental features, and calendar features were extracted, as shown in Table 1.

For the time series of HAs and environmental exposure, lag features were broadly considered in epidemiological studies and HAs predictions [27, 28]. In our study, single-day lag features, namely historical values on day x ($x \in \{1, 2, 3, \ldots, L\}$) before prediction, and cumulative lag features, including the moving average and standard deviation of historical values during the previous 1 to L days were extracted. Besides, L was set to 6 to represent the short-term effect of environmental exposure as most epidemiological studies [3, 4]. In calendar features, day of the week (DOW), month (MON), season (SEA), year (YEAR), and timestamp (TS) were used to depict the trends of HAs from short to long term. Holiday (HOL), workday (WD), first work day (FWD), and last work day (LWD) were extracted to present the impact of the work-rest schedule in hospitals.

We processed DOW, MON, SEA and YEAR with One-Hot Encoding [29] and normalized features using the min–max normalization [30] as formulated in Eq. (1),

$$X^{'} = \frac{X - min(X)}{max(X) - min(X)} \tag{1}$$

Model construction

Stacking ensemble method

In this study, a stacking ensemble model comprised of four base learners and a meta learner was proposed to accurately predict the daily number of HAs for CD. As shown in Fig. 1, the development of the stacking model consists of two phases.

Table 1 Feature descriptions

Feature Name	Feature Descriptions
HAs Features	
HAs Lag x	Historical HAs on day x before the day for prediction, $x \in \{1, 2, 3, \ldots, L\}$
HAs Lag 1L mean	Moving average of historical HAs during the previous 1 to L days
HAs Lag 1L std	Standard deviation of historical HAs during the previous 1 to L days
Environmental Features	
P[a] Lag x	Historical values of P on day x before the day for prediction, $x \in \{1, 2, 3, \ldots, L\}$
P Lag 1L mean	Moving average of historical P during the previous 1 to L days
P Lag 1L std	Standard deviation of historical P during the previous 1 to L days
Calendar Features	
DOW	Day of the week, {Mon., Tues., …, Sun.}—> {1, 2, …, 7}
MON	Month of the year, {Jan., Feb., …, Dec.}—> {1, 2, …, 12}
SEA	Season of the year, {spring, summer, fall, winter}—> {1, 2, 3, 4}
YEAR	The year, {2015, 2016, 2017, 2018}—> {1, 2, 3, 4}
TS	Timestamp, serial number from 1 to 1461
HOL	Holiday, [0,1], 1 represented the day is a holiday, while 0 represented not
WD	Workday, [0,1], 1 represented the day is a work day, while 0 represented not
FWD	First work day, [0,1], 1 represented the day is the first workday, while 0 represented not
LWD	Last work day, [0,1], 1 represented the day is the last work day, while 0 represented not

[a] $P \in \{PM_{2.5}, PM_{10}, PM_C, SO_2, NO_2, CO, O_3, AQI, TEM, RH\}$

Initially, we split the whole dataset into a training set, which covered the period from January 1, 2015 to December 31, 2017, and a testing set, which covered the period from January 1, 2018 to December 31, 2018. We denoted them as $S_{train} = (X_{train}, Y_{train})$ and $S_{test} = (X_{test}, Y_{test})$, respectively, where X represented the feature set and Y represented the corresponding label set.

In the first phase, the four base learners, including linear regression with L2 regularization (Ridge) [31], random forest (RF) [32], gradient boosting decision tree (GBDT) [33], and ANN [34], were trained and used to make predictions as the input features for the meta learner. To avoid overfitting and improve the generalization capability, a five-fold cross-validation was implemented. We split S_{train} into five subsets in chronological order, and the ith subset was denoted as $S_i = (X_i, Y_i)(i = 1, 2, 3, 4, 5)$. At the ith-fold cross-validation, the jth base learner $(j = 1, 2, 3, 4)$ was trained using the subsets except S_i and made predictions on S_i, which were recorded as $P_j(X_i)$. Consequently, this process was repeated a total of 20 (4×5) times, with each base learner making predictions once on each fold. Afterwards, the predictions generated by the jth base learner throughout the five-fold cross-validation were represented as $M_j(X_{train}) = \{P_j(X_1), P_j(X_2), P_j(X_3), P_j(X_4), P_j(X_5)\}$ and treated as a new feature in the new training set. At the meantime, the jth base learner trained at the ith-fold cross-validation made predictions using S_{test}, which were recorded as $Q_{ji}(X_{test})$, and the average of predictions were calculated as a new feature in the new testing set, namely $N_j(X_{test}) = \frac{1}{5}\sum_{i=1}^{5} Q_{ji}(X_{test})$.

Furthermore, to help the meta learner decide which model to apply under a certain circumstance [23], we merged the key features selected by the base learners, namely calendar features and HAs features (as shown in Additional file 1: Fig. S1), into the new training set and the new testing set, which were denoted as X_{train_key} and X_{test_key}, respectively. Hence, at the end of the first phase, we gained a new training set $S_{new\_train} = (X_{new\_train}, Y_{train})$, where $X_{new\_train} = (M_1(X_{train}), M_2(X_{train}), M_3(X_{train}), M_4(X_{train}), X_{train\_key})$, and a new testing set $S_{new\_test} = (X_{new\_test}, Y_{test})$, where $X_{new\_test} = (N_1(X_{test}), N_2(X_{test}), N_3(X_{test}), N_4(X_{test}), X_{test\_key})$.

In the second phase, the meta learner, i.e., the elastic net [35], was trained on S_{new_train} and then used to make the final predictions on S_{new_test}.

For a suitable architecture of the stacking model, we have tested eight widely utilized lightweight ML models in the preliminary experiment (see Additional file 1: Table S1), and Ridge, RF, GBDT, and ANN were picked

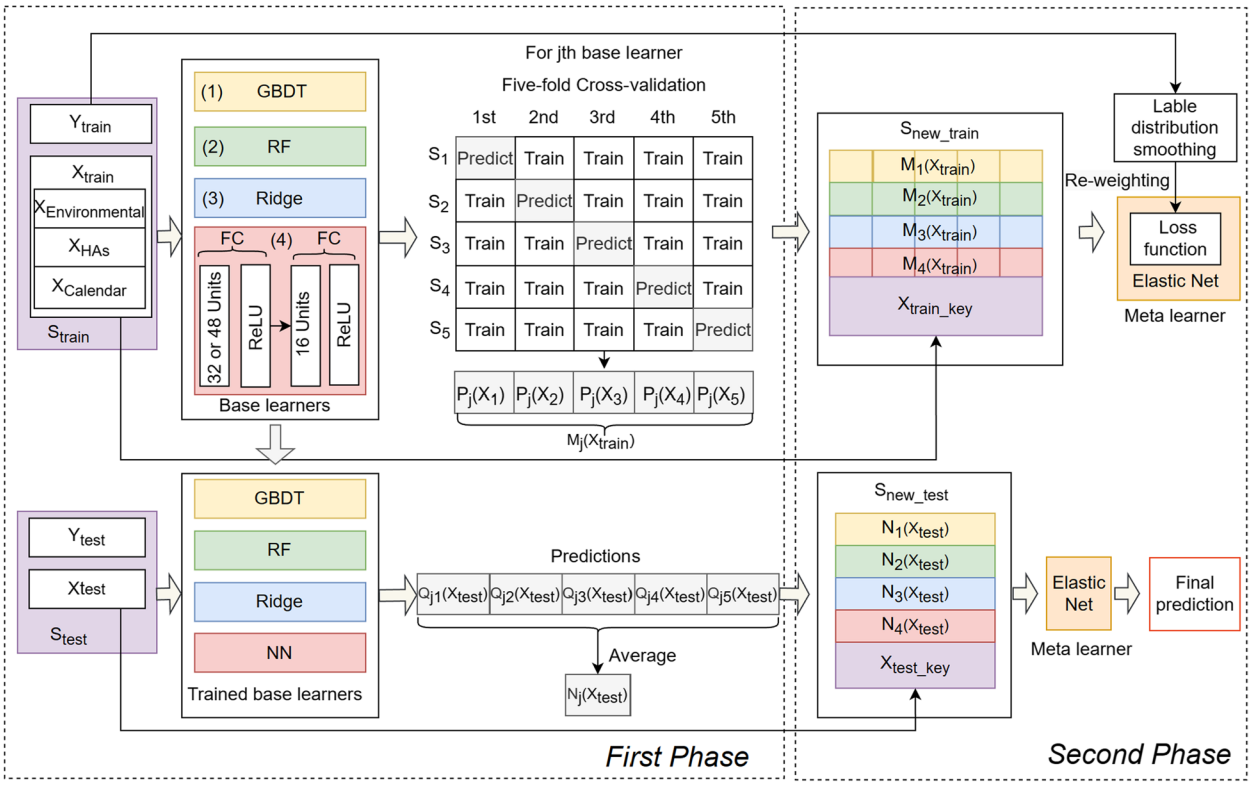

Fig. 1 Schematic diagram of stacking model development

as base learners for two reasons: 1) Each of them outperformed other models in daily HAs prediction. 2) Ridge and ANN are classical linear and network ML models, respectively. RF and GBDT are ensemble tree models based on the bagging and boosting methods, respectively. Because the theories of these models are highly heterogeneous, they can obtain insight into the training data from different perspectives and eventually increase the accuracy and robustness of the stacking model [36]. Linear regression with a combination of L1 and L2 regularization (elastic net) was selected as the meta learner because it was widely used in a similar context and can prevent overfitting to a large extent [21].

Re-weighting with Label distribution smoothing (LDS)

In our study, daily HAs data exhibit an imbalanced distribution, where certain target values, especially peaks and troughs, have strikingly fewer observations. For classification tasks, re-sampling and re-weighting are the two main methods to address data imbalance. However, methods based on re-sampling, such as SMOTE [37] and SMOGN [38], are not directly applicable to our task, because the distance between labels was not considered and the intrinsic seasonal pattern of HAs might be damaged. We adopt the LDS method [39] to extend re-weighting schemes to regression tasks, which includes the following steps: First, discretize the continuous target space into finite bins, which can be considered as the empirical label density distribution. Then convolve the empirical label density with a symmetric kernel to calculate the effective label density that accounts for the overlap in the information of nearby labels so that the cost-sensitive re-weighting method can be utilized based on the effective label distribution.

To integrate this approach into our proposed model, in the training process of the meta learner, we used the inverse of effective label density as the weight of training samples when calculating the loss function, and given that the HAs data show a yearly rising trend and an annual seasonal pattern, it is more reasonable to calculate the LDS estimated label density within each year as shown in Fig. 2.

Training details and parameters

In our experiment, the ANN were completed using Keras with Tensorflow 2.4.1 as the backend. Other base learners were implemented based on the Scikit-learn 0.24.2 Python library. The computation was performed using AMD Ryzen 74800U with Radeon Graphics 1.80 GHz. In the stacking model, the hyper-parameters of the base learners and the meta learner were tuned with the last 20% of the original training dataset and the last 20% of the new training dataset, respectively. The grid search was applied, and the best hyper-parameter combinations were illustrated in Additional file 1: Table S2.

Model evaluation

To demonstrate the superiority of our stacking model, we compared the stacking model with all base learners and LSTM [20] on the testing set. In the LSTM, the input variables only included historical HAs and environmental features, and the time window of the input sequence was set to 6, which was consistent with the lag days of other methods. The hyper-parameters of the benchmarks were also tuned with the last 20% of the original training dataset (shown as Additional file 1: Table S2).

Four metrics, including mean absolute error (MAE), root mean square error (RMSE), mean absolute percentage error (MAPE), and coefficient of determination (R^2) were used to evaluate the effectiveness of the prediction models as Hu et al. [22].

Model explainability

To explain the predictions of our final model, we made use of the permutation explainer implemented in the SHAP Python library (version 0.39.0). SHAP [40] is a unified approach based on the additive feature attribution method that interprets the difference between an actual prediction and the baseline as the sum of the attribution values, i.e., SHAP values, of each feature. In this study, the SHAP value for each feature in a given sample of CD dataset was calculated based on our proposed stacking model to present its contribution to the variation of HAs predictions. For the historical HAs and environmental features, their SHAP values were regarded as the sum of the SHAP values of all single-day lag and cumulative lag features, rendering their contributions during the previous 6 days.

A post-hoc interpretation was provided by analyzing the SHAP values from two perspectives. On the global scale, the SHAP values over all training samples were holistically analyzed to reveal how the stacking model fits the relationship between daily HAs and predictors. On the local scale, the SHAP values in several samples selected from the testing set were investigated to disclose how the predictions were generated in the effect of environmental features.

Results

Descriptive statistics

The summary statistics of daily HAs, air pollutants, and meteorological indicators are shown in Table 2, and the

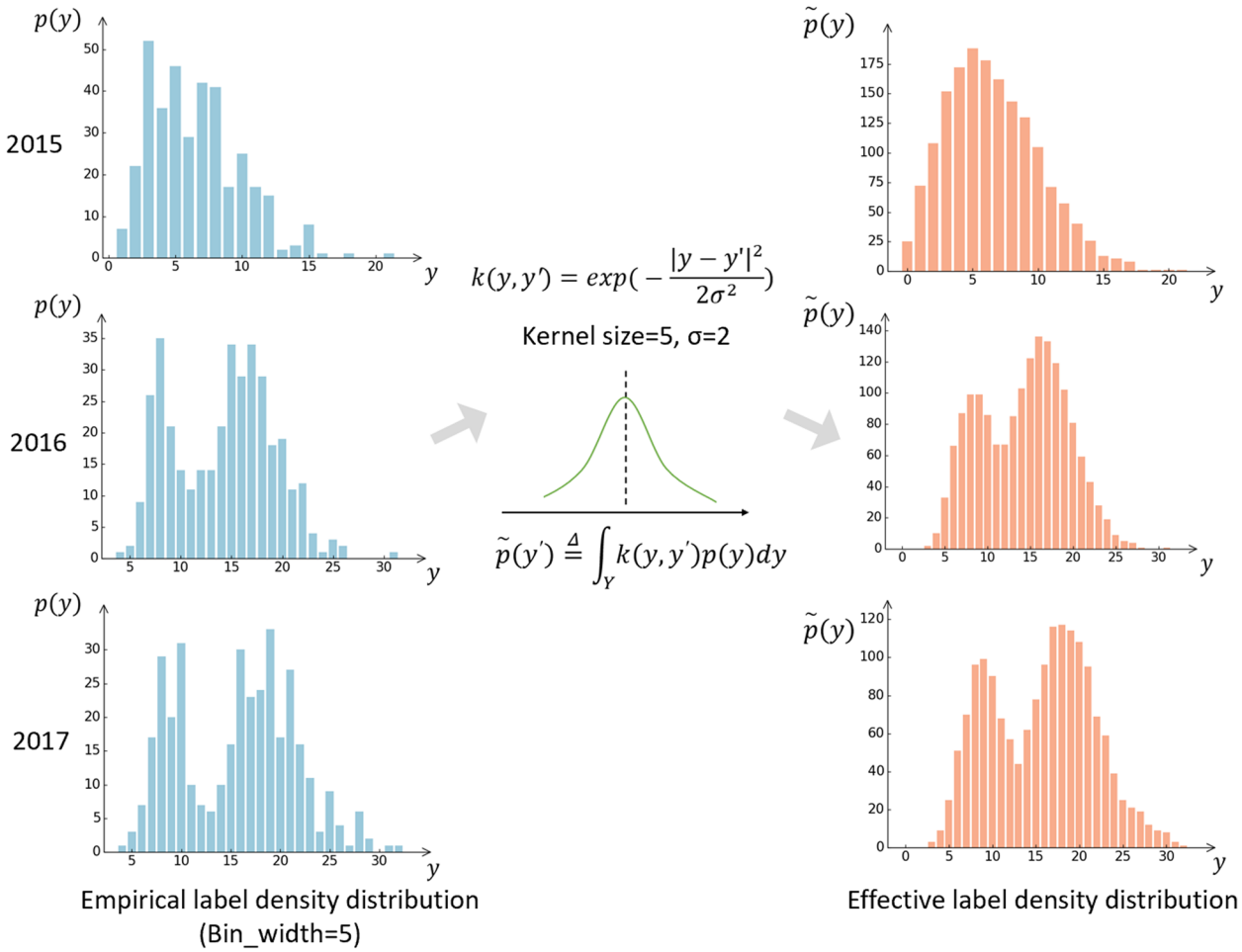

Fig. 2 Schematic diagram of label distribution smoothing

corresponding temporal variations of them are visualized in Additional file 1: Figs. S2 and S3, respectively. The correlations between environmental exposure variables are shown in Additional file 1: Table S3.

From 2015 to 2018, the total number of HAs for CD was 102,708, and the average number of daily HAs was 70 (std = 35). The daily mean ± std concentrations of $PM_{2.5}$, PM_{10}, PM_C, O_3, SO_2, NO_2, and CO were 57.9 ± 40.6, 95.7 ± 62.1, 37.8 ± 25.6, 96.6 ± 54.6, 12.7 ± 5.5, 53.9 ± 17.7, and 1030 ± 360 μg/m^3, respectively. The value of AQI ranged from 16.7 to 404.6, with a mean of 85.2. The mean TEM and RH were 16.9℃ and 80.5%, respectively.

Model performance

Table 3 compares the performance of base learners, LSTM, and the proposed stacking model on CD dataset and stroke dataset.

On both datasets, ANN and LSTM surpass other individual models in terms of MAE, RMSE, and R^2, but gain higher MAPE than tree-based models, and the stacking model is substantially superior to all individual models. After using LDS, the performance of the stacking model is further improved. On CD dataset, compared with the optimal results obtained by individual models, the MAE, RMSE, and MAPE of the stacking model with LDS remarkably reduced by 13.9%, 12.7%, and 5.8%, respectively, and the R^2 increased by 6.8%. Additionally, the results of the t-test indicate that, when evaluated by most metrics, the performance gap between the stacking model with LDS and the best individual model is significant, and the difference between the R^2 of them is visualized in Additional file 1: Fig. S4. Figure 3 shows a comparison between the observed HAs and the predictions of the stacking model with LDS on two datasets.

Model explanation

Global explanation

Figure 4 shows the distribution of SHAP values of each feature in chronological order, and the features are

Table 2 Descriptive statistics of daily HAs for CD and environmental exposure data in Chengdu, 2015–2018

	Variables	Units	Mean	Std[a]	Min	25%	50%	75%	Max
Daily HAs	CD HAs	persons	70	35	6	42	68	96	214
	Stroke HAs	persons	45	21	4	27	43	60	131
Air pollutants	$PM_{2.5}$	$\mu g/m^3$	57.9	40.6	6.1	29.6	46.3	74.5	324.5
	PM_{10}	$\mu g/m^3$	95.7	62.1	12.0	51.6	78.4	124.5	492.5
	PM_C	$\mu g/m^3$	37.8	25.6	3.9	20.3	30.9	48.0	238.2
	O_3	$\mu g/m^3$	96.6	54.6	5.6	54.2	86.2	135.2	290.4
	SO_2	$\mu g/m^3$	12.7	5.5	3.9	8.5	11.2	15.3	37.9
	NO_2	$\mu g/m^3$	53.9	17.7	13.9	41.0	51.9	64.6	130.4
	CO	mg/m^3	1.0	0.4	0.4	0.8	1.0	1.2	2.8
	AQI	1	85.2	48.5	16.7	52.5	71.4	103.8	404.6
Meteorological measures	TEM	℃	16.9	7.3	-1.1	10.1	17.4	23.2	30.2
	RH	%	80.5	9.2	43.0	74.4	80.8	87.7	99.3

[a] *Std* standard deviation

Table 3 Performance comparison of different methods in predicting HAs for CD and stroke

Datasets	Models	MAE	RMSE	MAPE	R^2
CD	RF	14.713	20.649	0.154	0.652
	GBDT	14.661	20.296	0.154	0.663
	Ridge	14.894	19.963	0.183	0.674
	ANN	14.408	18.407	0.191	0.723
	LSTM	13.774	18.421	0.165	0.739
	Stacking	12.467	17.053	0.153	0.762
	Stacking + LDS	**11.855**[*]	**16.078**[*]	**0.145**	**0.789**
Stroke	RF	10.889	14.660	0.175	0.51
	GBDT	11.251	14.995	0.178	0.487
	Ridge	10.357	13.278	0.210	0.598
	ANN	9.525	12.140	0.191	0.664
	LSTM	9.422	12.166	0.185	0.676
	Stacking	9.038	11.898	0.170	0.677
	Stacking + LDS	**8.961**[*]	**11.850**	**0.159**[*]	**0.680**

The best result for each metric is in bold.

[*] The differences in the MAE, RMSE, or MAPE between the stacking model with LDS and the best individual model are significant (*P*-value < 0.05) according to the t-test

ranked according to the average of their absolute SHAP values over all the training samples, which represents their global importance.

Most straightforwardly, a calendar feature nearly always had similar SHAP values when it remained at the same values, resulting in visually prominent color blocks with a periodic alternation. In contrast, the impact of environmental features varied over samples without explicit patterns. In light of additive feature attribution theory, the predicted HAs could be regarded as the additive combination of three parts: a baseline (generally the average of predictions), a regular difference attributed by calendar features and historical HAs, and a highly volatile difference attributed by environmental features. As shown in Table 4, the second part depicted the general trend of variations in the number of daily HAs, and the third part served to further enhance the model performance.

After further observing each feature shown in Fig. 4, we found that TS and historical HAs played a major part in profiling the growth trend in HAs, and DOW served to depict the periodic variation of HAs caused by the work-rest schedule in hospitals (see Additional file 1: Fig. S2). TEM in the fall and summer contributed to increasing the predicted HAs. Notably, the annual peak concentrations of several air pollutants, such as O3, CO, and $PM_{2.5}$, and their SHAP values pushing up the predicted HAs occurred at similar times (see Additional file 1: Fig. S3).

Local explanation

As the contributions of calendar features and HAs features are relatively straightforward and regular, it makes more sense to concentrate on how the involvement of environmental features improves the model performance. Thus, the samples on every Wednesday in August 2018 were selected to fix the calendar features except TS, where August was selected to further explore the impact of high temperatures on the risk of CD, and Wednesday was selected to reduce the interference caused by weekend breaks and Chinese holidays. The sum of the first two parts of the predictions mentioned above was set as a new baseline. Figure 5 shows how the SHAP values of environmental features were accumulated from the new baseline to reach the final predictions.

By comparing Fig. 5 and Additional file 1: Fig. S5, it was observed that the TEM in August and peak concentrations of $PM_{2.5}$ and O_3 that appeared in the six

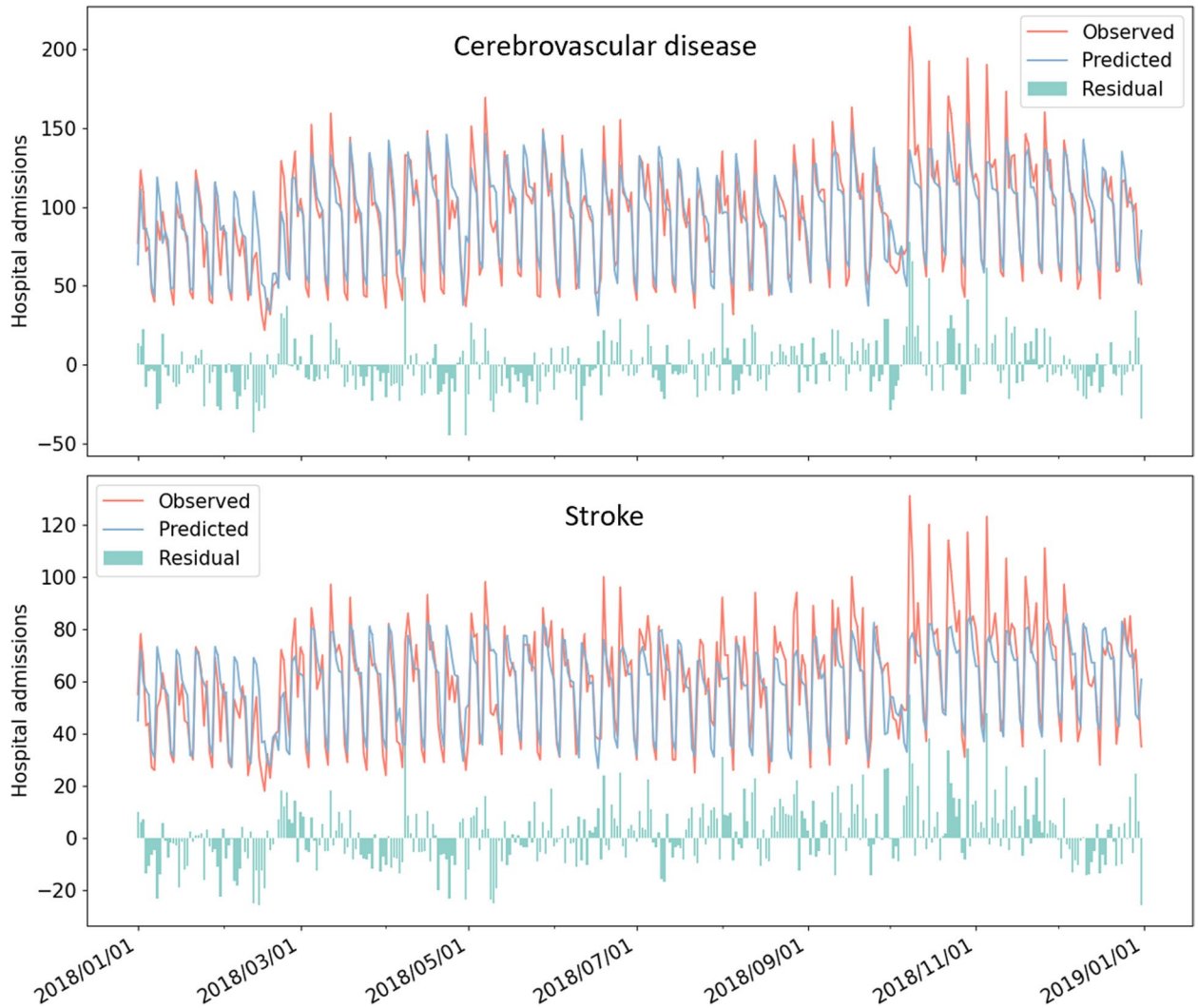

Fig. 3 The comparison and residual between the observed HAs and the predictions of the stacking model with LDS on CD dataset and stroke dataset

days leading up to August 28 served to increase the predicted values of HAs, while gradually declining RH over the previous 6 days lowered the predicted values on August 21 by around 1 count.

Discussion

Model performance analysis

A distinct improvement of the stacking model compared to the base learners can be attributed to three aspects: 1) ANN and tree-based models performed best when evaluated by different metrics, which reflected the heterogeneity of their predictions and laid the foundation for meta learner to combine their strengths and obtain a better generalization [36]. 2) The key features served to help the meta learner understand how to choose and combine the predictions of base learners under various circumstances. 3) The re-weighting method based on LDS reduced the error caused by label imbalance, as shown in Fig. 6.

As shown in Additional file 1: Table S3, there are significant correlations among environmental exposure variables, however, the performance of the stacking model was not impacted because the base learners and the meta learner we selected can effectively handle multicollinearity. In Ridge and elastic net, the L2 regularization term added to their loss function can help stabilize the estimates and reduce overfitting in the presence of collinearity [41].

LSTM has been successfully applied in many fields, but in our task, its prediction performance is still inferior to the proposed stacking model. There are two potential

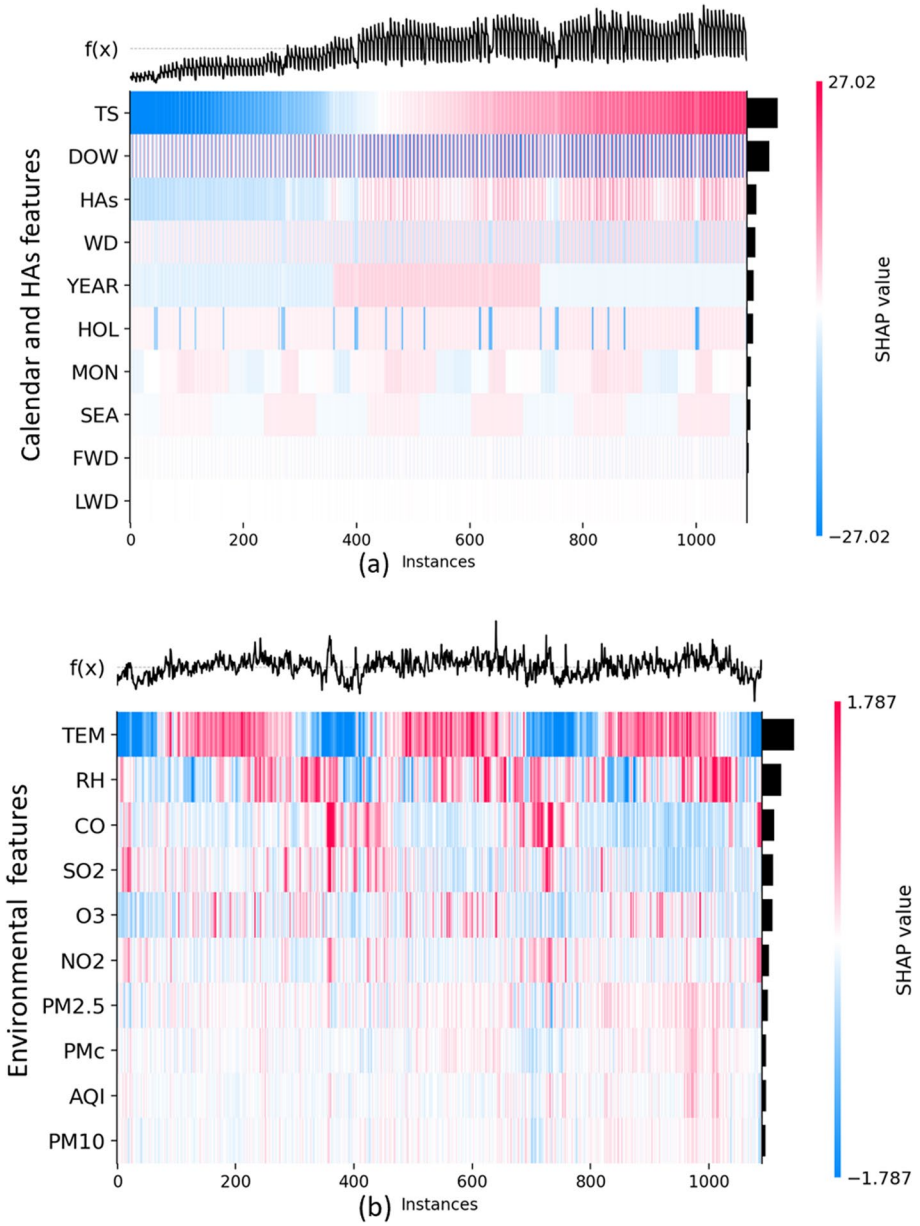

Fig. 4 Heatmap plot of SHAP values of all features across all samples in the CD training set. The width of the black bar on the right-hand side shows the global importance of each feature. **a** Calendar features and HAs features **b** Environmental features

Table 4 The improvement of model performance attributed to environmental features estimated by SHAP

	MAE	RMSE	MAPE	R^2
Baseline + SHAP values of HAs features and calendar features	12.306	16.584	0.152	0.775
Final prediction	11.855	16.078	0.145	0.789
Improvement	3.665%	3.054%	4.306%	1.743%

reasons: 1) The size of our dataset is limited, with a total of 1461 samples, therefore, using lightweight models with simple architectures and fewer parameters can help avoid overfitting. 2) In our task, the time lag was set to 6 days to consider the short-term effect of environmental exposure, which constrained the advantage of LSTM in learning the long-time dependencies. However, the calendar features we extracted, such as YEAR, SEA, and MON, can assist the stacking model in capturing long-term trends.

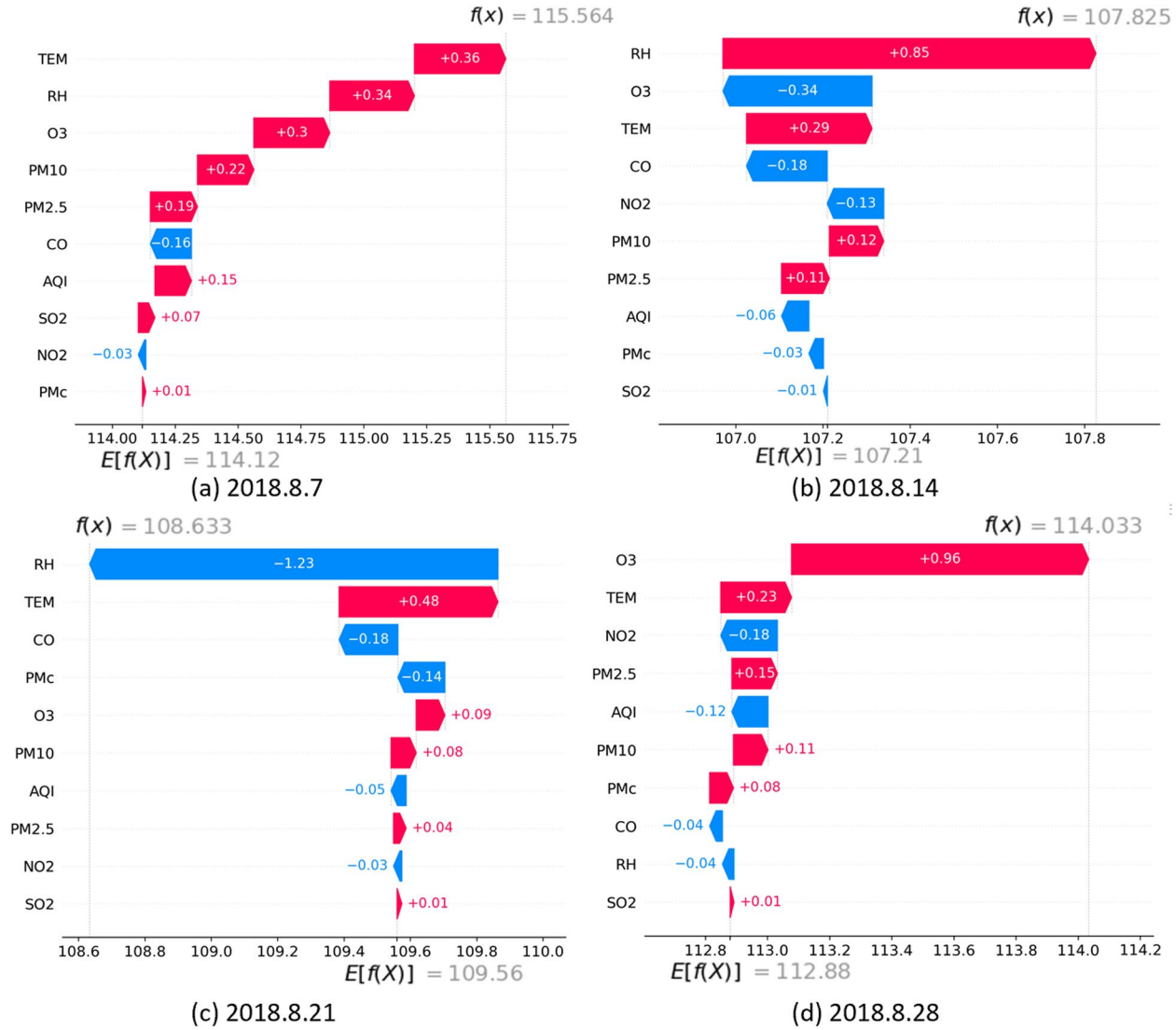

Fig. 5 Waterfall plot of SHAP values to four selected samples, i.e., samples on August 7, 14, 21 and 28, 2018. The new baselines and the final predictions are marked at the bottom and top of the image, respectively. The SHAP values of each feature are listed on the bar

Comparing SHAP explanation and conventional association analysis

In our study, the explanations obtained by SHAP displayed a strong agreement with the conclusion of association analysis using conventional analytical methods based on statistical models, such as GLMs and generalized additive models (GAMs). For example, the TEM in summer and the peak values of air pollutant concentrations always played a role in describing the increase in HAs. This is consistent with the previous studies, which found high levels of air pollution and high temperatures were associated with a high morbidity of stroke [4, 5, 42, 43], but we did not observe that cold weather served to drive up the predicted HAs, probably because the minimum TEM (-1.1℃) during our research period in Chengdu did not reach the extreme cold defined in related studies. This consistency indicates that our stacking model can accurately describe the relationship between daily HAs for CD and environmental factors, thus improving the model performance.

Moreover, if we associate SHAP values that push up the predicted HAs with an increased risk of CD, the model explanations can be regarded as a new perspective to explore the adverse health effects of air pollutants and extreme weather conditions. For example, the relationship between the risk of CD and the lagged TEM and RH can be depicted in Fig. 7. Furthermore, SHAP can comprehensively account for the combined impact of multiple

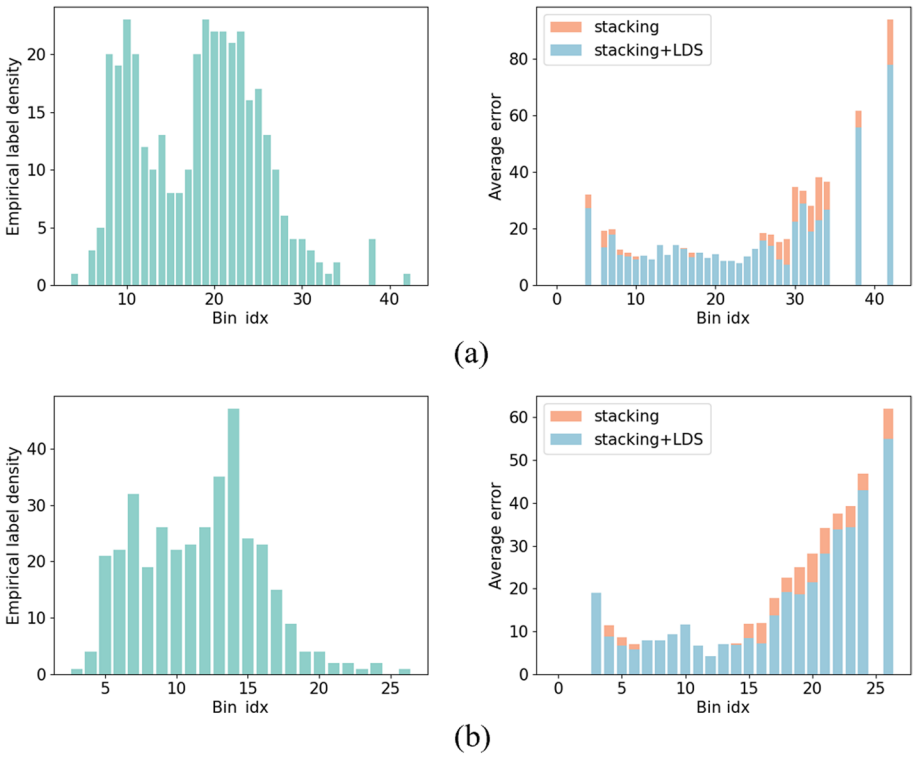

Fig. 6 The left side shows empirical label distribution plots, and the right side shows comparison plots of error before and after using LDS on two testing datasets: **a** CD and **b** stroke

environmental factors, whereas most traditional methods can only analyze the association between single environmental factors and HAs after controlling confounding effects among multiple covariates by smoothing functions.

Limitations

Several limitations should be addressed in our study. First, limited by the available data sources, we only considered the impact of ambient air pollutants and meteorological conditions on daily HAs for CD, but some other environmental factors and individual health behaviors may also play important roles in the development and severity of CD. Second, our proposed model is only applicable to predict HAs for non-communicable diseases, such as CD, which are associated with environmental exposure, and the model might not be suitable for forecasting the daily number of hospitalizations for infectious diseases. Third,

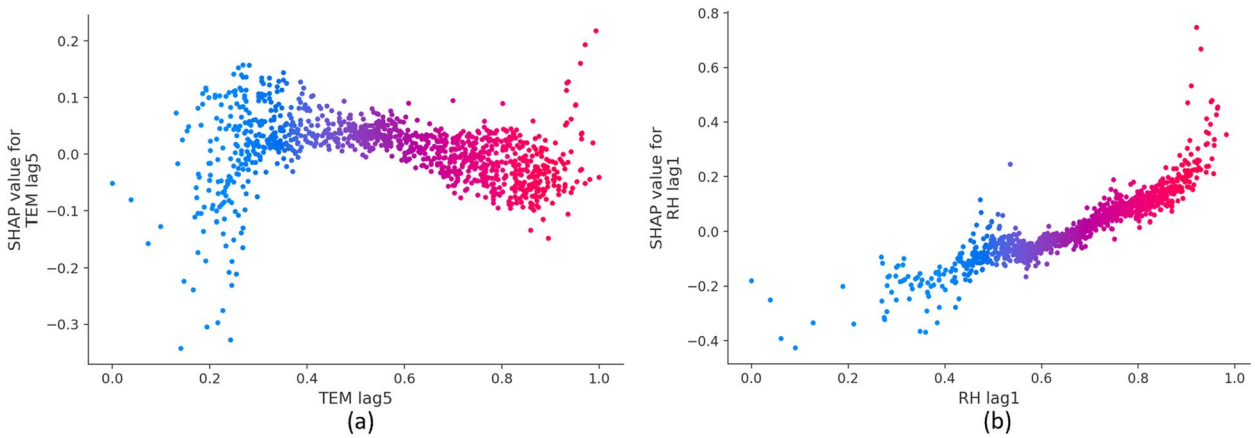

Fig. 7 SHAP dependence plots that show the effect of TEM lag5 and RH lag1 on the predictions of HAs

the peak values of HAs are not well predicted, which could lead to under-allocation of medical resources.

Conclusions

This study proposed a stacking ensemble model to predict the daily number of HAs for CD using the HAs data, air quality data, and meteorological data. The experimental results showed that our proposed model is superior to the base learners and LSTM on two datasets under four evaluation criteria. Moreover, the model explanation demonstrated that environmental factors played a role in further improving the model performance and identified that high TEM and high concentrations of gaseous air pollutant might strongly associate with an increased risk of CD. This study indicates that the proposed model considering environmental exposure factors is efficient in predicting daily HAs for CD and has practical value for hospital management teams in early warning and healthcare resource allocation.

Abbreviations

CD	Cerebrovascular disease
CVD	Cardiovascular disease
HAs	Hospital admissions
ML	Machine learning
SES	Simple exponential smoothing
ARIMA	Autoregressive integrated moving average
LSTM	Long Short-Term Memory
GLM	Generalized linear model
GAM	Generalized additive models
AQI	Air quality index
TEM	Temperature
RH	Relative humidity
DOW	Day of the week
MON	Month
SEA	Season
TS	Timestamp
HOL	Holiday
WD	Workday
FWD	First work day
LWD	Last work day
LDS	Label distribution smoothing
RF	Random forest
GBDT	Gradient boosting decision tree
ANN	Artificial neural network
MAE	Mean absolute error
RMSE	Root mean square error
MAPE	Mean absolute percentage error
R^2	Coefficient of determination
SHAP	SHapley Additive exPlanations

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12911-023-02159-7.

Additional file 1: An additional file provided supplementary figures and tables for comprehension of our research.

Acknowledgements

We thank the Health Information Center of Sichuan Province for its permission to use the data.

Authors' contributions

XL performed the experiments, analyzed the data and wrote the first draft of the manuscript. HQ conceived the study and revised the manuscript. All authors have read and approved the final manuscript.

Funding

This work has been supported by the National Natural Science Foundation of China (No. 71661167005) and the Key Research and Development Program of Sichuan Province, China (No. 2019YFS0271). The content is solely the responsibility of the authors and does not necessarily represent the official views of the funders. The funding bodies did not play any role in the design of the study, collection, analysis, and interpretation of the data, or writing the manuscript.

Availability of data and materials

The meteorological and air quality data are available at http://data.cma.cn/ and http://www.cnemc.cn/. Daily data of hospitalizations for cerebrovascular disease are available from the Health Information Center of Sichuan Province, but restrictions apply to the availability of these data, which were used under license for the current study, and so are not publicly available. The daily number of hospitalizations for cerebrovascular disease are however available from corresponding author on reasonable requests and with permission of the Health Information Center of Sichuan Province.

Declarations

Ethics approval and consent to participate

This study was conducted according to the ethical guidelines of the Helsinki Declaration and was approved by the Ethics Committee of Health Information Center of Sichuan Province. The Ethics Committee of Health Information Center of Sichuan Province exempted informed consent because this research only involved daily number of hospitalizations for cerebrovascular disease, not individual data.

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Received: 3 December 2022 Accepted: 23 March 2023
Published online: 06 April 2023

References

1. WHO mortality database: the number of deaths caused by cerebrovascular disease. https://platform.who.int/mortality/themes/theme-details/topics/indicator-groups/indicator-group-details/MDB/cerebrovascular-disease. Accessed 3 Sep 2022.
2. China TWC of the R on CH and D in. Report on Cardiovascular Health and Diseases in China. An Updated Summary. Biomed Environ Sci. 2021;2022(35):573–603.
3. Vered S, Paz S, Negev M, Tanne D, Zucker I, Weinstein G. High ambient temperature in summer and risk of stroke or transient ischemic attack: a national study in Israel. Environ Res. 2020;187:109678.
4. Liu H, Tian Y, Xu Y, Huang Z, Huang C, Hu Y, et al. Association between ambient air pollution and hospitalization for ischemic and hemorrhagic stroke in China: a multicity case-crossover study. Environ Pollut. 2017;230:234–41.
5. Abedi A, Baygi MM, Poursafa P, Mehrara M, Amin MM, Hemami F, et al. Air pollution and hospitalization: an autoregressive distributed lag (ARDL) approach. Environ Sci Pollut Res. 2020;27:30673–80.

6. Sun Q, Wang A, Jin X, Natanzon A, Duquaine D, Brook RD, et al. Long-term air pollution exposure and acceleration of atherosclerosis and vascular inflammation in an animal model. JAMA. 2005;294:3003–10.
7. Mills NL, Törnqvist H, Robinson SD, Gonzalez M, Darnley K, MacNee W, et al. Diesel exhaust inhalation causes vascular dysfunction and impaired endogenous fibrinolysis. Circulation. 2005;112:3930–6.
8. Kaufman JD, Adar SD, Barr RG, Budoff M, Burke GL, Curl CL, et al. Association between air pollution and coronary artery calcification within six metropolitan areas in the USA (the Multi-Ethnic Study of Atherosclerosis and Air Pollution): a longitudinal cohort study. Lancet Lond Engl. 2016;388:696–704.
9. Franchini M, Mannucci PM. Thrombogenicity and cardiovascular effects of ambient air pollution. Blood. 2011;118:2405–12.
10. Luo L, Luo L, Zhang X, He X. Hospital daily outpatient visits forecasting using a combinatorial model based on ARIMA and SES models. BMC Health Serv Res. 2017;17:469.
11. Ordu M, Demir E, Tofallis C. A comprehensive modelling framework to forecast the demand for all hospital services. Int J Health Plann Manage. 2019;34:e1257–71.
12. Jahan S, Wraith D. Immediate and delayed effects of climatic factors on hospital admissions for schizophrenia in Queensland Australia: a time series analysis. Environ Res. 2021;197:111003.
13. Zhang X, Yu Y, Xiong F, Luo L. Prediction of daily blood sampling room visits based on ARIMA and SES model. Comput Math Methods Med. 2020;2020:1720134.
14. Huck N. Large data sets and machine learning: applications to statistical arbitrage. Eur J Oper Res. 2019;278:330–42.
15. Qiu H, Luo L, Su Z, Zhou L, Wang L, Chen Y. Machine learning approaches to predict peak demand days of cardiovascular admissions considering environmental exposure. BMC Med Inform Decis Mak. 2020;20:83.
16. Bibi H, Nutman A, Shoseyov D, Shalom M, Peled R, Kivity S, et al. Prediction of emergency department visits for respiratory symptoms using an artificial neural network. Chest. 2002;122:1627–32.
17. Kassomenos P, Petrakis M, Sarigiannis D, Gotti A, Karakitsios S. Identifying the contribution of physical and chemical stressors to the daily number of hospital admissions implementing an artificial neural network model. Air Qual Atmosphere Health. 2011;4:263–72.
18. Schmidhuber J. Deep learning in neural networks: An overview. Neural Netw. 2015;61:85–117.
19. Khatibi T, Karampour N. Predicting the number of hospital admissions due to mental disorders from air pollutants and weather condition descriptors using stacked ensemble of Deep Convolutional models and LSTM models (SEDCMLM). J Clean Prod. 2021;280:124410.
20. Wang C, Qi Y, Zhu G. Deep learning for predicting the occurrence of cardiopulmonary diseases in Nanjing. China Chemosphere. 2020;257:127176.
21. Breiman L. Stacked regressions. Mach Learn. 1996;24:49–64.
22. Hu Z, Qiu H, Su Z, Shen M, Chen Z. A stacking ensemble model to predict daily number of hospital admissions for cardiovascular diseases. IEEE Access. 2020;8:138719–29.
23. Navares R, Díaz J, Linares C, Aznarte JL. Comparing ARIMA and computational intelligence methods to forecast daily hospital admissions due to circulatory and respiratory causes in Madrid. Stoch Environ Res Risk Assess. 2018;32:2849–59.
24. Zhang S, Wang J, Pei L, Liu K, Gao Y, Fang H, et al. Interpretability analysis of one-year mortality prediction for stroke patients based on deep neural network. IEEE J Biomed Health Inform. 2022;26:1903–10.
25. Ministry of Environmental Protection of the People's Republic of China, Ambient air quality standards. 2016. https://www.mee.gov.cn/ywgz/fgbz/bz/bzwb/dqhjbh/dqhjzlbz/201203/W020120410330232398521.pdf. Accessed 9 Aug 2022.
26. Norazian MN, Shukri YA, Azam RN, Al Bakri AM. Estimation of missing values in air pollution data using single imputation techniques. ScienceAsia. 2008;34:341.
27. Ho AFW, Lim MJR, Zheng H, Leow AS-T, Tan BY-Q, Pek PP, et al. Association of ambient air pollution with risk of hemorrhagic stroke: A time-stratified case crossover analysis of the Singapore stroke registry. Int J Hyg Environ Health. 2022;240:113908.
28. Polezer G, Tadano YS, Siqueira HV, Godoi AFL, Yamamoto CI, de André PA, et al. Assessing the impact of PM2.5 on respiratory disease using artificial neural networks. Environ Pollut. 2018;235:394–403.
29. Qiao Y, Yang X, Wu E. The research of BP neural network based on one-hot encoding and principle component Analysis in determining the therapeutic effect of diabetes mellitus. IOP Conf Ser Earth Environ Sci. 2019;267:042178.
30. Han J, Kamber M, Pei J. 3 - Data Preprocessing. In: Han J, Kamber M, Pei J, editors. Data Mining. 3rd ed. Boston: Morgan Kaufmann; 2012. p. 83–124.
31. Hilt DE, Seegrist DW, States U, Northeastern Forest Experiment Station (Radnor P). Ridge, a computer program for calculating ridge regression estimates. Upper Darby, Pa: Dept. of Agriculture, Forest Service, Northeastern Forest Experiment Station; 1977.
32. Breiman L. Random Forests. Mach Learn. 2001;45:5–32.
33. Friedman JH. Greedy function approximation: a gradient boosting machine. Ann Stat. 2001;29:1189–232.
34. Araujo LN, Belotti JT, Alves TA, Tadano Y de S, Siqueira H. Ensemble method based on artificial neural networks to estimate air pollution health risks. Environ Model Softw. 2020;123:104567.
35. Zou H, Hastie T. Regularization and variable selection via the elastic net. J R Stat Soc Ser B Stat Methodol. 2005;67:301–20.
36. Chen L, Wang C, Song S. Software defect prediction based on nested-stacking and heterogeneous feature selection. Complex Intell Syst. 2022;8:3333–48.
37. Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP. SMOTE: synthetic minority over-sampling technique. J Artif Intell Res. 2002;16:321–57.
38. Branco P, Torgo L, Ribeiro RP. SMOGN: a pre-processing approach for imbalanced regression. 2017.
39. Yang Y, Zha K, Chen Y-C, Wang H, Katabi D. Delving into deep imbalanced regression. 2021.
40. Lundberg S, Lee S-I. A unified approach to interpreting model predictions. 2017.
41. McDonald GC. Ridge regression. Wiley Interdiscip Rev Comput Stat. 2009;1:93–100.
42. Tian Y, Liu H, Si Y, Cao Y, Song J, Li M, et al. Association between temperature variability and daily hospital admissions for cause-specific cardiovascular disease in urban China: a national time-series study. PLoS Med. 2019;16:e1002738.
43. Aklilu D, Wang T, Amsalu E, Feng W, Li Z, Li X, et al. Short-term effects of extreme temperatures on cause specific cardiovascular admissions in Beijing. China Environ Res. 2020;186:109455.

Publisher's Note