Machine Learning Approaches to Predict Peak Demand Days of Cardiovascular Admissions Considering Environmental Exposure

Hang Qiu (qiuhang@uestc.edu.cn)
University of Electronic Science and Technology of China https://orcid.org/0000-0002-5380-2870
Lin Luo
University of electronic science and technology of China
Ziqi Su
University of British Columbia
Li Zhou
Health Information Center of Sichuan Province
Liya Wang
University of Electronic Science and Technology of China
Yucheng Chen
Sichuan University

Research article

Keywords: machine learning, cardiovascular disease, hospital admission, prediction, environmental exposure

Posted Date: December 28th, 2019

DOI: https://doi.org/10.21203/rs.2.19636/v1

Version of Record: A version of this preprint was published on May 1st, 2020. See the published version at https://doi.org/10.1186/s12911-020-1101-8.

Abstract

Background

Accumulating evidence has linked environmental exposures, such as ambient air pollution and meteorological factors to the development and severity of cardiovascular diseases (CVDs), resulting in increased healthcare demand. Effective prediction of situations of demand for healthcare services particularly those associated with peak events of CVDs can be useful in optimizing the allocation of medical resources. However, few studies have attempted to adopt machine learning approaches with excellent predictive abilities to forecast the healthcare demand for CVDs. This study aims to develop machine learning models to predict the peak demand days of CVDs admissions using the hospital admissions data, air quality data and meteorological data in Chengdu, China from 2015 to 2017.

Methods

Six machine learning algorithms, including logistic regression (LR), support vector machine (SVM), random forest (RF), extreme gradient boosting (XGBoost), light gradient boosting machine (LightGBM) and artificial neural network (ANN), were applied to build the predictive models. The area under a receiver operating characteristic curve (AUC), logarithmic loss function, accuracy, sensitivity, specificity and F1 score were used to evaluate the predictive performances among the six models.

Results

The LightGBM model exhibited the highest AUC (0.940, 95% CI: 0.900-0.980), which was significantly higher than that of LR (0.842, 95% CI: 0.783-0.901), SVM (0.834, 95% CI: 0.774-0.894) and ANN (0.890, 95% CI: 0.836-0.944), but did not differ significantly from that of RF (0.926, 95% CI: 0.879-0.974) and XGBoost (0.930, 95% CI: 0.878-0.982). In addition, the LightGBM has the optimal logarithmic loss function (0.218), accuracy (91.3%), specificity (94.1%) and F1 score (0.725). Feature importance identification based on LightGBM indicated that the contribution rate of meteorological conditions and air pollutants for the prediction was 32% and 43%, respectively.

Conclusion

This study suggests that ensemble learning models especially the LightGBM model can be used to effectively predict the peak events of CVDs, which provide decision making for medical resource management.

Background

Cardiovascular diseases (CVDs) are the leading cause of worldwide mortality [1]. It has been estimated that 17.9 million deaths were attributable to CVDs in 2016, representing approximately 31% of all global deaths in that year [1]. China is one of the countries in the world with the highest burden of CVDs [2]. According to the Report on Cardiovascular Diseases released by the National Center for Cardiovascular

Diseases 2017, there were about 290 million people with CVDs in that year and the prevalence of CVDs is trending upwards [3].

Even though behavioral factors, including physical inactivity, smoking, unhealthy diets and obesity, are the well-known risk factors for CVDs, a large body of studies have indicated that environmental exposure [4], such as ambient air pollution [5–9] and temperature variability [10–12], also makes significant contributions to CVDs, resulting in increased risk of morbidity. For example, using conditional logistic regression models, Liu et al. [13] conducted a multi-city study in 26 Chinese cities, and the results showed that elevated concentrations of sulfur dioxide (SO_2), nitrogen dioxide (NO_2), carbon monoxide (CO), and ozone (O_3) were associated with increased risk of hospitalization for heart failure. Another national time-series study conducted in 184 Chinese cities linked temperature variability to the increase of hospital admissions for CVDs and its subtypes using over-dispersed Poisson regression models [14]. Although these statistical regression models can assess the associations of environmental exposure with CVDs morbidity [15–17], they are often incapable of providing sufficiently accurate morbidity prediction for healthcare management. Moreover, we lack information on the effect of complex mixture of environmental exposure on CVDs morbidity.

With the increasing number of CVDs patients, the contradiction between the growing demand of patients for healthcare services and the limited medical resources is becoming prominent, which makes the prediction of future healthcare demand particularly those associated with peak event has gained greater attention. Time series forecasting approaches, such as the autoregressive integrated moving average (ARIMA) model and the seasonal ARIMA (SARIMA) model, are widely applied in predicting problems about emergency department visits [18, 19], new admission inpatients [20] and inpatients discharge [21]. However, these models seem to be difficult to solve the complex nonlinear relationship among multi-factors and their forecasting abilities to extrapolate are limited.

Recently, machine learning algorithms, which can solve the nonlinear relationship among multi-dimensional variables, have been shown to be effective in prediction, and are being used successfully in various healthcare applications, such as medical diagnosis [22, 23] and disease risk prediction [24, 25]. Nevertheless, only a very limited number of studies have attempted to adopt machine learning-based data-driven approaches to forecast the demand for healthcare services associated with environmental exposure, and these few studies predominately focused on the application of artificial neural network (ANN) [26–29]. For instance, Kassomenos et al. [30] applied ANN and stepwise regression models to predict the daily number of hospital admissions for CVDs and respiratory diseases considering air pollution and meteorological conditions, and ANN performed better than regression model. Khatri et al. [31] used ANN to forecast peak demand days of emergency department visits for chronic respiratory diseases based on weather and environmental pollution. Although part of other machine learning algorithms performed better than ANN in other fields [32], it is unclear how effective the other machine learning approaches are in predicting the healthcare services demand associated with environmental exposure.

In the present study, we aimed to develop several machine learning models to predict the peak demand days of CVDs admissions based on hospital admissions data, air quality data and meteorological data in Chengdu, China from 2015 to 2017. Six types of machine learning models, including logistic regression (LR), support vector machine (SVM), random forest (RF), extreme gradient boosting (XGBoost), light gradient boosting machine (LightGBM) and ANN, were constructed and their predictive performances were also compared. The study shows the potential of machine learning approaches for predicting peak events of CVDs, and provide decision-making for medical resource management.

Method

Overview of the research framework

This study attempted to predict the peak demand days of CVDs admissions by adopting machine learning techniques. The block diagram of the classified prediction process is shown in Fig. 1. In brief, the time series dataset, which was comprised of CVDs admissions, meteorological data and air quality data, was pre-processed. Second, the generalized additive model (GAM) was built to choose the lag day of meteorological conditions and air pollutants for CVDs admission. Then, six machine learning algorithms, including LR, SVM, ANN, RF, XGBoost and LightGBM, were applied to construct the predictive models, and the models' parameters were optimized with 10-fold cross validation. After that, the predictive models were validated, then the performances of these models were compared. Finally, we predicted the peak demand days of CVDs admissions based on the optimal machine learning model.

The details are discussed in the following sub-sections.

Data collection and preprocessing

Hospital admissions data

Data for the daily number of hospital admissions for patients with CVDs who lived in urban areas of Chengdu was obtained from the Health Information Center of Sichuan Province. This data contains aggregate numbers of CVDs admissions in all the tertiary and secondary hospitals of Chengdu each day with primary diagnosis of CVDs (International Classification of Diseases, 10th Revision codes: I00-I99) from 1 January 2015 to 31 December 2017, which is 1096 days of continuous data.

Additionally, we focused on the peak demand of CVDs admissions, and the binary variable was generated from the daily number of CVDs admissions. In the situation of absence of a known threshold for daily CVDs admissions, the peak demand was defined on the basis of 85th percentile threshold (304 hospital admissions per day) by reference to the previous studies [31, 33]. Specifically, the days on which the daily number of CVDs admissions were equal to or above the 85th percentile threshold were defined as peak demand days. Thus, the binary variable of CVDs admissions is highly imbalanced with 931 samples of non-peak demand and 165 samples of peak demand. This binary variable of CVDs admissions was used as the primary dependent variable in the analysis.

Meteorological data and air quality data

Meteorological data, including temperature, relative humidity and rainfall, were derived from the Chengdu Meteorological Monitoring Database (http://data.cma.cn/).

Hourly data of air pollutants including $PM_{2.5}$ (particulate matter with aerodynamic diameter ≤ 2.5 μm), PM_{10} (particulate matter with aerodynamic diameter ≤ 10 μm), SO_2, NO_2, CO and O_3 were obtained from the China National Environmental Monitoring Center (http://www.cnemc.cn/), which provides real-time monitoring of hourly concentrations of air pollutants to the general public. We averaged the 24-h mean concentrations for $PM_{2.5}$, PM_{10}, SO_2, NO_2, CO and calculated maximum 8-h moving average concentrations for O_3 from the air quality monitoring stations interspersed among the urban areas of Chengdu. Concentrations of particulate matter with an aerodynamic diameter between 2.5 and 10 μm (PM_C) were calculated by subtracting daily average concentrations of $PM_{2.5}$ from PM_{10} [9, 34].

Data preprocessing

Data for the daily number of hospital admissions for CVDs, meteorological data and air quality data were collected from different data sources. We merged these three datasets to form a time series dataset by date (i.e. 1 January 2015 to 31 December 2017). The time series features were extracted from date, including year, month (month of year), day (day of month), holiday (public holidays) and DOW (day of week).

During the study period, the percentages of missing values from the monitoring stations were 1.28% (14/1096) for meteorological conditions, and 3.19% (35/1096) for air pollutants. The linear interpolation which has acceptable performance and reliability was used to fill in the missing values of meteorological conditions and air pollutants [35, 36].

Feature extraction

As illustrated in the above section, the features for predicting the peak demand days of CVDs admissions included time series features, meteorological conditions features and air pollutants features. Accumulating epidemiological studies have suggested that the effect of meteorological conditions and air pollutants on CVDs admissions is delayed, and the lag effect is related with regional environment [8, 12, 37]. Hence, we employed an over-dispersed GAM which allowed the quasi-Poisson distribution to analyze the lag effects of daily meteorological conditions and air pollutants on CVDs admissions, and chose the lag day based on the minimum Generalized Cross-Validation (GCV) values which measure models fit [5, 34]. The lag effects of single day lags (from lag0 to lag6) and cumulative day lags (from lag01 to lag06) were taken into consideration. The penalized spline approaches were applied to control for potential confounding of long-term trends, seasonality and meteorological effects [38]. Moreover, dummy variables of holiday and DOW were controlled.

The results demonstrated that temperature, relative humidity, rainfall, $PM_{2.5}$, PM_{10}, PM_C, SO_2, NO_2, CO and O_3 were associated with CVDs admissions, with the minimum GCV values at lag04, lag06, lag06, lag3, lag3, lag3, lag0, lag0, lag0 and lag6, respectively.

Finally, the independent variables for forecasting the peak demand days of CVDs admissions included fifteen features, which are shown in Table 1.

Table 1
The features for prediction

Feature category	Features	Description
time series features	year	year of the date of hospital admission
	month	month of year
	day	day of month
	holiday	public holidays
	DOW	day of week
meteorological conditions features	Tem_lag04	mean temperature for the moving average of current day and previous four days (lag04)
	RH_lag06	relative humidity for the moving average of current day and previous six days (lag06)
	Rain_lag06	rainfall for the moving average of current day and previous six days (lag06)
air pollutants features	PM2.5_lag3	$PM_{2.5}$ at the previous three days (lag3)
	PM10_lag3	PM_{10} at the previous three days (lag3)
	PMC_lag3	PM_C at the previous three days (lag3)
	SO2_lag0	SO_2 at the current day (lag0)
	NO2_lag0	NO_2 at the current day (lag0)
	CO_lag0	CO at the current day (lag0)
	O3_lag6	O_3 at the previous six days (lag6)

Machine learning methods

In this study, six well-accepted machine learning algorithms, including LR, SVM, ANN, RF, XGBoost and LightGBM, were applied to develop predictive models. These machine learning methods were considered according to their following characteristics.

LR is a common and basic algorithm, which is widely used in disease risk prediction and epidemiology [39]. LR is one of the most important basic model and often uses as a baseline comparison method among machine learning studies. It has advantages in the high interpretable of results, simple implementation and low computational cost.

SVM is a discriminative classification technique, which has been widely applied in the medical diagnostics and other fields, especially with small sample sets [40]. SVM constructs the optimal hyperplane or set of hyperplanes in a high dimensional space to divide the data to different classes with the largest separation between the classes. SVM has four kernel functions (i.e. linear function, polynomial function, radial based function, and sigmoid function) which are used to maximize margins between hyperplanes.

ANN, which is inspired by biological neural networks, has a remarkable ability to determine the meaning and rules of complicated data [41, 42]. ANN is composed of three-layer neurons: input layer, hidden layer and output layer. The input layer transmits the original data to one or more hidden layers, and the output layer accepts the result of the hidden layers. The transmitting procedure processes the data through weighted connections and activation functions. ANN can automatically learn from the multiple samples of training data until each input matches to output correctly, therefore achieving the best prediction.

RF is an ensemble algorithm that is composed of multiple decision trees [43]. RF applies bootstrap algorithm to extract multiple samples from the training set randomly, and trains the samples with the weak classifier (i.e. decision tree). The final result of RF is determined by the majority of votes over all decision trees, thereby improving the predictive accuracy and preventing the model from over-fitting.

XGBoost is a distributed gradient boosting algorithm and has gained wide popularity and attention in machine learning competitions [44, 45]. XGBoost chooses classification and regression trees (CART) as a weak classifier, which facilitates efficient optimization algorithms. XGBoost achieves lower variance for adding an L2 regularization term of leaf weights. XGBoost uses the second-order Taylor series as the cost function to retain more information about the target function, thereby improving the predictive accuracy.

LightGBM is a distributed and high-performance gradient lifting framework based on decision tree algorithm designed for fast computational time, especially with very large data sets [46]. It utilizes two novel techniques: gradient-based one-side sampling and exclusive feature bundling, which is used to deal with the huge number of data samples and massive amount of features, respectively [47].

All above mentioned models were trained and tested on a partitioned 80/20 percentage split of the dataset by stratified random sampling. Simultaneously, in situations where there was imbalanced class data combined with unequal error costs, these models' performance metrics are not representative of reasonable performances. Therefore, it is necessary to balance the dataset to get true performance values for the classifier. We adjusted weights inversely proportional to class frequencies in the input data when training the machine learning models.

The parameters of these six predictive models were determined by grid search and 10-fold cross-validation in training dataset. To be specific, we partitioned the training dataset into ten equal-size pieces, and we utilized the grid search with nine pieces to tune the parameters, while the remaining piece was used as the validation set. We repeated this process for ten times. The best parameters for predictive models were obtained with the best score, which itself was obtained by averaging the process of repetition mentioned in the previous sentence. Table 2 shows the values of parameters in each model.

Table 2
Summary of parameter values in each model

Models	Parameters	Values	Parameters Mean
LR	penalty	L1	penalty function
SVM	kernel	linear	kernel function
	C	5	penalty parameter of the error term
ANN	kernel_initializer	uniform	kernel initializer function
	activation1	relu	activation of hidden layer
	activation2	sigmoid	activation of output layer
	optimizer	Adam	training optimization algorithm
	epochs	300	number of times shown to the network
	batch_size	20	batch size
	dropout	0.0	dropout rate
RF	n_estimators	695	number of iterations
	max_depth	4	maximum depth of variable interactions
	max_features	7	number of features for the best split
XGBoost	learning_rate	0.1	learning rate
	n_estimators	100	number of iterations
	eta	0.01	control of learning rate
	max_depth	3	maximum depth of variable interactions
	gamma	0.6	minimum loss reduction required to make a further partition on the tree' leaf node
	subsample	0.7	subsample ratio
	colsample_bytree	0.6	subsample ratio of columns when constructing each tree
	min_child_weight	2	sum of the minimum weights that leaf node need to observe
LightGBM	learning_rate	0.1	learning rate
	n_estimators	100	number of iterations
	max_depth	8	maximum depth of variable interactions
	num_leaves	10	number of leaves in each tree

Models	Parameters	Values	Parameters Mean
	bagging_fraction	0.7	percentage of sampling used each iteration
	feature_fraction	0.9	ratio of features to build the tree in each iteration
	min_data_in_leaf	5	minimum number of records a leaf
	min_split_gain	0.0	smallest gain of the split

Model Assessment

We calculated the AUC from receiver operating characteristic (ROC) analysis to evaluate the predictive utilities of the models, and the AUC of the six machine learning models was compared based on DeLong method (p-value < 0.05 was deemed to indicate statistical significance) [48]. Meanwhile, logarithmic loss function (log-loss) was applied to quantify the accuracy of the classifier by punishing the wrong classification. Furthermore, the evaluation indicators of the confusion matrix, including accuracy, sensitivity, specificity and F1 score, were used to analyze the relationship between the actual values and the predicted values for the peak demand of CVDs admissions.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (1)$$

$$Sensitivity = \frac{TP}{TP + FN} \quad (2)$$

$$Specificity = \frac{TN}{TN + FP} \quad (3)$$

$$F1score = \frac{2 * Precision * Recall}{Precision + Recall} \quad (4)$$

where, TP = True Positive, FP = False Positive, TN = True Negative, FN = False Negative; $Precision = \frac{TP}{TP + FP}$, $Recall = \frac{TP}{TP + FN}$.

Result

Descriptive Statistics

The statistical information of daily CVDs hospital admissions, meteorological conditions and air pollutants concentrations is summarized in Table 3. During the study period, the average of daily hospital admissions for CVDs was 208 inpatients, the minimum value was 33 and the maximum value was 476. The daily average level of temperature, relative humidity and rainfall was 17.0°C, 80.4% and 2.6 mm, respectively. The daily average concentrations were 60.3 μg/m³ for $PM_{2.5}$, 99.3 μg/m³ for PM_{10}, 39.0 μg/m³ for PM_C, 13.9 μg/m³ for SO_2, 55.0 μg/m³ for NO_2, 96.0 μg/m³ for O_3 and 1.1 mg/m³ for CO.

Table 3
Summary statistics of daily CVDs hospital admissions, meteorological conditions and air pollutants concentrations in Chengdu, 2015–2017

	Mean	Standard Deviation	Minimum	Median	Maximum
CVDs hospital admissions (n)	208	90	33	206	476
Meteorological Conditions					
Temperature (°C)	17.0	7.2	-1.1	17.8	30.2
Relative Humidity (%)	80.4	8.8	43.0	80.8	98.3
Rainfall (mm)	2.6	8.7	0.0	0.0	122.0
Air Pollutants Concentrations					
$PM_{2.5}$ ($\mu g/m^3$)	60.3	42.4	6.1	48.4	324.5
PM_{10} ($\mu g/m^3$)	99.3	64.7	14.3	79.8	492.5
PM_C ($\mu g/m^3$)	39.0	25.8	4.8	31.6	238.2
SO_2 ($\mu g/m^3$)	13.9	5.8	3.9	12.7	37.9
NO_2 ($\mu g/m^3$)	55.0	17.3	15.7	53.0	130.4
O_3 ($\mu g/m^3$)	96.0	54.6	5.6	85.3	290.4
CO (mg/m^3)	1.1	0.4	0.4	1.0	2.8
CVDs (Cardiovascular diseases, ICD-10 codes: I00-I99)					

Evaluation and Comparison of the Predictive Models

Based on the above-mentioned features in Table 1, we constructed six machine learning models to predict the peak demand days for CVDs admissions. Using the optimal parameters for each model, the predictive models were corroborated via a validation set which was divided from the training dataset by 10-fold cross-validation. The box plot of AUC for each model with 10-fold cross-validation in training dataset is shown in Fig. 2.

The AUC for LR, SVM, ANN, RF, XGBoost and LightGBM was 0.817 (95% confidence interval (CI): 0.795–0.839), 0.814 (95% CI: 0.792–0.836), 0.844 (95% CI: 0.814–0.875), 0.929 (95% CI: 0.906–0.951), 0.945 (95% CI: 0.922–0.967) and 0.9454 (95% CI: 0.921–0.967), respectively. The XGBoost model achieved the best AUC, and the performance was significantly better than LR ($p < 0.001$), SVM ($p < 0.001$) and ANN ($p < 0.001$), but did not differ significantly from RF ($p = 0.264$) and LightGBM ($p = 0.933$).

Based on the validation result for the training dataset, we predicted the peak demand days for CVDs admissions in an independent testing dataset. The ROC curve for the predictive models in testing dataset is shown in Fig. 3. The AUC of LR, SVM, ANN, RF, XGBoost and LightGBM was 0.842 (95% CI: 0.783–0.901), 0.834 (95% CI: 0.774–0.894), 0.890 (95% CI: 0.836–0.944), 0.926 (95% CI: 0.879–0.974), 0.930 (95% CI: 0.878–0.982) and 0.940 (95% CI: 0.900–0.980), respectively. The LightGBM model had the highest AUC value among all these predictive models, and the performance was significantly better than LR ($p < 0.001$), SVM ($p < 0.001$), ANN ($p = 0.03$), but did not differ significantly from RF ($p = 0.222$) and XGBoost ($p = 0.489$).

Furthermore, we used log-loss, accuracy, sensitivity, specificity and F1 score to compare the performances of these six machine learning models in independent testing dataset (Table 4). The LightGBM model exhibited the best AUC (0.940), log-loss (0.218), accuracy (0.913), specificity (0.941) and F1 score (0.725) in testing dataset, and the RF model had the best sensitivity (0.909). Thus, the LightGBM model achieved the best performance among the six machine learning models.

Table 4
The evaluation indicators of machine learning models in testing dataset

Models	AUC	log-loss	Accuracy	Sensitivity	Specificity	F1 score
LR	0.842 (95% CI: 0.783–0.901)	0.513	0.766	0.848	0.751	0.523
SVM	0.834 (95% CI: 0.774–0.894)	0.344	0.748	0.879	0.724	0.513
ANN	0.890 (95% CI: 0.836–0.944)	0.296	0.858	0.333	0.951	0.415
RF	0.926 (95% CI: 0.879–0.974)	0.358	0.862	0.909	0.854	0.667
XGBoost	0.930 (95% CI: 0.878–0.982)	0.277	0.876	0.818	0.886	0.667
LightGBM*	0.940 (95% CI: 0.900–0.980)	0.218	0.913	0.758	0.941	0.725

font bold: the optimal values; *: the optimal model. LR: logistic regression; SVM: support vector machine; ANN: artificial neural network; RF: random forest; XGBoost: extreme gradient boosting; LightGBM: light gradient boosting machine.

The identification of feature importance

As illustrated in the above section, the LightGBM model achieved the best performance, which can offer the most powerful predictors for predicting the peak demand days of CVDs admissions. The identification of feature importance based on LightGBM is shown in Fig. 4. The contribution rate of time series features, meteorological conditions and air pollutants for predicting the peak demand days of

CVDs admissions was 25%, 32% and 43%, respectively. Among the meteorological conditions features, the top-rank 2 features was Tem_lag04 and RH_lag06, respectively. Similarly, the top-rank 2 features among the air pollutants was NO2_lag0 and SO2_lag0, respectively.

Discussion

In the current study, six machine learning models were developed to predict the peak demand days for CVDs admissions and find the optimal model. To the best of our knowledge, this is the first study to construct and compare various machine learning models in terms of predicting the peak events of CVDs using meteorological data, air quality data and hospital admission data.

Our study found that the ensemble learning models, including LightGBM, RF and XGBoost, outperformed ANN, SVM and LR, achieving overall accuracies of > 0.86 and AUCs of > 0.92. This implies that the ensemble learning models have better generalization capabilities compared to other models on predictions of the peak demand days of CVDs admissions. The LightGBM exhibited the best performance among the ensemble learning models. Compared with ANN, SVM and LR, the AUC of LightGBM significantly improved by 5.65%, 12.66% and 11.61% respectively. The results of our study indicate that ensemble learning models are well suited for the prediction of peak demand for healthcare services.

The lag patterns of meteorological conditions and air pollutants have been well-documented in epidemiological studies [8, 12, 16] and suggested that the lag effect of environmental exposure has regional differences. However, to date, very few machine learning-based studies have analyzed the lag effect of environmental exposure when predicting the peak demand for healthcare services. Krishan et al. [31] applied representative lags to predictors based on the results from other studies to forecast the peak demand days of emergency department visits, but did not combined with the actual situation of the study area. In this study, we utilized GAM to analyze the lag effect of meteorological conditions and air pollutants on CVDs admissions in our study areas. It is useful to detect and provide early warning signals of likely peak demand in future.

Environmental exposure, such as ambient air pollution and extreme temperature, is an important but underappreciated risk factor contributing to the development and severity of CVDs [4]. Accumulating evidence from epidemiological studies has linked environmental exposure to increased risk of CVDs morbidity [5–12]. However, evidence on the effect of complex mixture of environmental exposure on CVDs morbidity is still limited. Machine learning techniques provide opportunity for developing algorithms that classify individuals with complex interaction factors. In our study, the contribution of the special ambient air pollutants and climatic characteristics of the area to the peak demand days of CVDs admissions was successfully modeled. The identification of feature importance based on the optimal model showed that among the environmental exposure features, the top-rank 4 feature was Tem_lag04, RH_lag06, NO2_lag0 and SO2_lag0, respectively, and the contribution rate of meteorological conditions and air pollutants for the prediction was 32% and 43%, respectively. These results suggested that environmental exposure is important predictor and should be brought to the attention.

This study has several strengths. First, we applied six well-accepted machine learning algorithms to construct predictive models, which shows certain representativeness, where LR represents the basic machine learning model, SVM and ANN are widely used in prediction, and RF, XGBoost and LightGBM are ensemble learning models. In addition, we took account of the effect of complex mixture of environmental exposure on CVDs admissions and analyzed the lag effect of meteorological conditions and air pollutants on hospitalizations for CVDs. Finally, we studied all the hospital admissions of patients with CVDs who were exposed to the environment in the whole region.

Our study also has some limitations that need to be addressed. First, we considered only two well-studied environmental exposures: meteorological conditions and ambient air pollutants, but some other environmental factors, such as exposure to the metals arsenic, cadmium and lead, also play important roles in the development and severity of CVDs [4]. Second, we just constructed the classification models to predict the peak demand days of CVDs admissions. Further study is required to forecast the number of admissions for CVDs accurately based on regression models. Third, the current model is designed for non-communicable disease such as CVDs which is associated with environmental exposure and it might not be suitable for forecasting the peak events of infectious diseases.

Conclusion

This study used machine learning approaches to forecast the peak demand days for CVDs admissions based on hospital admissions data, air quality data and meteorological data. The results revealed that ensemble learning models especially the LightGBM model can accurately predict the peak events of CVDs admissions. Meanwhile, the identification of feature importance based on LightGBM indicated that meteorological conditions and air pollutants made great contributions to the prediction. These findings show that machine learning approaches have potential in predicting the peak events of CVDs, and the predictive capacity of ensemble learning model makes it a valid tool supporting decisions regarding medical resource management.

Declarations

Ethics approval and consent to participate

This study was approved by the Health Information Center of Sichuan Province. Informed consent was waived because this research did not involve individual data.

Consent for publication

Not applicable. The study does not include details relating to an individual person.

Availability of data and materials

The meteorological and air quality datasets analysed during the current study are available in http://data.cma.cn and http://www.cnemc.cn/. Daily data of hospital admissions for CVDs are available from the Health Information Center of Sichuan Province but restrictions are applied to the availability of these data, which were used under license for the current study, and so are not publicly available. Daily number of hospital admissions for patients with CVDs are however available from authors upon reasonable request and with permission of Health Information Center of Sichuan Province, China.

Competing interests

The authors declare that they have no competing interests.

Funding

This research was supported by the National Natural Science Foundation of China (No. 71661167005), the Key Research and Development Program of Sichuan Province (No. 2018SZ0114, No. 2019YFS0271) and the 1·3·5 Project for Disciplines of Excellence–Clinical Research Incubation Project, West China Hospital, Sichuan University (Grant No. 2018HXFH023, ZYJC18013).

Authors' Contributions

HQ proposed and designed the study. HQ, LL and ZQS performed the experiments and analyzed the data. LYW and LZ collected the data and performed the statistical analyses. HQ and LL wrote the manuscript. ZQS and YCC revised the manuscript. All authors have read and approved the final manuscript.

Acknowledgements

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

References

1. WHO: **https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds) (accessed on 1 September 2019)**.
2. **Institute for Health Metrics and Evaluation (IHME)**. *GBD Compare | Viz Hub VizHub http://vizhub healthdataorg/gbd-compare* (2018).
3. **Report on cardiovascular diseases in China (2017)**. *National center for cardiovascular diseases, China*.
4. Cosselman KE, Navas-Acien A, Kaufman JD: **Environmental factors in cardiovascular disease**. *Nature reviews Cardiology* 2015, **12**(11):627-642.
5. Zhu X, Qiu H, Wang L, Duan Z, Yu H, Deng R, Zhang Y, Zhou L: **Risks of hospital admissions from a spectrum of causes associated with particulate matter pollution**. *Science of The Total Environment* 2019, **656**:90-100.

6. Hui L, Yaohua T, Xiao X, Juan J, Jing S, Yaying C, Chao H, Man L, Yonghua H: **Ambient Particulate Matter Concentrations and Hospital Admissions in 26 of China's Largest Cities: A Case-Crossover Study**. *Epidemiology* 2018, **29**(5):649-657.
7. Tatiane F, Maria F, Clarice dF, Felipe N, Washington J, Nelson G: **Effects of Particulate Matter and Its Chemical Constituents on Elderly Hospital Admissions Due to Circulatory and Respiratory Diseases**. *International Journal of Environmental Research & Public Health* 2016, **13**(10):947-957.
8. Soleimani Z, Darvishi Boloorani A, Khalifeh R, Griffin DW, Mesdaghinia A: **Short-term effects of ambient air pollution and cardiovascular events in Shiraz, Iran, 2009 to 2015**. *Environmental science and pollution research international* 2019, **26**(7):6359-6367.
9. Chen M, Qiu H, Wang L, Zhou L, Zhao F: **Attributable risk of cardiovascular hospital admissions due to coarse particulate pollution: A multi-city time-series analysis in southwestern China**. *Atmospheric Environment* 2019, **218,117014**.
10. Zhao Q, Zhao Y, Li S: **Impact of ambient temperature on clinical visits for cardio-respiratory diseases in rural villages in northwest China**. *Science of the Total Environment* 2018, **612**:379-385.
11. Ha S, Nguyen K, Liu D, Mannisto T, Nobles C, Sherman S, Mendola P: **Ambient temperature and risk of cardiovascular events at labor and delivery: A case-crossover study**. *Environmental Research* 2017, **159**:622-628.
12. Phung D, Thai PK, Guo Y, Morawska L, Rutherford S, Chu C: **Ambient temperature and risk of cardiovascular hospitalization: An updated systematic review and meta-analysis**. *Science of the Total Environment* 2016, **550**:1084-1102.
13. Liu H, Tian Y, Song J, Cao Y, Hu Y: **Effect of Ambient Air Pollution on Hospitalization for Heart Failure in 26 of China's Largest Cities**. *American Journal of Cardiology* 2017, **121**(5):628-633.
14. Tian Y, Liu H, Si Y, Cao Y, Song J, Li M, Wu Y, Wang X, Xiang X, Juan J: **Association between temperature variability and daily hospital admissions for cause-specific cardiovascular disease in urban China: A national time-series study**. *PLoS Medicine* 2019, **16**(1):e1002738.
15. Hsu WH, Hwang S-A, Kinney PL, Lin S: **Seasonal and temperature modifications of the association between fine particulate air pollution and cardiovascular hospitalization in New York state**. *Science of the Total Environment* 2017, **578**:626-632.
16. Ma Y, Zhao Y, Yang S, Zhou J, Yang D: **Short-term effects of ambient air pollution on emergency room admissions due to cardiovascular causes in Beijing, China**. *Environmental Pollution* 2017, **230**:974-980.
17. Vahedian M, Khanjani N, Mirzaee M, Koolivand A: **Ambient air pollution and daily hospital admissions for cardiovascular diseases in Arak, Iran**. *Arya Atherosclerosis* 2017, **13**(3):117-134.
18. Juang WC, Huang S-J, Huang F-D, Cheng P-W, Wann S-R: **Application of time series analysis in modelling and forecasting emergency department visits in a medical centre in Southern Taiwan**. *Bmj Open* 2017, **7**(11):e018628.
19. Jilani T, Housley G, Figueredo G, Tang PS, Hatton J, Shaw D: **Short and Long term predictions of Hospital emergency department attendances**. *International journal of medical informatics* 2019,

129:167-174.

20. Zhou L, Ping Z, Dongdong W, Cheng C, Hao H: **Time series model for forecasting the number of new admission inpatients**. *Bmc Medical Informatics & Decision Making* 2018, **18**(1):39-49.
21. Zhu T, Luo L, Zhang X, Shi Y, Shen W: **Time Series Approaches for Forecasting the Number of Hospital Daily Discharged Inpatients**. *IEEE Journal of Biomedical & Health Informatics* 2017, **21**:515-526.
22. Esteva A, Kuprel B, Novoa RA, Ko J, Swetter SM, Blau HM, Thrun S: **Dermatologist-level classification of skin cancer with deep neural networks**. *Nature* 2017, **542**(7639):115-118.
23. Gunčar G, Kukar M, Notar M, Brvar M, Černelč P, Notar M, Notar M: **An application of machine learning to haematological diagnosis**. *Sci Rep* 2018, **8**(1):411.
24. Qiu H, Yu HY, Wang LY, Yao Q, Wu SN, Yin C, Fu B, Zhu XJ, Zhang YL, Xing Y *et al*: **Electronic Health Record Driven Prediction for Gestational Diabetes Mellitus in Early Pregnancy**. *Sci Rep* 2017, **7**(1):16417.
25. Lim J, Kim J, Cheon S: **A Deep Neural Network-Based Method for Early Detection of Osteoarthritis Using Statistical Data**. *International journal of environmental research and public health* 2019, **16**(7):1281.
26. 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 Quality Atmosphere & Health* 2011, **4**(3-4):263-272.
27. Shakerkhatibi M, Dianat I, Jafarabadi MA, Azak R, Kousha A: **Air pollution and hospital admissions for cardiorespiratory diseases in Iran: artificial neural network versus conditional logistic regression**. *International Journal of Environmental Science & Technology* 2015, **12**(11):3433-3442.
28. Moustris KP, Larissi IK, Nastos PT, Paliatsos AG: **Seven-days-ahead forecasting of childhood asthma admissions using artificial neural networks in Athens, Greece**. *Int J Environ Health Res* 2012, **22**(2):93-104.
29. Polezer G, Tadano YS, Siqueira HV, Godoi AFL, Yamamoto CI, de André PA, Pauliquevis T, Andrade MdF, Oliveira A, Saldiva PHN: **Assessing the impact of PM 2.5 on respiratory disease using artificial neural networks**. *Environmental Pollution* 2018, **235**:394-403.
30. Kassomenos P, Papaloukas C, Petrakis M, Karakitsios S: **Assessment and prediction of short term hospital admissions: the case of Athens, Greece**. *Atmospheric Environment* 2008, **42**(30):7078-7086.
31. Khatri KL, Tamil LS: **Early Detection of Peak Demand Days of Chronic Respiratory Diseases Emergency Department Visits Using Artificial Neural Networks**. *IEEE Journal of Biomedical & Health Informatics* 2017(99):285-290.
32. Wu C-C, Yeh W-C, Hsu W-D, Islam MM, Nguyen PA, Poly TN, Wang Y-C, Yang H-C, Li Y-C: **Prediction of fatty liver disease using machine learning algorithms**. *Computer Methods and Programs in Biomedicine* 2019, **170**:23-29.
33. Soyiri IN, Reidpath DD, Sarran C: **Forecasting peak asthma admissions in London: an application of quantile regression models**. *Int J Biometeorol* 2013, **57**(4):569-578.

34. Qiu H, Zhu X, Wang L, Pan J, Pu X, Zeng X, Zhang L, Peng Z, Zhou L: **Attributable risk of hospital admissions for overall and specific mental disorders due to particulate matter pollution: A time-series study in Chengdu, China**. *Environmental Research* 2019, **170**:230-237.
35. Junninen H, Niska H, Tuppurainen K, Ruuskanen J, Kolehmainen M: **Methods for imputation of missing values in air quality data sets**. *Atmospheric Environment* 2004, **38**(18):2895-2907.
36. Qiu H, Tan K, Long F, Wang L, Yu H, Deng R, Long H, Zhang Y, Pan J: **The Burden of COPD Morbidity Attributable to the Interaction between Ambient Air Pollution and Temperature in Chengdu, China**. *International Journal of Environmental Research & Public Health*, **15**(3).
37. Ma Y, Zhang H, Zhao Y, Zhou J, Yang S, Zheng X, Wang S: **Short-term effects of air pollution on daily hospital admissions for cardiovascular diseases in western China**. *Environmental Science & Pollution Research* 2017, **24**(16):14071-14079.
38. Chen G, Zhang Y, Zhang W, Li S, Guo Y: **Attributable risks of emergency hospital visits due to air pollutants in China: A multi-city study**. *Environmental Pollution* 2017, **228**:43-49.
39. Dreiseitla S, Ohno-Machadob L: **Logistic regression and artificial neural network classification models: a methodology review**. *Journal of Biomedical Informatics* 2002, **35**(5-6):352-359.
40. Cortes C, Vapnik VN: **Support Vector Networks**. *Machine Learning* 1995, **20**(3):273-297.
41. Marcel VG, Sander B: **Editorial: Artificial Neural Networks as Models of Neural Information Processing**. *Frontiers in Computational Neuroscience* 2017, **11**:114-.
42. White H: **Learning in Artificial Neural Networks: A Statistical Perspective**. *Neural Computation* 2014, **1**(4):425-464.
43. Breiman L: **Random Forests**. *Machine Learning* 2001, **45**(1):5-32.
44. Chen T, Guestrin C: **XGBoost: A Scalable Tree Boosting System**. In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining: 2016*; 2016.
45. Friedman JH: **Greedy Function Approximation: A Gradient Boosting Machine**. *The Annals of Statistics* 2001, **29**(5):1189-1232.
46. Ke GL, Meng Q, Finley T, Wang TF, Chen W, Ma WD, Ye QW, Liu TY: **LightGBM: A Highly Efficient Gradient Boosting Decision Tree**. *Adv Neur In* 2017, **30**:46-54.
47. Deng L, Pan J, Xu X, Yang W, Liu C, Liu H: **PDRLGB: precise DNA-binding residue prediction using a light gradient boosting machine**. *BMC Bioinformatics* 2018, **19**:136-145.
48. Delong ER, Delong DM, Clarke-Pearson DL: **Comparing the Areas under Two or More Correlated Receiver Operating Characteristic Curves: A Nonparametric Approach**. *Biometrics* 1988, **44**(3):837-845.

Figures

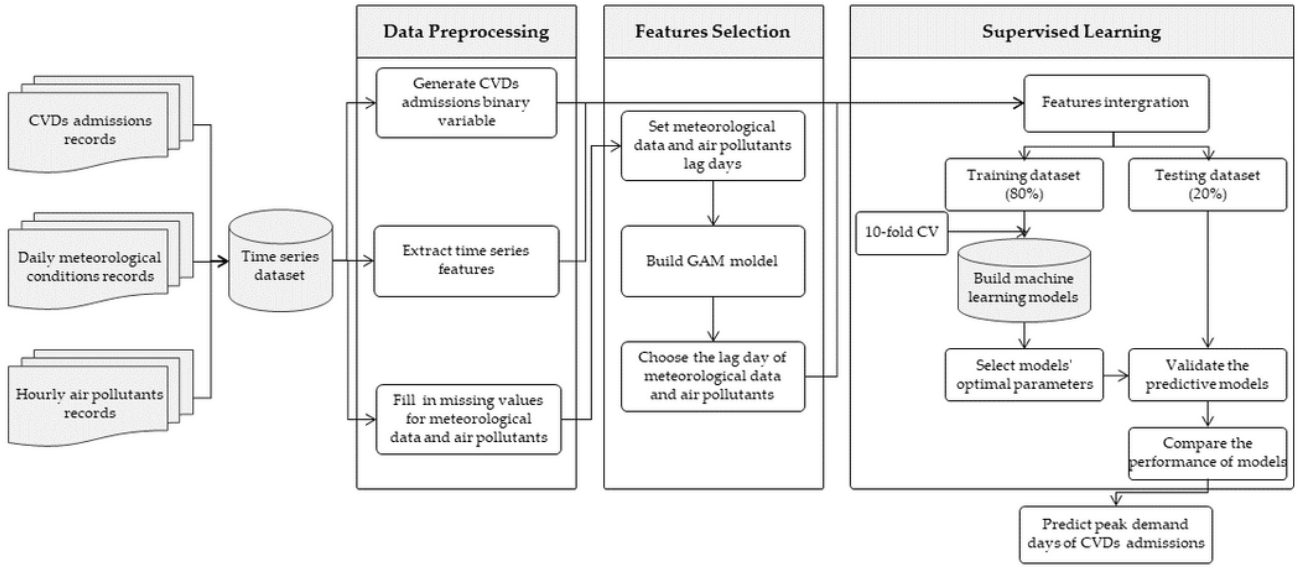

Figure 1

Block Diagram of Classified Prediction Process

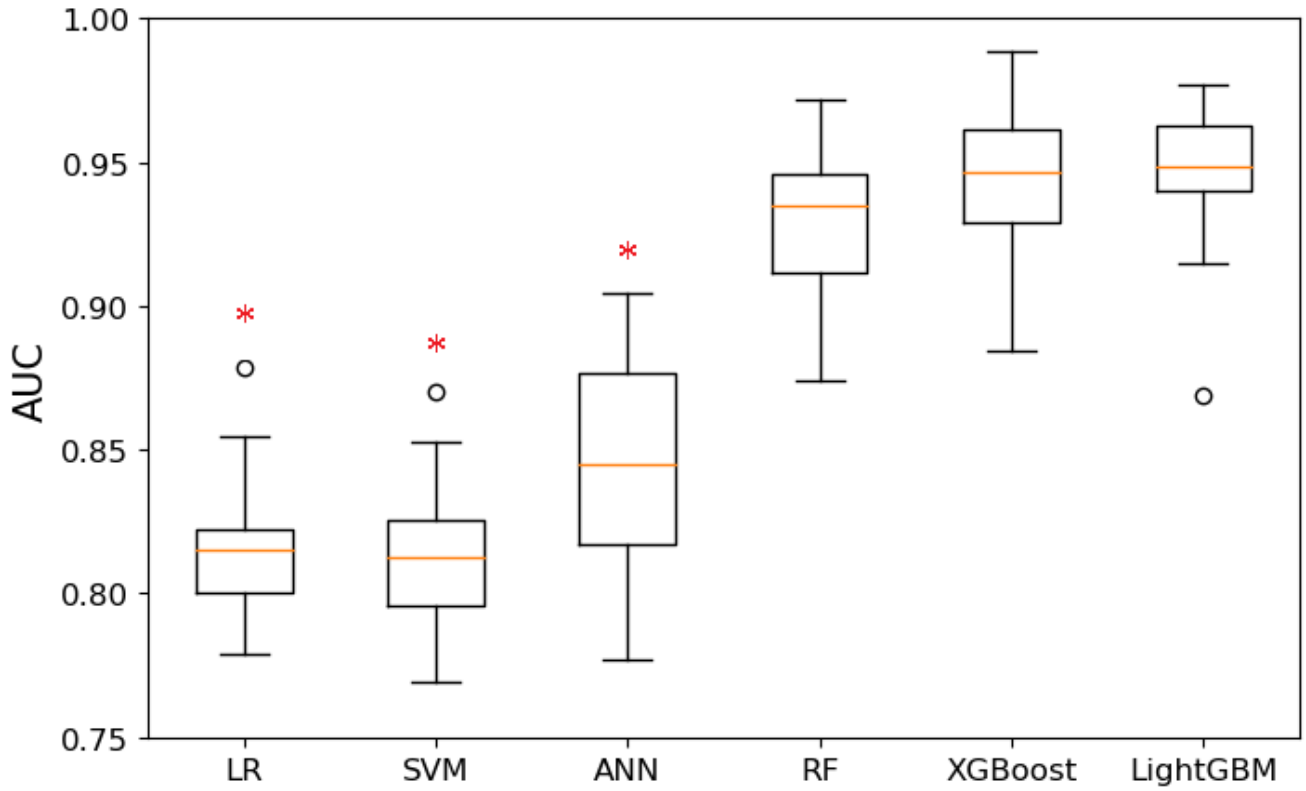

Figure 2

Box plot of AUC for machine learning models with 10-fold cross-validation in training dataset. °: the outliers of box plot; *: the model is significantly different from the XGBoost model. LR: logistic regression; SVM: support vector machine; ANN: artificial neural network; RF: random forest; XGBoost: extreme gradient boosting; LightGBM: light gradient boosting machine.

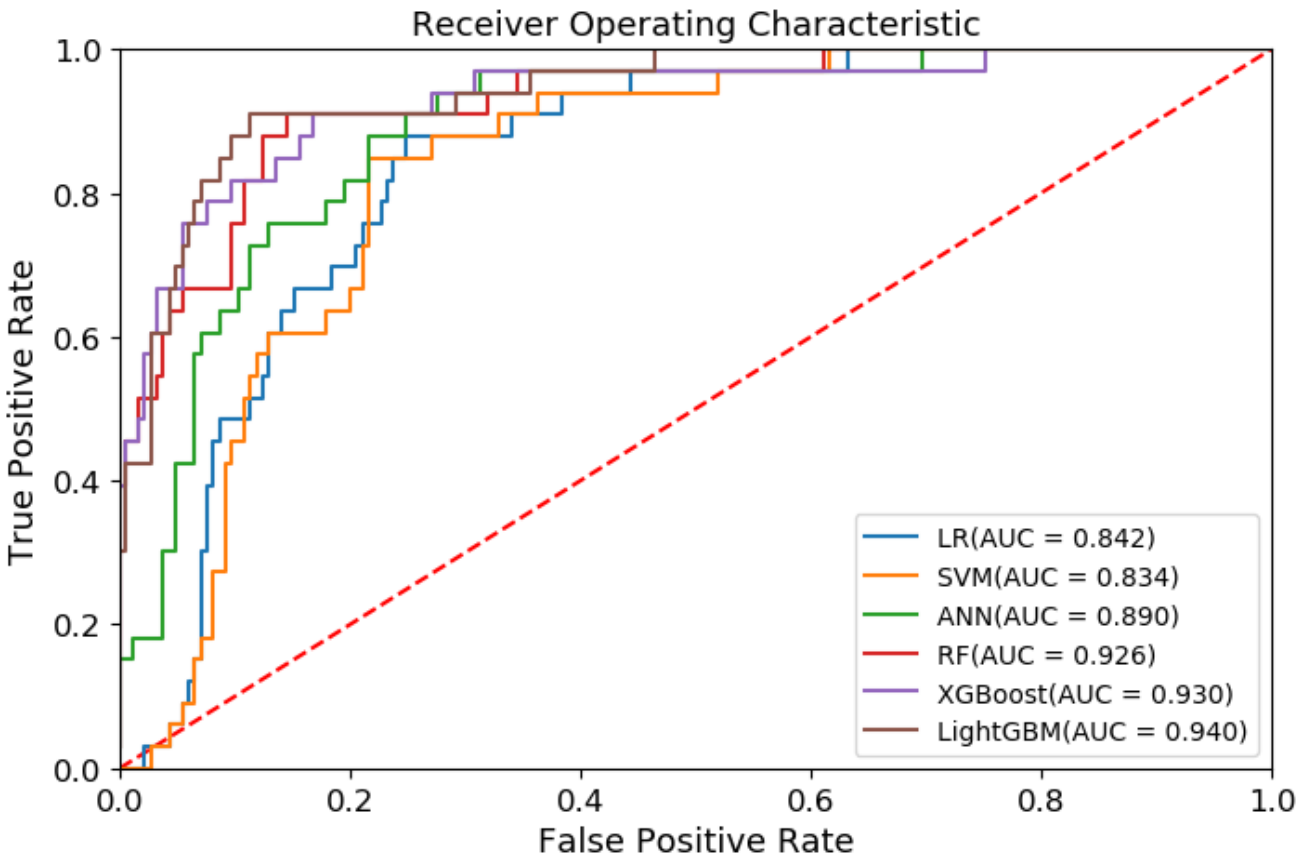

Figure 3

ROC curve of machine learning models in testing dataset. LR: logistic regression; SVM: support vector machine; ANN: artificial neural network; RF: random forest; XGBoost: extreme gradient boosting; LightGBM: light gradient boosting machine.

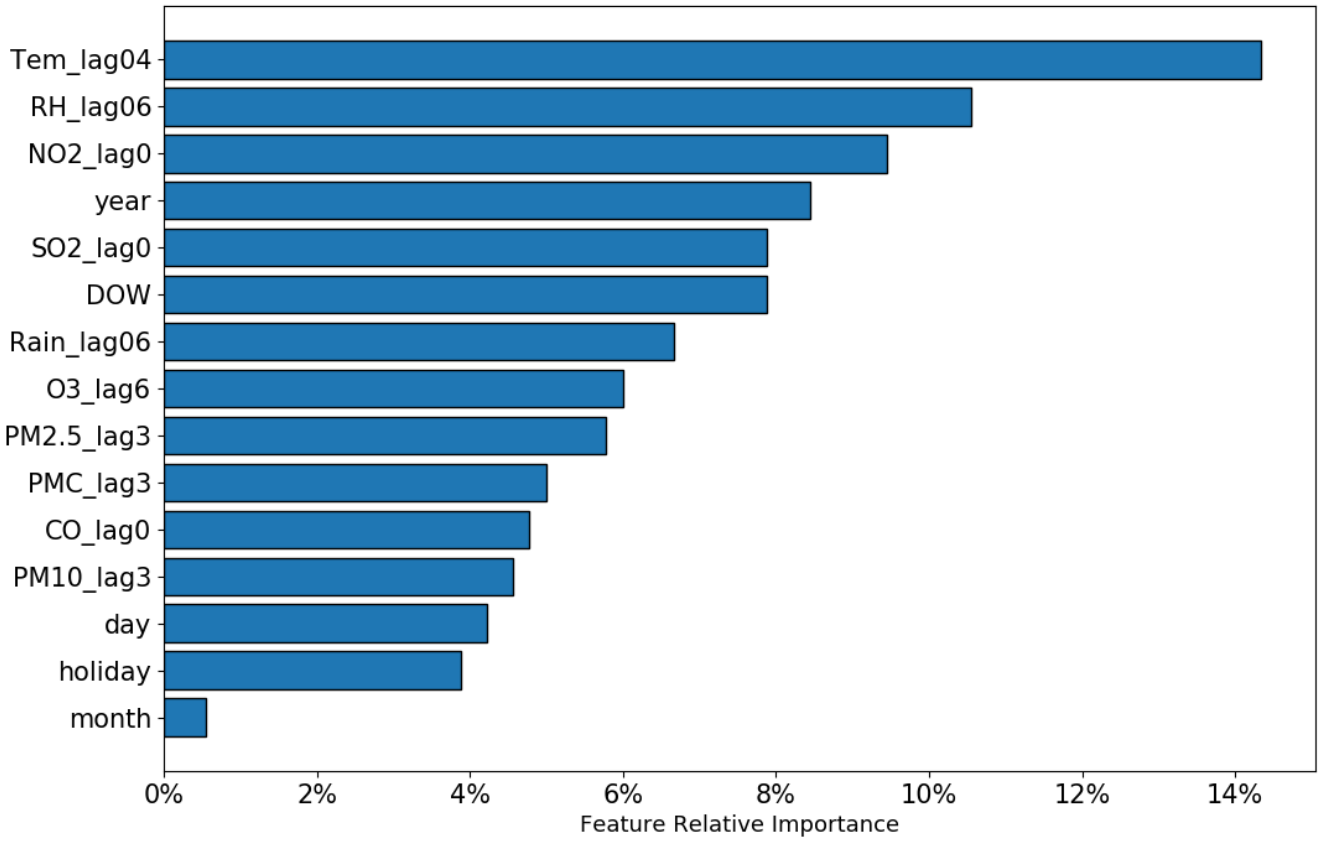

Figure 4

Features importance ranking based on LightGBM model