BMC Medical Informatics and Decision Making

RESEARCH ARTICLE

Open Access

The prediction of asymptomatic carotid atherosclerosis with electronic health records: a comparative study of six machine learning models

Jiaxin Fan[1], Mengying Chen[1], Jian Luo[2], Shusen Yang[3], Jinming Shi[1], Qingling Yao[1], Xiaodong Zhang[1], Shuang Du[1], Huiyang Qu[1], Yuxuan Cheng[1], Shuyin Ma[1], Meijuan Zhang[1], Xi Xu[1], Qian Wang[4] and Shuqin Zhan[1*]

Abstract

Background: Screening carotid B-mode ultrasonography is a frequently used method to detect subjects with carotid atherosclerosis (CAS). Due to the asymptomatic progression of most CAS patients, early identification is challenging for clinicians, and it may trigger ischemic stroke. Recently, machine learning has shown a strong ability to classify data and a potential for prediction in the medical field. The combined use of machine learning and the electronic health records of patients could provide clinicians with a more convenient and precise method to identify asymptomatic CAS.

Methods: Retrospective cohort study using routine clinical data of medical check-up subjects from April 19, 2010 to November 15, 2019. Six machine learning models (logistic regression [LR], random forest [RF], decision tree [DT], eXtreme Gradient Boosting [XGB], Gaussian Naïve Bayes [GNB], and K-Nearest Neighbour [KNN]) were used to predict asymptomatic CAS and compared their predictability in terms of the area under the receiver operating characteristic curve (AUCROC), accuracy (ACC), and F1 score (F1).

Results: Of the 18,441 subjects, 6553 were diagnosed with asymptomatic CAS. Compared to DT (AUCROC 0.628, ACC 65.4%, and F1 52.5%), the other five models improved prediction: KNN + 7.6% (0.704, 68.8%, and 50.9%, respectively), GNB + 12.5% (0.753, 67.0%, and 46.8%, respectively), XGB + 16.0% (0.788, 73.4%, and 55.7%, respectively), RF + 16.6% (0.794, 74.5%, and 56.8%, respectively) and LR + 18.1% (0.809, 74.7%, and 59.9%, respectively). The highest achieving model, LR predicted 1045/1966 cases (sensitivity 53.2%) and 3088/3566 non-cases (specificity 86.6%). A tenfold cross-validation scheme further verified the predictive ability of the LR.

Conclusions: Among machine learning models, LR showed optimal performance in predicting asymptomatic CAS. Our findings set the stage for an early automatic alarming system, allowing a more precise allocation of CAS prevention measures to individuals probably to benefit most.

Keywords: Machine learning, Asymptomatic carotid atherosclerosis, Electronic health records, Prediction

Background

Carotid atherosclerosis (CAS) is a complex disease [1], which reflects cerebral atherosclerosis to a certain extent and can trigger ischemic stroke. The atherosclerotic

*Correspondence: sqzhan@mail.xjtu.edu.cn
[1] Department of Neurology, The Second Affiliated Hospital of Xi'an Jiaotong University, No. 157 West Five Road, Xi'an 710004, Shaanxi, China
Full list of author information is available at the end of the article

process usually originates early in life and the condition remains asymptomatic for several decades. The standardised prevalence of asymptomatic CAS in China is 36.2% [2]. One of the key measures to delay the development of asymptomatic CAS into symptomatic CAS and cerebrovascular events is to identify apparently healthy individuals with risk factors and control them as early as possible [3]. However, the early diagnosis of asymptomatic individuals remains a challenge for clinicians.

Machine learning can effectively configure multimodal data and achieve a precise predictive ability to assess diagnostic and prognostic outcomes [4]. Medicine is undergoing an electronic revolution, more and more electronic medical records are available, laying the cornerstone for personalised medicine mediated by computer technology. Mounting studies [5–8] have shown that machine learning yields satisfactory results in biomedicine. One meta-analysis [9] revealed that the diagnostic performance of deep learning models was comparable to that of healthcare professionals. Driven by market forces and a strong public interest, such machine- learning-based predictive tools require rapid development.

In neurology, machine learning has been increasingly applied in disease diagnosis, treatment, and outcome prediction [10, 11]. To date, no study has applied machine learning algorithms to predict asymptomatic CAS. In this context, we used and compared multiple machine learning models to predict asymptomatic CAS subjects using electronic health records.

Methods

Study design and data collection

Electronic health records of medical check-up subjects were retrospectively extracted from the Department of Health Management at The Second Affiliated Hospital of Xi'an Jiaotong University from April 19, 2010 to November 15, 2019. Inclusion criteria: (1) Data documented in "Rocket Frog" (Beijing, China), an electronic health record system; (2) Age ≥ 18 years old; (3) Lack of symptoms, such as limb weakness, aphasia, transient monocular blindness, dizziness, crooked mouth, dysphagia, and coma; (4) Complete carotid B-mode ultrasonography examination; (5) No missing values; (6) Provision of informed consent (refer to the Ethics approval and consent to participate and Statement sections for more details). Subjects were excluded if they did not meet the above criteria. If the subject had completed more than one check-up, the most recent report was included, as this would be more closely related to the physical condition of people. The sample selection process is summarised in Fig. 1. To ensure data accuracy, data were collected independently by two clinicians. If the data were not identical after assessment using the assert_frame_equal module from pandas, a third clinician would reconfirm the data.

CAS was diagnosed by carotid B-mode ultrasonography [12], which was defined as a carotid intima-media thickness of 1.0 mm or greater, or plaque formation. We used a carotid ultrasound machine (Preirus, Hitachi,

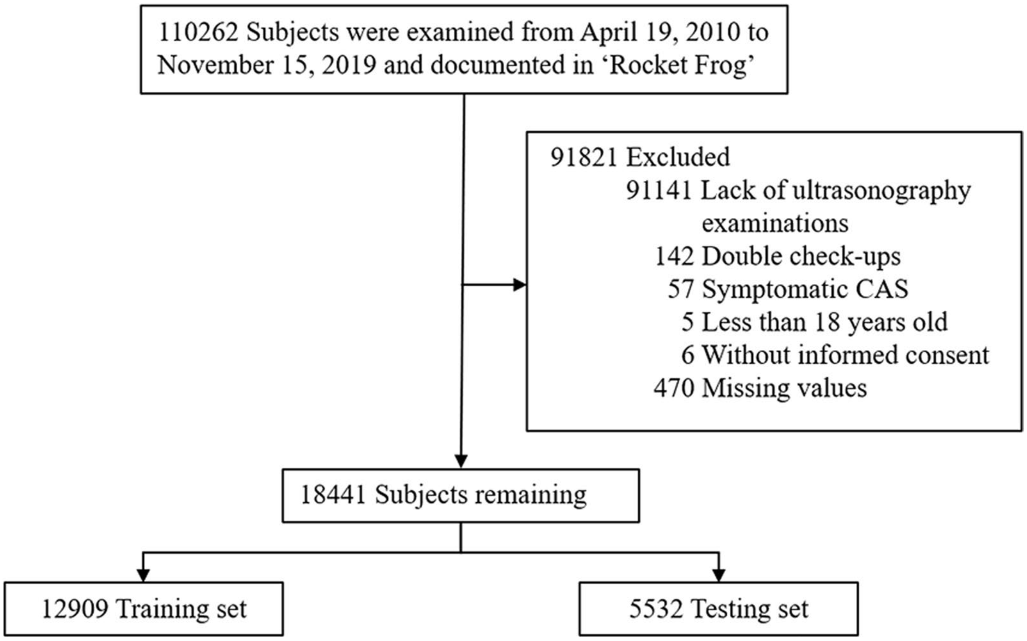
110262 Subjects were examined from April 19, 2010 to November 15, 2019 and documented in 'Rocket Frog'
91821 Excluded
91141 Lack of ultrasonography examinations
142 Double check-ups
57 Symptomatic CAS
5 Less than 18 years old
6 Without informed consent
470 Missing values
18441 Subjects remaining
12909 Training set
5532 Testing set

Fig. 1 Flowchart illustrating sample selection. (CAS, carotid atherosclerosis)

Tokyo, Japan) with a probe (5–10 MHz; L7-3, Tokyo, Japan). To maintain a balance between positive and negative cases, the participants were randomly divided into a training set (70%) and a testing set (30%), as stratified by CAS. Briefly, we first separated the positive and negative cases from the entire dataset. Then, 70% of the positive cases were randomly assigned to the training set and the remaining 30% were assigned to the testing set. The division of negative cases was the same as that in positive cases.

Table 1 The characteristics of 18,441 participants

Feature	Training set (n = 12,909)	Testing set (n = 5532)	Total population (n = 18,441)
Characteristics			
Age	50.88 (19, 96)	50.81 (18, 93)	50.86 (18, 96)
Age subgroup, y			
18–64	11,170 (86.6)	4815 (87.0)	15,991 (86.7)
> 64	1733 (13.4)	717 (13.0)	2450 (13.3)
Gender (male)	7738 (59.9)	3297 (59.6)	11,035 (59.8)
SBP, mmHg	128.69 ± 17.76	128.04 ± 17.43	128.49 ± 17.66
Heart rate, beats/min	75.71 ± 7.80	75.48 ± 7.93	75.64 ± 7.84
Pulse, beats/min	77.40 ± 10.88	76.92 ± 10.94	77.25 ± 10.90
Waistline, cm	83.66 ± 10.10	83.70 ± 9.89	83.67 ± 10.04
Co-morbid conditions			
Hypertension	1502 (11.6)	868 (15.7)	2370 (12.9)
Diabetes mellitus	524 (4.1)	326 (5.9)	850 (4.6)
Hyperlipidemia	177 (1.4)	210 (3.8)	387 (2.1)
Family history	287 (2.2)	235 (4.2)	522 (2.8)
Ever-smoker	563 (4.4)	284 (5.1)	847 (4.6)
Laboratory values			
Glucose, mmol/L	5.42 ± 1.36	5.31 ± 1.36	5.39 ± 1.36
HDL, mmol/L	1.23 ± 0.21	1.21 ± 0.25	1.22 ± 0.22
TC, mmol/L	4.40 ± 0.62	4.40 ± 0.80	4.40 ± 0.68
Total protein, g/L	68.89 ± 4.21	69.15 ± 3.68	68.96 ± 4.06
Albumin, g/L	44.30 ± 2.78	44.21 ± 2.40	44.27 ± 2.67
Albumin/Globulin	1.84 ± 0.28	1.82 ± 0.29	1.84 ± 0.28
γ-GLT, U/L	27.33 ± 26.86	27.12 ± 25.18	27.27 ± 26.37
Platelets, 10^9/L	217.36 ± 58.42	218.57 ± 57.25	217.72 ± 58.07
Carotid atherosclerosis			
Yes	4587 (35.5)	1966 (35.5)	6553 (35.5)
Age 18–64 y	3246 (25.2)	1418 (25.6)	4670 (25.3)
Age > 64 y	1335 (10.3)	548 (9.9)	1883 (10.2)
No	8322 (64.5)	3566 (64.5)	11,888 (64.5)
Age 18–64 y	7924 (61.4)	3397 (61.4)	11,321 (61.4)
Age > 64 y	398 (3.1)	169 (3.1)	567 (3.1)

Categorical features represented as frequency (%). Continuous features represented as median ± SD, except age, which was median (minimum, maximum). (SBP, systolic blood pressure; HDL, high density lipoprotein; TC, total cholesterol; γ-GLT, γ-glutamyl transpeptidase)

Features preprocessing

Before implementing the models, we first preprocessed candidate features on the complete dataset, which represented the personal characteristics, vital signs, co-morbid conditions, laboratory values, and physical examinations (see Additional file 1: Table 1 for more details). Using asymptomatic CAS as a dependent variable, we performed binary logistic regression, a widely and traditionally used method in the medical and biological sciences [13] to select statistically significant features. Categorical features were encoded as binary input features and continuous features were scaled to span the range [0, 1].

Models comparison for binary classification problem

After that, we used six state-of-the-art machine learning algorithms to predict the probability of a binary outcome (asymptomatic CAS or non-asymptomatic CAS): logistic regression (LR), random forest (RF), decision tree (DT), eXtreme Gradient Boosting (XGB), Gaussian Naïve Bayes (GNB), K-Nearest Neighbour (KNN), because they are touted as currently widely and successfully classifiers for clinical data [14–16].

LR, a generalised linear regression analysis model, is used to predict a categorical dependent variable based on one or more predictor (independent) variables. That is, it is used to estimate the expectation values of each parameter in a qualitative response model [17]. This algorithm was implemented using the LogisticRegression module of Scikit-Learn. RF is a recursive method using randomization and bagging to increase the variance of ensemble trees. The outcome is directly related to the number of trees in a forest [18]. The higher the number of trees, the more exact the results obtained. We implemented it using Scikit-Learn's RandomForestClassifier module. DT [19] is used for discriminant analysis and constructed by recursive partition, whose information theory includes ID3, C4.5 and CART. ID3 cannot handle continuous data and when processing continuous attribute data, the efficiency of C4.5 is easily negatively affected by data discretization [20]. In the process of constructing the classification tree, CART uses the discretized continuous attribute derived from the Ginigain minimum of the selection criterion as the cut-off point, and the dichotomy can simplify the DT and improve its efficiency [21]. In this study, we used the CART algorithm to construct an asymptomatic CAS prediction model. XGB [7] is an optimized distributed gradient boosting library designed to be excellently scalable and highly efficient. As a modified algorithm based on the traditional gradient boosting decision tree, XGB reduces the risk of overfitting by adding regular terms and directly uses the first and the two order derivatives of the loss function [22, 23]. This algorithm was implemented using Scikit-Learn's XGBClassifier module. Naïve Bayes (NB) [24] applies Bayes' theorem with the "naive" assumption of independence between every set of features, meaning that all features contribute independently to the probability of the target outcome [25]. When the likelihood of features is presumed to be Gaussian, GNB is obtained [26]. This algorithm was implemented using Scikit-Learn's GaussianNB module. KNN [27] is based on a wealth of information among the k-closest neighbours of existing data to predict new data. In fact, it does not construct a model to predict asymptomatic CAS. Instead, the prediction is based on the largest proportion of the k-closest point, so it is often called a lazy classifier [28]. This algorithm was implemented using Scikit-Learn's KNeighborsClassifier module.

Since the values of hyperparameters must be set in advance and cannot be automatically obtained from data [11], tuning parameters are critical and specific for each model. GridSearchCV, which is implemented by estimators, is a traditional way for hyperparameter adjustment in any classification method. Consistent with the study reported by Puneet et al. [29], we optimized hyperparameters using Scikit-Learn's GridSearchCV module and fitted them to the training set without a specific validation set. Briefly, before GridSearchCV performed all necessary model fitting and outperformed the best hyperparameters, a dictionary was defined to store the hyperparameters which needed to be searched first. Fitting the GridSearchCV object not only searched for the best hyperparameters, but also obtained a new training model which automatically fitted the best cross-validation performance hyperparameters of all training sets [30]. After obtaining the optimal hyperparameter combination of each algorithm, we evaluated the model using a 30% hold-out testing set.

Considering that 30% hold-out validation may also suffer from overfitting, we performed tenfold cross-validation scheme to avoid this problem [31]. To achieve this, first, the data were partitioned into 10-equal parts. The model was trained on 9 parts and leaving 1 part for testing. This process was repeated 10 folds while changing the test part one-by-one until testing was performed on all the 10 parts.

Predictive performance measurements

Several metrics were used to evaluate performance: accuracy, F1 score, specificity, precision, recall and we visualized the area under the receiver operating characteristic curve (AUCROC). Accuracy refers to the ratio of the number of correctly predicted asymptomatic CAS to the total number of participants [32]. F1 score is composed of a weighted average of precision and recall [33]. Compared to commonly used performance metrics (including recall and specificity), AUCROC better reflected model performance. Hence, AUCROC was the main metric, while accuracy and F1 score were considered as the secondary priorities. Furthermore, a confusion matrix was used to evaluate the performance of the best model.

Statistical and machine learning analysis

Statistical analysis was performed using SPSS 23.0. Characteristics are presented as mean (± SD) for continuous features and frequencies (%) for categorical features. Binary logistic regression analysis was used to select significant features ($p < 0.2$). Machine learning models were implemented using the Scikit-Learn toolkit in Python version 3.7.4.

Results

Data description

We preprocessed 40 continuous features and 19 categorical features, and a total of 19 features were used as input features to develop models (see Additional file 1: Table 2 for more details). Among the 18,441 participants, 6553 were diagnosed with asymptomatic CAS. 59.8% were male, mean age was 50.86 years old and 13.3% were older than 64 years old. Characteristics of the participants are presented in Table 1. The training set consisted of 12,909 subjects (13.4% aged > 64 years old; 59.9% male; 35.5%

Table 2 Comparison of the predictive performance for six models (testing set)

Model	Acc (%)	Sp (%)	Pp (%)	Re (%)	F1 (%)	AUCROC
LR	74.7	86.6	68.6	53.2	59.9	0.809
RF	74.5	89.5	71.3	47.2	56.8	0.794
DT	65.4	71.8	51.2	53.8	52.5	0.628
XGB	73.4	87.8	68.0	47.2	55.7	0.788
GNB	67.0	88.0	63.1	37.2	46.8	0.753
KNN	68.8	81.5	57.7	45.6	50.9	0.704

Acc, accuracy; Sp, specificity; Pp, precision; Re, recall; F1, F1 score; AUCROC, the area under the receiver operating characteristic curve; LR, logistic regression; RF, random forest; DT, decision tree; XGB, eXtreme Gradient Boosting; GNB, Gaussian Naïve Bayes; KNN, K-Nearest Neighbour with asymptomatic CAS), while the testing set consisted of 5532 subjects (13.0% aged > 64 years old; 59.6% male; 35.5% with asymptomatic CAS). In the testing set, hypertension was present in 15.7%, diabetes mellitus in 5.9%, hyperlipidemia in 3.8%, family history in 4.2%, and ever-smoker in 5.1%. All of these values were higher than those in the training set or total participants. The laboratory values of the testing or training sets were similar to those of the total participants. In the subgroup of people over 64 years old, the number of subjects with asymptomatic CAS was significantly higher than that of those without asymptomatic CAS, in both the training and testing sets.

Model comparison for binary classification problem

A comparison of the receiver operating characteristic curve for six models is shown in Fig. 2 and Table 2. The differences between these curves were slight, but we could still clearly recognize each model. DT showed the poorest predictive performance, with the lowest AUCROC of 0.628, an accuracy of 65.4%, and a F1 score of 52.5%. In comparison the other five models improved prediction: KNN + 7.6% (AUCROC: 0.704, accuracy: 68.8% and F1 score: 50.9%); GNB + 12.5% (0.753, 67.0%, and 46.8% respectively); XGB + 16.0% (0.788, 73.4%, and 55.7% respectively); RF + 16.6% (0.794, 74.5%, and 56.8% respectively); and LR + 18.1%, which had the highest AUCROC of 0.809, an accuracy of 74.7%, and a F1 score of 59.9%. Detailed predictions of LR were presented in the form of a confusion matrix (see Additional file 1: Table 3 for more details). In the present context, LR was able to predict 1045/1966 asymptomatic CAS (sensitivity 53.2%) and 3088/3566 non-asymptomatic CAS (specificity 86.6%). The results of the tenfold cross-validation showed that LR had a better discriminative ability for asymptomatic CAS than the other five models (Table 3).

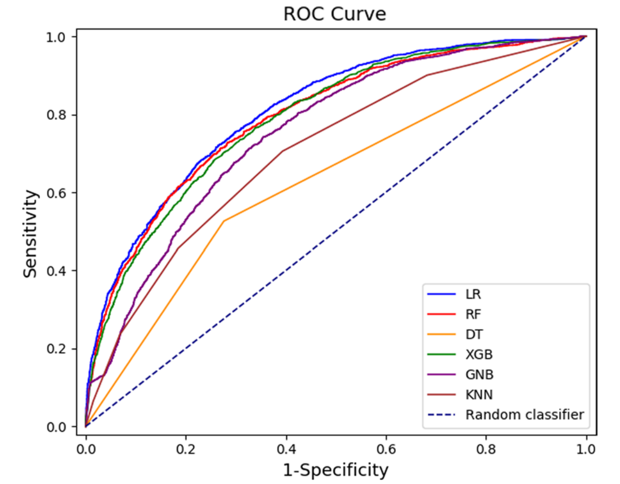

Fig. 2 Performance characteristic curves for six models (Testing set). (LR, logistic regression; RF, random forest; DT, decision tree; XGB, eXtreme Gradient Boosting; GNB, Gaussian Naïve Bayes; KNN, K-Nearest Neighbour)

Table 3 Comparison of the performance for six models (tenfold cross-validation)

Model	AUCROC	Model	AUCROC
LR	0.812	XGB	0.797
RF	0.799	GNB	0.755
DT	0.630	KNN	0.701

In order to confirm the independence between 19 features, Pearson correlation analysis for NB was performed, as shown in Fig. 3. No significant cross-correlation (Pearson correlation coefficient > 0.8) was observed, which proved that the NB worked well.

Moreover, according to the information gain values of DT model, we further ranked those 19 features, as shown in Fig. 4. Age contributed the most to the asymptomatic CAS outcome, followed by systolic blood pressure, glucose, high density lipoprotein, platelets and so on.

Discussions

In this retrospective analysis, we used six state-of-the-art machine learning models to predict asymptomatic CAS subjects based on 19 input features, which were selected according to binary logistic regression. Among the six models, LR achieved an AUCROC of 0.809, an accuracy of 74.7%, and a F1 score of 59.9% when validating in the separate hold-out testing set, generating the optimal predictive model for data, which was in keeping with other studies using machine learning techniques to assess clinical events [34, 35]. This accuracy was equivalent to or exceeded that of other evaluating tools yet used in neurology [36].

Similar to previous reports, approximately 35.5% of the included subjects presented with asymptomatic CAS. Furthermore, in the subgroup of people over 64 years old, we found that the number of subjects with asymptomatic CAS was significantly higher than that of those without asymptomatic CAS. In accordance with several studies [37, 38], age is one of the robust risk factors of asymptomatic CAS. Population aging poses a threat, and the number of people with CAS will continue to increase, which deserves more attention from the entire society. Perhaps stakeholders should put more effort into intervention measures for this threat in the future.

While machine learning algorithms help us to deal with several problems, they also present an inherent problem increasingly more evident. More often than not, machine learning is still "black box", lacking sufficient

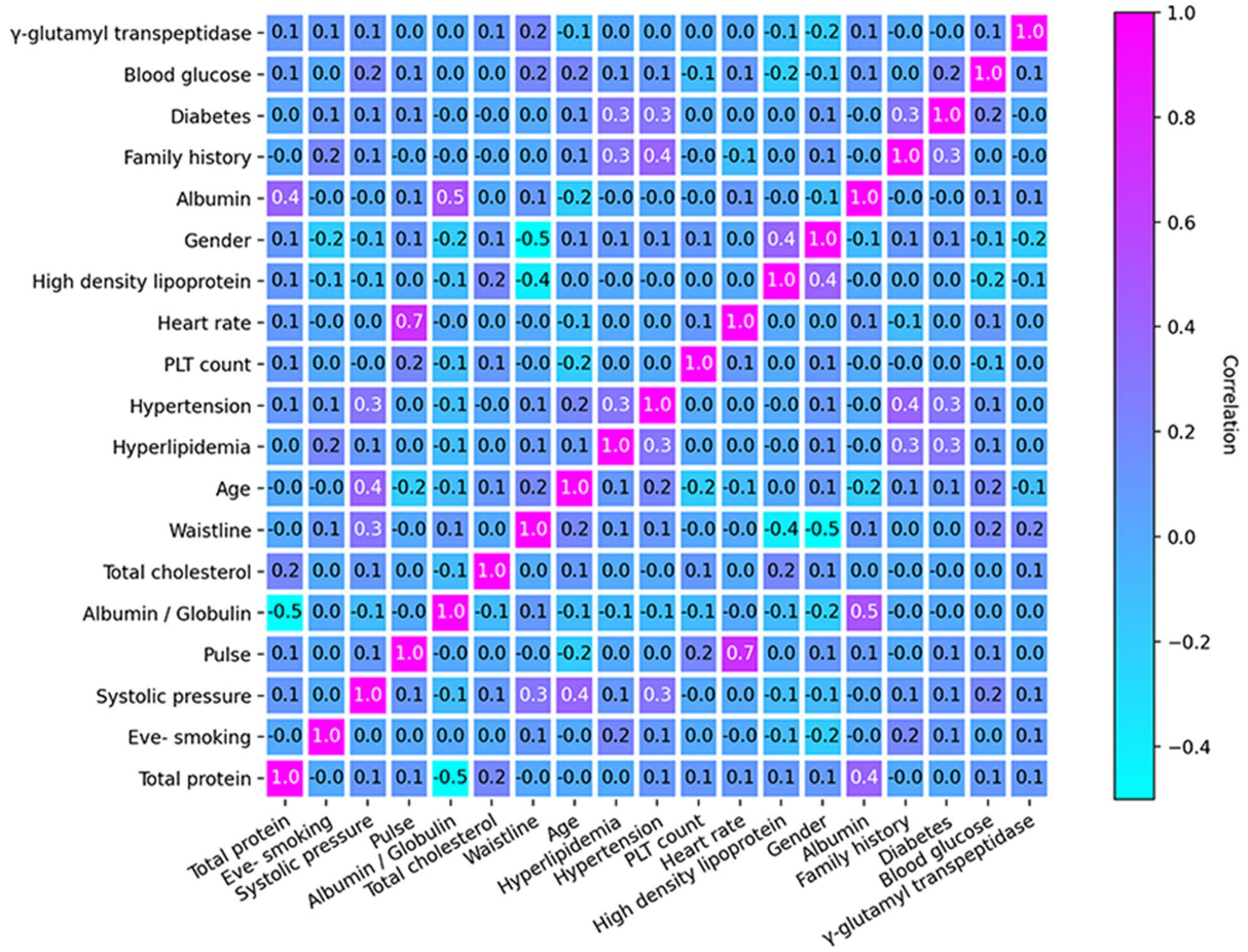

Fig. 3 Pearson correlation analysis regarding 19 features for Naïve Bayes. (*PLT*, platelets)

interpretability. It means that researchers or clinicians are progressively relying on the "black box", achieving results without even knowing what is going on in the machinery of the classifiers. If we had knowledge about some important details about the classifiers, we might achieve one result that could be much more effective. For example, when we used DT, the model already selected the most important features according the information gain values and then split, ranking of the features selected by the DT potentially provided us with additional information. In other words, by focusing on and controlling those high-risk predictors, such as age, systolic blood pressure, and glucose, we would see a more positive tendency throughout the individual's entire CAS treatment. Delaying the progression of CAS would be a tremendous relief for individuals, clinicians, and healthcare systems.

According to complex multidimensional clinical data, Xia Hu, and colleagues [39] constructed a Bayes-based learning framework to reveal predictive insights into the rapid progression of atherosclerosis in prediabetics at risk. And they found that NB was the best, obtaining an AUCROC of < 0.800. Combined with machine learning techniques, Sebastian Okser et al. [40] used the "grey zone" of genetic variation to predict increased risk of preclinical CAS. After spending a full 6-year period, they achieved AUCROCs of 0.844 and 0.761 when predicting the extreme classification of CAS risk and progression. The major disadvantages of the above assessment models, lie not only in the specific selection of samples (such as for prediabetics) but also in the requirement of a longer period. Whereas our proposed framework predicted the imminent future for medical check-up subjects, making it a much more pragmatic prescreening tool for clinicians.

Because of binary logistic regression-based feature selection and considerable hyperparameter tuning, machine learning approaches were compared at their best performance. We believe that the input features in the population-based models should depend on the availability of clinical evidence and clinical data, rather than statistical significance, thus feature selection was set at a

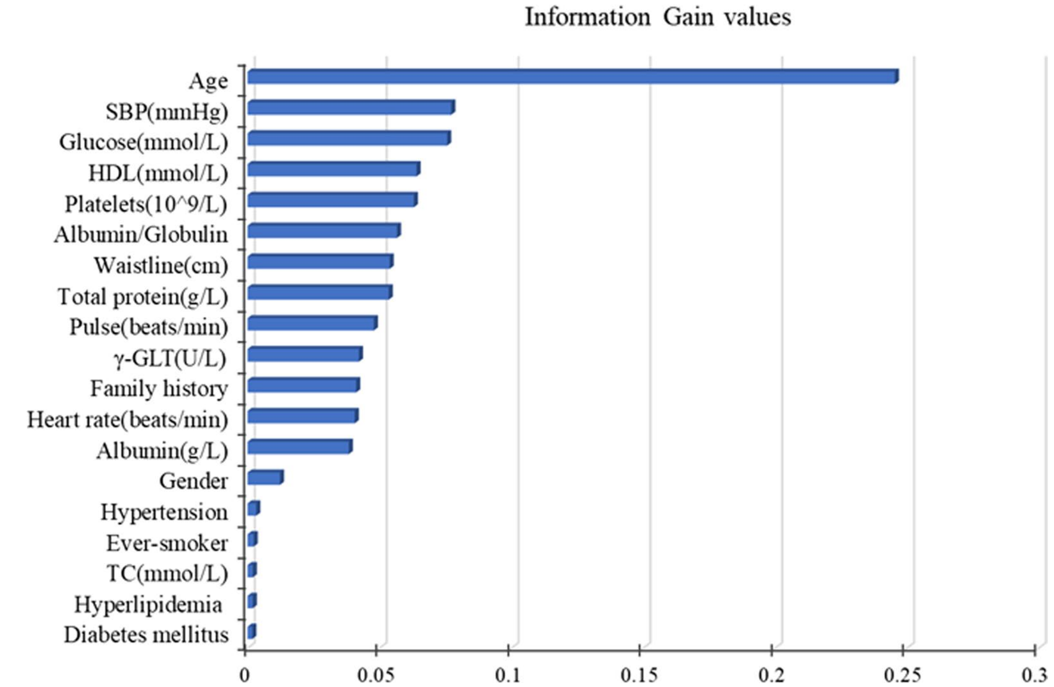

Fig. 4 19 features used for decision tree model generation and their information gain values. Features were ranked according to their information gain values which reflect the entropy gain with respect to the predictive outcome. The longer the blue transverse column (the higher the value), the greater importance on the outcome. (SBP, systolic blood pressure; HDL, high density lipoprotein; TC, total cholesterol; γ-GLT, γ-glutamyl transpeptidase)

filter *p* value of 0.2. One of the key results was that the addition of the tenfold cross-validation further verified the superior predictive ability of LR. As a widely used predictive tool in the real-world clinic, LR has extended its potential to improve prediction of precise asymptomatic CAS. Consequently, we envision that the predictions from our models will warn clinicians to pay attention to individuals who are at elevated risk of asymptomatic CAS and make them the major beneficiaries.

Based on MINimum Information for Medical AI Reporting (MINIMAR) [41], there were indeed some limitations to be noted. We believe that the data from multi centers would provide reliable predictive value on how our models identify asymptomatic CAS without selection bias. Due to the limited electronic records available from the "Rocket Frog" system, we did not include the image information, lending our inability to accurately predict the location of atherosclerosis. Moreover, the lack of external validation based on other health systems may limit the generalizability of our models.

Conclusions

In this study, we demonstrated that the logistic regression model produced a more accurate and effective prediction for asymptomatic CAS among six machine learning models. These findings set the stage for an early automatic alarming system, allowing a more precise allocation of CAS prevention measures to individuals probably to benefit most. Future large-scale studies are needed to provide more reliable and precise data for the prediction of asymptomatic CAS. Using more image information and advanced machine learning schemes are also promising.

Abbreviations

CAS: Carotid atherosclerosis; XGB: EXtreme Gradient Boosting; GNB: Gaussian Naïve Bayes; KNN: K-Nearest Neighbour; NB: Naïve Bayes; LR: Logistic regression; RF: Random forest; DT: Decision tree; AUCROC: The area under the receiver operating characteristic curve; ACC: Accuracy; F1: F1 score; Sp: Specificity; Pp: Precision; Re: Recall; SBP: Systolic blood pressure; HDL: High density lipoprotein; TC: Total cholesterol; γ-GLT: γ-Glutamyl transpeptidase.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12911-021-01480-3.

Additional file 1. Table 1–3: Candidate features, feature selection, and confusion matrix.

Acknowledgements

Not applicable.

Authors' contributions

JXF and SQZ conceived and designed the study. JXF and MYC analyzed the data, wrote, and revised the manuscript. JL and SSY developed the models and guided the Python code. JMS, QLY, and XDZ collected the data. SD, HYQ, and YXC conducted the data. SYM, MJZ, and XX obtained ethical approval. QW supported the data resources. SQZ monitored the entire planning, revision, and provided funding. All authors read and approved the final manuscript.

Funding

This study is supported by the National Natural Science Foundation of China (No. 81070999); the Foundation of Shaanxi social development and technology research project (No. 2016SF-020); the Foundation of Xi'an Science and technology plan project (No. 2019114613YX001SF039(2)); the new medical technology of the Second Affiliated Hospital of Xi'an Jiaotong University (No. 2019-32, 2018-16, 2010-22); the Fundamental Research Funds for the Central Universities (Xi'an Jiaotong University, No. xjj2014153, 2009-95); the Foundation of Second Affiliated Hospital of Xi'an Jiaotong University (No. RC(GG)201109).

Availability of data and materials

The datasets generated and/or analysed during the current study are not publicly available due privacy but are available from the corresponding author on reasonable request.

Declarations

Ethics approval and consent to participate

The ethics committee at The Second Affiliated Hospital of Xi'an Jiaotong Universityhad approved this study. Written informed consent was obtained from individual orguardian participants.

Statement

All machine learning methods were carried out in accordance with the "Guidelines for Developing and Reporting Machine Learning Predictive Models in Biomedical Research". And the protocol conformed to the Declaration of Helsinki.

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Author details

[1] Department of Neurology, The Second Affiliated Hospital of Xi'an Jiaotong University, No. 157 West Five Road, Xi'an 710004, Shaanxi, China. [2] Faculty of Electronic and Information Engineering, Xi'an Jiaotong University, Xi'an, China. [3] School of Mathematics and Statistics, Xi'an Jiaotong University, Xi'an, China. [4] Department of Health Management, The Second Affiliated Hospital of Xi'an Jiaotong University, Xi'an, China.

Received: 14 December 2020 Accepted: 26 March 2021
Published online: 05 April 2021

References

1. Sirimarco G, Amarenco P, Labreuche J, Touboul PJ, Alberts M, Goto S, Rother J, Mas JL, Bhatt DL, Steg PG, et al. Carotid atherosclerosis and risk of subsequent coronary event in outpatients with atherothrombosis. Stroke. 2013;44(2):373–9.
2. de Weerd M, Greving JP, Hedblad B, Lorenz MW, Mathiesen EB, O'Leary DH, Rosvall M, Sitzer M, Buskens E, Bots ML. Prevalence of asymptomatic carotid artery stenosis in the general population: an individual participant data meta-analysis. Stroke. 2010;41(6):1294–7.
3. McGill HC Jr, McMahan CA, Herderick EE, Malcom GT, Tracy RE, Strong JP. Origin of atherosclerosis in childhood and adolescence. Am J Clin Nutr. 2000;72(5 Suppl):1307s–15s.
4. Darcy AM, Louie AK, Roberts LW. Machine learning and the profession of medicine. JAMA. 2016;315(6):551–2.
5. Byra M, Galperin M, Ojeda-Fournier H, Olson L, O'Boyle M, Comstock C, Andre M. Breast mass classification in sonography with transfer learning using a deep convolutional neural network and color conversion. Med Phys. 2019;46(2):746–55.
6. Danielsen AA, Fenger MHJ, Østergaard SD, Nielbo KL, Mors O. Predicting mechanical restraint of psychiatric inpatients by applying machine learning on electronic health data. Acta Psychiatr Scand. 2019;140(2):147–57.
7. Yu D, Liu Z, Su C, Han Y, Duan X, Zhang R, Liu X, Yang Y, Xu S. Copy number variation in plasma as a tool for lung cancer prediction using extreme gradient boosting (XGBoost) classifier. Thorac Cancer. 2020;11(1):95–102.
8. Jiang H, Mao H, Lu H, Lin P, Garry W, Lu H, Yang G, Rainer TH, Chen X. Machine learning-based models to support decision-making in emergency department triage for patients with suspected cardiovascular disease. Int J Med Inform. 2021;145:104326.
9. Liu X, Faes L, Kale A, Wagner S, Fu D, Bruynseels A, Mahendiran T, Moraes G, Shamdas M, Kern C, et al. A comparison of deep learning performance against health-care professionals in detecting diseases from medical imaging: a systematic review and meta-analysis. Lancet Digit Health. 2019;1:271–97.
10. Jiang F, Jiang Y, Zhi H, Dong Y, Li H, Ma S, Wang Y, Dong Q, Shen H, Wang Y. Artificial intelligence in healthcare: past, present and future. Stroke Vasc Neurol. 2017;2(4):230–43.
11. van Os HJA, Ramos LA, Hilbert A, van Leeuwen M, van Walderveen MAA, Kruyt ND, Dippel DWJ, Steyerberg EW, van der Schaaf IC, Lingsma HF, et al. Predicting outcome of endovascular treatment for acute Ischemic Stroke: potential value of machine learning algorithms. Front Neurol. 2018;9:784.
12. Katakami N, Kaneto H, Shimomura I. Carotid ultrasonography: a potent tool for better clinical practice in diagnosis of atherosclerosis in diabetic patients. J Diabetes Investig. 2014;5(1):3–13.
13. Bittencourt HR, Moraes DADO, Haertel V. A binary decision tree classifier implementing logistic regression as a feature selection and classification method and its comparison with maximum likelihood. In: 2007 IEEE international geoscience and remote sensing symposium: 23–28 July 2007; 2007. p. 1755–1758.
14. Deo RC. Machine learning in medicine. Circulation. 2015;132(20):1920–30.
15. Singal AG, Mukherjee A, Elmunzer BJ, Higgins PD, Lok AS, Zhu J, Marrero JA, Waljee AK. Machine learning algorithms outperform conventional regression models in predicting development of hepatocellular carcinoma. Am J Gastroenterol. 2013;108(11):1723–30.
16. Kop R, Hoogendoorn M, Teije AT, Buchner FL, Slottje P, Moons LM, Numans ME. Predictive modeling of colorectal cancer using a dedicated pre-processing pipeline on routine electronic medical records. Comput Biol Med. 2016;76:30–8.
17. Kim J, Lee J, Lee C, Park E, Kim J, Kim H, Lee J, Jeong H. Optimal feature selection for pedestrian detection based on logistic regression analysis. In: 2013 IEEE international conference on systems, man, and cybernetics: 13–16 Oct. 2013; 2013. p. 239–242.
18. Kabiraj S, Raihan M, Alvi N, Afrin M, Akter L, Sohagi SA, Podder E. Breast cancer risk prediction using XGBoost and random forest algorithm. In: 2020 11th international conference on computing, communication and networking technologies (ICCCNT): 1–3 July 2020; 2020. p. 1–4.
19. Sherafatian M, Arjmand F. Decision tree-based classifiers for lung cancer diagnosis and subtyping using TCGA miRNA expression data. Oncol Lett. 2019;18(2):2125–31.
20. Thakur D, Markandaiah N, Raj DS. Re optimization of ID3 and C4.5 decision tree. In: 2010 International conference on computer and communication technology (ICCCT): 17–19 Sept. 2010; 2010. p. 448–450.
21. Li M. Application of CART decision tree combined with PCA algorithm in intrusion detection. In: 2017 8th IEEE international conference on software engineering and service science (ICSESS): 24–26 Nov. 2017; 2017. p. 38–41.
22. Xu H, Wang H. Identifying diseases that cause psychological trauma and social avoidance by Xgboost. In: 2019 IEEE international conference on bioinformatics and biomedicine (BIBM): 18–21 Nov. 2019; 2019. p. 1809–1813.
23. Zhang D, Gong Y. The comparison of LightGBM and XGBoost coupling factor analysis and prediagnosis of acute liver failure. IEEE Access. 2020;8:220990–1003.

24. Lou W, Wang X, Chen F, Chen Y, Jiang B, Zhang H. Sequence based prediction of DNA-binding proteins based on hybrid feature selection using random forest and Gaussian naive Bayes. PLoS ONE. 2014;9(1):e86703.
25. Abdul Muhsin Zambang M, Jiang H, Wahab L. Modeling vehicle ownership with machine learning techniques in the Greater Tamale Area, Ghana. PLoS ONE. 2021;16(2):e0246044.
26. Zhang H, Jiang T, Shan G. Identification of hot spots in protein structures using gaussian network model and Gaussian Naive Bayes. Biomed Res Int. 2016;2016:4354901.
27. Oh J, Yun K, Maoz U, Kim TS, Chae JH. Identifying depression in the National Health and Nutrition Examination Survey data using a deep learning algorithm. J Affect Disord. 2019;257:623–31.
28. Džeroski S. Data mining. In: Jørgensen SE, Fath BD, editors. Encyclopedia of ecology. Oxford: Academic Press; 2008. p. 821–30.
29. Chauhan A. Detection of lung cancer using machine learning techniques based on routine blood indices. In: 2020 IEEE international conference for innovation in technology (INOCON): 6–8 Nov. 2020; 2020. p. 1–6.
30. Shuai Y, Zheng Y, Huang H. Hybrid software obsolescence evaluation model based on PCA-SVM-GridSearchCV. In: 2018 IEEE 9th international conference on software engineering and service science (ICSESS): 23–25 Nov. 2018; 2018. p. 449–453.
31. Tworoger SS, Zhang X, Eliassen AH, Qian J, Colditz GA, Willett WC, Rosner BA, Kraft P, Hankinson SE. Inclusion of endogenous hormone levels in risk prediction models of postmenopausal breast cancer. J Clin Oncol. 2014;32(28):3111–7.
32. Teixeira F, Montenegro JLZ, Costa CAD, Righi RDR. An analysis of machine learning classifiers in breast cancer diagnosis. In: 2019 XLV Latin American computing conference (CLEI): 30 Sept.–4 Oct. 2019; 2019. p. 1–10.
33. Yao X, Sun K, Bu X, Zhao C, Jin Y. Classification of white blood cells using weighted optimized deformable convolutional neural networks. Artif Cells Nanomed Biotechnol. 2021;49(1):147–55.
34. Witteveen A, Nane GF, Vliegen IMH, Siesling S. MJ IJ: comparison of logistic regression and bayesian networks for risk prediction of breast cancer recurrence. Med Decis Making. 2018;38(7):822–33.
35. Mwanga EP, Minja EG, Mrimi E, Jimenez MG, Swai JK, Abbasi S, Ngowo HS, Siria DJ, Mapua S, Stica C, et al. Detection of malaria parasites in dried human blood spots using mid-infrared spectroscopy and logistic regression analysis. Malar J. 2019;18(1):341.
36. Thornhill RE, Lum C, Jaberi A, Stefanski P, Torres CH, Momoli F, Petrcich W, Dowlatshahi D. Can shape analysis differentiate free-floating internal carotid artery thrombus from atherosclerotic plaque in patients evaluated with CTA for stroke or transient ischemic attack? Acad Radiol. 2014;21(3):345–54.
37. Song P, Xia W, Zhu Y, Wang M, Chang X, Jin S, Wang J, An L. Prevalence of carotid atherosclerosis and carotid plaque in Chinese adults: a systematic review and meta-regression analysis. Atherosclerosis. 2018;276:67–73.
38. Alizargar J, Bai C-H. Factors associated with carotid Intima media thickness and carotid plaque score in community-dwelling and non-diabetic individuals. BMC Cardiovasc Disord. 2018;18(1):1–9.
39. Hu X, Reaven PD, Saremi A, Liu N, Abbasi MA, Liu H, Migrino RQ, Investigators ANS. Machine learning to predict rapid progression of carotid atherosclerosis in patients with impaired glucose tolerance. EURASIP J Bioinform Syst Biol. 2016;2016(1):14.
40. Okser S, Lehtimaki T, Elo LL, Mononen N, Peltonen N, Kahonen M, Juonala M, Fan YM, Hernesniemi JA, Laitinen T, et al. Genetic variants and their interactions in the prediction of increased pre-clinical carotid atherosclerosis: the cardiovascular risk in young Finns study. PLoS Genet. 2010;6(9):e1001146.
41. Hernandez-Boussard T, Bozkurt S, Ioannidis JPA, Shah NH. MINIMAR (MINimum Information for Medical AI Reporting): developing reporting standards for artificial intelligence in health care. J Am Med Inform Assoc. 2020;27(12):2011–5.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.