Original Article

A study of aortic dissection screening method based on multiple machine learning models

Lijue Liu[1,2#], Caiwang Zhang[1#], Guogang Zhang[3#], Yan Gao[1#], Jingmin Luo[3#], Wei Zhang[3#], Yi Li[1,2], Yang Mu[2]

[1]School of Information Science and Engineering, Central South University, Changsha 410075, China; [2]Hunan Zixing Artificial Intelligence Research Institute, Changsha 410007, China; [3]Xiangya Hospital, Central South University, Changsha 410008, China

Contributions: (I) Conception and design: L Liu, C Zhang, Y Gao; (II) Administrative support: Y Li, Y Mu; (III) Provision of study materials or patients: G Zhang, J Luo, W Zhang; (IV) Collection and assembly of data: G Zhang, J Luo, W Zhang; (V) Data analysis and interpretation: G Zhang, J Luo, W Zhang; (VI) Manuscript writing: All authors; (VII) Final approval of manuscript: All authors.

[#]These authors contributed equally to this work.

Correspondence to: Guogang Zhang. Office Building of Xiangya Hospital of Central South University, Changsha 410008, China. Email: ljliu@csu.edu.cn.

Background: The main purpose of the study was to develop an early screening method for aortic dissection (AD) based on machine learning. Due to the rarity of AD and the complexity of symptoms, many doctors have no clinical experience with it. Many patients are not suspected of having AD, which lead to a high rate of misdiagnosis. Here, we report the preliminary study and feasibility of rapid and accurate screening method of AD with machine learning methods.

Methods: The dataset analyzed was composed by examination data provided by the Xiangya Hospital Central South University of China which include a total of 60,000 samples, including aortic patients and non-aortic ones. Each sample has 76 features which is consist of routine examinations and other easily accessible information. Since the proportion of people who are affected is usually imbalanced compared to non-diseased people, multiple machine learning models were used, include AdaBoost, SmoteBagging, EasyEnsemble and CalibratedAdaMEC. They used different methods such as ensemble learning, undersampling, oversampling, and cost-sensitivity to solve data imbalance problems.

Results: AdaBoost performed poorly with an average recall of 16.1% and a specificity of 99.8%. SmoteBagging achieved a statistically significant better performance for this problem with an average recall of 78.1% and a specificity of 79.2%. EasyEnsemble reached the values of 77.8% and 79.3% for recall and specificity respectively. CalibratedAdaMEC's recall and specificity are 75.8% and 76%.

Conclusions: It was found that the screening performance of the models evaluated in this paper had a misdiagnosis rate lower than 25% except AdaBoost. The data used in these methods are only routine inspection data. This means that machine learning methods can help us build a fast, cheap, worthwhile and effective early screening approach for AD.

Keywords: Aortic dissection (AD); machine learning; class imbalance; screening performance

Submitted Jun 10, 2019. Accepted for publication Dec 20, 2019.
doi: 10.21037/jtd.2019.12.119
View this article at: http://dx.doi.org/10.21037/jtd.2019.12.119

Introduction

Aortic dissection (AD) is a very rare clinical emergency, the pathogenesis of which is that the blood of the aorta enters the aortic wall under the pressure of the aorta, then a dissecting hematoma is formed in the wall of the aorta, and the longitudinal axis of the aorta is extended to form a "double lumenal aorta" (1). This is a very dangerous cardiovascular disease of which the death rate is 1–2% per hour in the first 24 hours of the onset of the disease and

© Journal of Thoracic Disease. All rights reserved.

up to 60–70% in one week (2). Most patients who are not treated will die within a year (3). Although AD is an acute disease in urgent need of surgical treatment (4,5), the rate of misdiagnosis is relatively high (6). The clinical misdiagnosis rate of AD described in (7-9) is between 35% and 45%. At present, the golden criterion of AD diagnosis is CTA (computer tomography angiography), and its sensitivity is over 90% and specificity is close to 100% (10). Meanwhile, due to the rarity of AD, many doctors lack clinical experience for this disease, they usually don't suspect that the patients have this disease. In fact, most patients with AD are found because they have obvious symptoms such as severe chest and/or back pain and undergo CTA on the advice of a doctor. Few people are directly suspected of having this disease. Patients with no obvious symptoms are difficult to detect because doctors will not think about sending these people to do CTA. Once a doctor cannot tell from the symptoms that the patient has an AD, the patient will not be able to receive proper follow-up diagnosis and treatment. In summary, the current screening of patients is mainly through the subjective identification of symptoms judged by clinicians, and in most cases, it is not directly suspected of this disease. So from the current clinical perspective, a basic, cheap, and fast early screening method for AD is needed urgently. If we can detect the majority of patients at high risk by screening and then recommend that they have a CTA diagnosis, we can greatly increase the detection rate for the disease.

At present, data mining and machine learning technologies are widely used in various fields such as spam identification, credit card fraud identification, and medical diagnosis. Nowadays, the process of collecting and mining useful knowledge from big data is becoming increasingly important. With the popularity of electronic medical records, more and more valuable digital patient information is available. If machine learning techniques can be successfully used to diagnose a patient's disease, then doctors will receive useful guidance on time, which will effectively reduce the rate of misdiagnosis, missed diagnosis and the economic burden caused by these reasons. Applying machine learning to medical diagnostics is nothing new. For example, Kukar *et al.* (11) applied machine learning to the diagnosis of ischemic heart disease. Hilario *et al.* (12) applied machine learning to the prediction of lung cancer. The conclusion in paper (13) proved that the use of routine examination data can improve the risk prediction of cardiovascular disease compares to the guidelines formulated by the American College of Cardiology. Huo *et al.* (14) applied machine learning algorithms to the diagnosis of emergency patients with AD and obtaining higher accuracy than the benchmark. Inspired by these researches, we used machine learning techniques to build an early screening model of AD. Our purpose is to find a basic, cheap, and fast method, so on the advice of the doctor, we chose some routine medical examination data, lifestyle habits and other data as features. All of these examination items are the most basic and can be done in any hospital.

As mentioned above, due to the rareness of the disease, our data set is imbalanced. Many standard machine learning models, such as SVMs and decision trees, tend to improve overall prediction accuracy, which will lead to classifying samples into a majority class. That means, in this case, all test samples in our data set will be judged as non-AD patients. In the work of this study, we had verified this opinion using some standard models mentioned above. However, in early screening, we focus on finding more patients as much as possible in order to advise them to do further checks. That means we need to have a higher true positive rate rather than just a higher accuracy. Therefore, we need models that can deal with imbalance problems. Therefore, in this article, we evaluate several machine learning models and learning strategies using different sampling methods to predict whether a patient has an AD and discuss the predictive value of each method.

Methods

Dataset

Xiangya Hospital is one of the top hospitals in China. It was founded in 1906 by the Yale-China Association. There are about 6,000 outpatient visits in this hospital per day, and over 3,500 beds in the hospital's inpatients' department. In 2016, the total number of emergency visits reached 2,993,100; the number of discharged patients was more than 123,500, and more than 65,500 operations were performed. The study used data from 60,000 inpatients at Xiangya Hospital of Central South University from 2008 to 2016. The use of all data authorized by the Institute of Hypertension, Xiangya Hospital, Central South University. The 2014 ESC Guidelines on the diagnosis and treatment of aortic diseases (15) provides a simple decision method: if the patient has obvious symptoms such as chest pain or back pain, he or she may be suspected of AD, then the D-dimer will be considered to exclude AD or need further examination. It recommends the diagnosis or exclusion of

© Journal of Thoracic Disease. All rights reserved.

Table 1 Indicators recommended by 2014 ESC Guidelines

Laboratory tests	Abbreviations
Red blood cell count	RBC
White blood cell count	WBC
C-reactive protein	CRP
Procalcitonin	PCT
Creatine kinase	CK
Troponin I or T	TnI or TnT
D-Dimer	D-Dimer
Creatinine	CRE
Alanine aminotransferase	ALT
Lactate dehydrogenase	LDH
Glucose	GLU
Blood gases	BG

aortic diseases by the indicators listed in *Table 1*.

According to this table, almost all tests are belonging to routine medical examination except BG. Even BG has many identical tests items as routine medical examination, such as Ca^+, K^+, Na^+ and Cl^-. The other tests in routine medical examination which are not list in *Table 1* do not appear to be directly related to the disease, but the human body is a complex object, each indicator will affect each other. Maybe the impact of one indicator is not obvious, but combining multiple indicators will be different. Because one of the characteristics of machine learning methods is that they can deal with very high-dimensional data, and building a non-linear model between input and output, it may get some associations that not yet found in clinical statistics. Some method even can rank the importance of the features. So we refer to the important index of ESC guide and consider establishing an early screening model from the perspective of machine learning, some routine inspection result such as the patient's routine blood examination, biochemical examination, and clotting routine examination items are selected as features. All these examinations are basic and can be done in any hospital even in most primary hospitals with weak facilities. The cost of doing these examinations is relatively low and the check time is relatively short. Some living habits, family history of genetic disease, and some other data were also chosen according to the doctor's experiences. If we can build a model to predict the patients with relatively high sensitive through these features, it will help people detect the risk of AD with a basic, cheap and fast method. Finally we used 76 features and obtained corresponding data from the electronic medical record. Detail of these features are described in *Table 2*.

Although we tried to collect as much data as possible, there are still some missing data in our dataset. For the vacant values in the samples set, we used the stratified random sampling method to fill the data, which means that the missing data was filled based on the random sampled values in the positive and negative categories. This method of filling completes the dataset and makes the populated dataset's distribution as close as possible to the actual distribution of the data set. Data collection is not the focus of this study, so it is not described in detail.

Among 60,000 patients involved in this study, there were only 1,000 aortic patients that had positive samples, which means the imbalanced ratio of positive to negative samples was 60:1. The age of patients ranged from 14 to 89 years. In the positive patients, there were 294 patients over 60 years old, and 716 patients over 40 years old. This shows that middle-aged and elderly people seem to be at greater risk of AD. There were 724 male and 276 female in the AD patients, the number of men is 2.5 times that of women, Crawford (3) shows that men are at higher risk of suffering from AD than women.

Study methods

Our purpose is to establish a model to predict whether a patient is at risk for AD. Some supervised learning methods are chosen to achieve this object. Since the predict result of our problem is yes or no, this kind of problem is a binary classification problem in machine learning. In order to use this kind of methods, we need a lot of "labeled data" of both class, which means each sample in the data set is composed of two parts (X, y), where X is a vector of features, y is the "label", that is the right answer of X. Different machine learning models use different loss function to describe the distance between the right answer and the predict result. The purpose of learning is to continuously use training data to minimize the loss function. When this goal is achieved, that means we can establish a hyperplane which can divide different classes into different sides of the plane.

Since our dataset is highly imbalanced and there are only 1,000 positive samples among 60,000 patients, the standard machine learning model cannot meet the requirements. For example, if all samples are predicted to be negative samples in our case, the accuracy can be up to

© Journal of Thoracic Disease. All rights reserved.

Table 2 Features of dataset

BR (blood routine)	Biochemical examination	Clotting routine examination	Others
1.1 WBC*	2.1 TP	3.1 PT	4.1 Chest pain*
1.2 RBC*	2.2 ALB	3.2 APTT	4.2 Stomach ache
1.3 HGB	2.3 GLO	3.3 TT	4.3 Aortic valve area murmur
1.4 HCT	2.4 GLU*	3.4 PT%	4.4 Dizziness and headache
1.5 MCV	2.5 BUN	3.5 D-Dimer*	4.5 Hypertension
1.6 MCH	2.6 UA	3.6 INR	4.6 Family history of hypertension
1.7 MCHC	2.7 CRE*	3.7 FIB	4.7 Family history of aortic dissection
1.8 PLT	2.8 TBIL		4.8 Chest trauma history
1.9 NEUT	2.9 DBIL		4.9 Smoking and duration
1.11 MONO	2.10 CO2CP		4.10 Diastolic pressure, systolic pressure
1.11 EO	2.11 Ca+*		4.11 Heart rate
1.12 BASO	2.12 P+		4.12 Heart disease
1.13 LYMPH	2.13 K+*		4.13 Family history of heart disease
1.14 LYMPH%	2.14 Na+*		
1.15 MONO%	2.15 Cl–*		
1.16 NEUT%	2.16 Mg+		
1.17 EO%	2.17 CHO		
1.18 BASO%	2.18 TG		
1.19 RDW	2.19 HDL		
1.20 PCT*	2.20 LDL		
1.21 MPV	2.21 CK*		
1.22 PDW	2.22 LDH*		
	2.23 CKMB		
	2.24 MB		
	2.25 HBA1C		
	2.26 AG		
	2.27 ALP		
	2.28 TBA		
	2.29 TnI*		
	2.30 TnT*		
	2.31 CRP*		
	2.32 ESR		
	2.33 ALT*		
	2.33 AST		
	2.34 PCT		

*, the tests suggested by ESC guide.

© Journal of Thoracic Disease. All rights reserved.

(60,000–1,000)/60,000=98.3%, but that means nothing, because no positive samples would be found. In medical diagnosis, the cost of misdiagnosing the patient as having no disease is usually much greater than the opposite. So in order to handle this situation, we need some special way to deal with imbalanced data. To solve the problem of prediction based on imbalanced datasets, the following two methods are commonly used:

- Processing at the data level

  The commonly used methods are oversampling and undersampling. Oversampling increases the number of minority classes. For example, if there are too few positive samples, some of the positive samples can be simply cloned or some new positive samples are generated by some special methods. The commonly used method to generate new samples is the SMOTE (16) method. This method creates several minority samples randomly between a minority sample and several of its nearest neighbors, which are of the same category. Due to the increased number of samples, the oversampling methods usually cause an increase in training cost. On the other hand, the undersampling methods only take part of the majority class samples into training. Commonly used undersampling methods are ENN (17), RENN (18), and sampling with replacement. The disadvantage of undersampling is that it can result in partial loss of the class information. The method of undersampling and oversampling can both improve the imbalance situation to some extent. The classical algorithms which using the undersampling or oversampling methods are SmoteBoost (19), SmoteBagging (20), RusBoost (21), and EasyEnsemble (22).

- Processing at the model level

  It is also a common method to deal with the imbalance problem which modifies the existing model so that the algorithms can handle the problem of imbalanced datasets. The commonly used method is the cost-sensitive method (23,24). The decision tree method for cost-sensitivity is further divided into a cost-sensitive decision tree (25) and a sample cost-sensitive decision tree (26). The class cost-sensitive decision tree increases the cost of misclassifying the minority class and decreases the cost of misclassifying the majority samples. The sample cost-sensitive decision tree is a decision tree which is sensitive to the sample. It gives more weight to some important samples in the procedure of constructing the decision tree. One class learning is also a model level method to deal with the problem of class imbalance. Unlike ordinary algorithms, there are only samples in one class in the training set. In addition, ensemble learning also takes place in dealing with class imbalance problems. Ensemble learning algorithms are integrations of basic classifiers. Each base classifier is trained using a subset obtained from oversampling or undersampling from the original data set. SmoteBagging, RusBoost, and EasyEnsemble, are all ensemble learning algorithms dealing with class imbalance.

In this study, we tried several different machine learning algorithms to analyze the samples in our dataset. These algorithms all used ensemble learning, and the latter three adopt undersampling, oversampling and cost sensitivity method respectively. Then we compared their screening effect of AD. The detailed description of the algorithm used is as follows.

AdaBoost

One of the most famous algorithm in ensemble learning is AdaBoost (27) because of its simplicity and high-precision. It is an algorithm that promotes weak classifiers into strong classifiers. It first trains a base classifier from the initial training dataset and then modifies the weights of the initial training data samples based on the classifier's performance. The weights of the samples which are misclassified become larger, and a new base classifier will be trained with the samples whose weights have been changed. Repeatedly until the iteration stop condition is satisfied, all the base classifiers are finally weighted and combined to obtain the final classifier. The decision function of AdaBoost is as follows:

$$H(X)=\sum_{t=1}^{T} a_t h_t(x) \quad [1]$$

Where $h^t(x)$ is the t^{th} base classifier α_t is the weight of the t^{th} base classifier, and T is the total number of base classifiers.

EasyEnsemble

EasyEnsemble is an algorithm dealing with imbalanced data by undersampling. It is an algorithm based on the ensemble of AdaBoost algorithm.

In the EasyEnsemble algorithm, suppose the training dataset of the minority class is *P*, and the training dataset of the majority class is *N* where $|N|>>|P|$, *N* is divided

© Journal of Thoracic Disease. All rights reserved.

into N_1, N_2, …, N_T and $|N_i|=|P|(i=1, 2, \ldots, T)$. The datasets N_i and P are used to train a base classifier H_i, which is an AdaBoost classifier. The EasyEnsemble classifier is combined with the T AdaBoost classifiers.

SmoteBagging

SmoteBagging is an algorithm dealing with imbalanced data by oversampling. SmoteBagging is also an ensemble learning algorithm that uses a voting strategy and the methods of Smote and Bagging. As mentioned earlier, smote is a method of synthesizing some new samples artificially, while bagging is an ensemble strategy which uses sampling with replacement method and ensemble the base classifiers by voting. Assuming N_n is the number of majority samples, and N_p is the number of minority samples, during the oversampling process, N_n*b minority samples are selected randomly with replacement, where $b\in\{x \mid x=0.1*k, k=1,2,\ldots,10\}$. Meanwhile, $N_n*(1-b)$ minority samples are synthesized by the Smote algorithm. After the procedure of oversampling, the number of minority samples equals the number of majority. SmoteBagging is an oversampling algorithm, but it does not only use Smote or Bagging alone, on the contrary, it use them both. The advantage of this is that it reduces the risk of overfitting the Bagging method and reduces the negative impact caused by the synthetic data, which is produced by Smote. Since different base classifiers use different b values, the diversity of base classifiers is guaranteed, which is required to improve the classification accuracy and the ability of the ensemble learning model to generalize.

CalibratedAdaMEC

CalibratedAdaMEC is both an ensemble learning algorithm and a cost-sensitive algorithm, which means there is a cost matrix in this algorithm.

$$c=\begin{bmatrix} 0 & c_{FN} \\ c_{FP} & 0 \end{bmatrix} \quad [2]$$

Where c_{FN} is the cost of misclassifying positive samples as negative samples, and c_{FP} is the cost of misclassifying negative samples as positive samples. If the positive samples are the minority class, c_{FN} is usually larger than c_{FP}. From 'Ada' we can see that this algorithm is an improved version of the AdaBoost algorithm. In this algorithm, Platt scaling is used to "calibrat" the output of AdaMEC to the probability space. Whether the sample is positive or negative is determined by probability. AdaMEC is an algorithm which modifies the output of AdaBoost according to the cost matrix. The cost-sensitive algorithm is also a commonly used method to deal with the imbalance problem. It avoids the risk of losing some information caused by undersampling and overfitting caused by oversampling.

Results

Evaluation method

As we have described earlier, due to the imbalanced nature of this data set, the traditional evaluation indicator such as accuracy and precision are no longer appropriate. The accuracy and precision are calculated as follows:

$$accuracy=\frac{TP}{P+N} \quad [3]$$

$$precision=\frac{TP}{TP+FP} \quad [4]$$

Where TP is the number of true positive, P is the number of positive samples, N is the number of negative samples, and FP is the number of false positive.

Because of the imbalance, that is $P<<N$, even $TP=P$, which means all positive samples are found, the accuracy still relatively low. Precision has similar problems, even the false positive rate is not high, the number of false positive (FP) can be still large, as a result, the precision value is not satisfactory.

The commonly used indicators in this case are recall rate, also called sensitivity, and specificity. The recall rate indicates the proportion of samples that correctly predict as positive to all positive samples. The higher the recall rate is, the lower the rate of missed diagnosis is. The recall rate is calculated as follows:

$$(recall)\ r=\frac{TP}{TP+FN} \quad [5]$$

Where FN is the number of false negative.

Specificity is the correct classification rate of the negative samples, which is also known as the true negative rate; it is equivalent to the recall rate of negative samples. The higher the specificity, the lower the misdiagnosis rate. The specificity is calculated as follows:

$$(specificity)\ S=\frac{TN}{TN+FP} \quad [6]$$

Where TN is the number of true negative.

© Journal of Thoracic Disease. All rights reserved.

Table 3 The models' training time (unit: s)

Methods	1st	2nd	3rd	4th	5th	6th	7th	Average
ada	9	9	9	8	9	9	8	9
easy	97	97	99	98	97	99	97	98
smote	1,890	1,866	1,882	1,875	1,864	1,865	1,869	1,873
mec	101	100	100	101	100	100	101	100

Experimental settings

For the convenience of writing, we will call AdaBoost, EasyEnsemble, SmoteBagging and CalibratedAdaMEC as *ada*, *easy*, *smote* and *mec* without affecting the expression.

For the *mec* method, the cost of the predicted error of the AD patients was set to 4,000; the number of decision tree used in AdaBoost was 200.

The number of training sets for the *easy* model was set to 60 because the imbalance ratio between positive samples and negative samples was 1:60, and the number of decision tree used in each AdaBoost was set to 200.

AdaBoost was chosen as the base classifier of smote, and the number of base classifier was set to 200.

In addition, we used a 7-fold cross validation method to verify the validity of our model. Due to having too few AD samples in this study, a 7-fold cross validation rather than a 5-fold cross validation or a 10-fold cross validation was used so as not to make the training or test sets too small.

Training time analysis

The models we used include oversampling models, undersampling models and cost-sensitive models. Because these models use different strategies to deal with class imbalances, their training times are different. In this section we analyze the training time of each model through experiments.

Since different hardware devices and programming environments affect the run time of the program, this test can only represent the training time of the equipment used in this paper (cpu: i5 6500, RAM: 8 GB, Hard Disk: 1 TB, 7200 rpm, operation: windows 10, 64 bit, python: 3.6). In order to minimize errors, we used the average of seven-fold cross validation training time which list in *Table 3*.

The *ada* algorithm is a traditional ensemble method. It does not have any special sampling processing. The base classifier of *easy* is *ada*, so the training time of *ada* is very short, far less than easy.

The *easy* algorithm is an undersampling method. It does not add extra data during data sampling and model training and although it needs to build more decision trees than other models, it only uses a very small portion of all the samples in the dataset, so the *easy* model's training time is much shorter than *smote*.

As an oversampling algorithm, the *smote* algorithm uses both Smote and Bagging methods. For the high imbalance of our dataset, the training data almost doubled in our experiment, so its training time is the longest in the three models.

The *mec* model is a cost sensitive model, and it does not produce additional training data like *smote*. It uses all samples when training every decision tree while the *easy* algorithm only takes a small part of all the samples. Thus, although the *mec* algorithm builds fewer decision trees than easy, the training time of the *mec* model is longer than *easy*.

In terms of time efficiency, *ada* has the shortest training time. It only has one idea of ensemble learning. At the same time, the latter three add some operations to deal with imbalanced data based on ensemble learning. It is obvious that the time of the latter three is longer than *ada*, but how about the performance of the results of these four algorithms?

Experimental results and analysis

In the following, AD is the minority/positive class and non-AD is the majority/negative class.

The experiments were performed with seven-fold cross-validation, and the recall rate and specificity were used as evaluation indicators. To evaluate the model as accurately as possible, the seven-fold cross validation was repeated five times. *Table 4* shows the result of cross validation and their average value-each row represents the average of a seven-fold cross validation result, and the final column represents the average of the result of multiple seven-fold cross validations.

© Journal of Thoracic Disease. All rights reserved.

Table 4 The five 7-fold cross validation result of each model

Algorithms	Evaluation	1st	2nd	3rd	4th	5th	Average
ada	Recall (%)	14.4	16.9*	15.3	16.1	15.2	15.6
	Specificity (%)	99.8*	99.1	99.2	99.7	99.3	99.4*
easy	Recall (%)	77.4	78.4*	77.9	77.5	77.7	77.8*
	Specificity (%)	80.6*	80.1	78.2	77.2	77.3	79.3
smote	Recall (%)	77.9	78.0	78.3	77.8	78.5*	78.1*
	Specificity (%)	79.5	78.1	79.9	77.5	81.1*	79.2
mec	Recall (%)	74.9	73.7	74.3	75.9	80.0*	75.8*
	Specificity (%)	74.8	76.7	76.3	75.3	76.8*	76.0

*, the best results in 5 tests and the best average result.

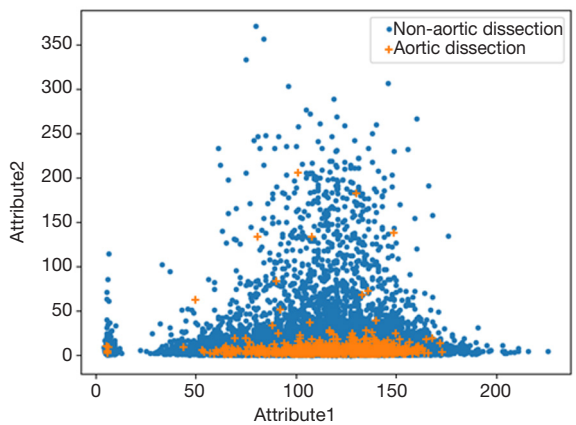

Figure 1 The overlap of two attributes of our dataset in different classes.

From *Table 4*, We can see that *ada* has a high specificity but a very low recall rate, which means that most of the non-patients were screened out while only a small part of the patients were screened out. That because without special method to deal with imbalance, the model tends to identify samples as majority classes to improve precision and accuracy.

The *smote* model adds a lot of additional training data during training and makes the training time of the model very long. However, in *Table 3*, we can see that the two models of *smote* and *easy* had comparable screening effects on the dataset. This is inconsistent with the view of Ali *et al.*'s article (28), which said *smote* can obtain better classification results than other ensemble learning algorithms by adding a large number of additional training samples.

What makes sense is that from the last line in *Table 3*, the *smote* model is 0.5% higher on recall than *easy*. Although *smote* is 0.1% lower than *easy* in terms of specificity, what is more important in this study is the increase in the recall rate. It should be pointed out that the recall rate of *smote* is not significantly higher than that of *easy*, but the training time of the *smote* model is 5 times that of the *easy* model. Whether it is necessary to achieve a slight increase in recall rate at the expense of a large amount of training time requires additional analysis.

As can be seen from *Figure 1*, the AD samples in this dataset are more concentrated, and there is a serious overlap between different classes, so the majority class is more sensitive to the dataset's size. The positive class has the same amount of samples compared to the negative class in each base classifier's training set, so the base classifiers prefer negative classes, which will increase the specificity of the model and suppress the recall rate. From *Table 3* we can see the recall rates for both *smote* and *easy* are lower than the specificity, which is because of the reason above.

The classification result of the *mec* model changes with the change of the cost matrix. If we want to improve the recall rate, we can just improve the misclassification cost of the positive class, and *easy* and *smote* cannot be so flexible due to their own characteristics. Unfortunately, the screening effect of the *mec* model is 2% to 3% worse than that of the *smote* and *easy* models. The reason for this may be that the samples' distribution of this dataset is very complicated, while the cost-sensitive method does not have a processing strategy to sample distribution.

Conclusions

Class imbalance problems are common in the real world. In

© Journal of Thoracic Disease. All rights reserved.

cases such as identification of credit card fraud, most users are normal. In medical diagnosis, the problem of imbalance is prominent, especially for some rare diseases such as AD. The imbalance rate of those existing datasets is usually higher.

AD, as a serious acute disease, needs to be discovered in time, but it is often missed in practice. Some patients even failed to detect suspected AD and lost the possibility of further diagnosis. The results of this study provide a reliable and effective diagnostic model for the early diagnosis of AD. Through simple and basic examination items, such as routine blood, biochemical, routine blood coagulation examinations, knowledge of the patient's habits, past medical history and genetic history, we can get early screening result of AD through these models. They can greatly reduce the misdiagnosis rate of AD. The predicted result can be provided to the doctor as a very important reference.

In the several machine learning models evaluated in this paper, the SmoteBagging model and the EasyEnsemble model get the best results in the experiments. The CalibratedAdaMEC algorithm as a cost sensitive method is obviously poor in this study, while it is the most flexible model. However, it should be pointed out that even the CalibratedAdaMEC model with the worst screening effect among the latter three models also obtained much better missed rate than that described in the paper (7-9), and the missed rate of CalibratedAdaMEC was less than 25%, while the missed rate in the paper (7-9) was between 35% and 45%.

This study, as the first study of using machine learning in the early screening of AD [Huo *et al.* (27) published in nature focused on the re-confirmation of machine learning after initial doctor's judgment to avoid doctor's misdiagnosis], explored the performance of several different ensemble learning algorithms on AD patient screening. With the current popularity of electronic medical records, we are able to collect more and more patient information. If we can make a comprehensive disease diagnosis system through data mining and machine learning algorithms for a variety of serious, uncommon, easily misdiagnosed and often missed diseases, in addition to reducing the economic burden on patients, we can provide an early warning to patients, which will enable patients to be diagnosed before the onset of the disease.

Acknowledgments

We acknowledge the data set supported by the Xiangya Hospital Central South University in China and the students of Central South University for their help and support—Shihao Li, Xuejian Sun and Chengyu Lin who collect the data used in this paper.

Funding: We acknowledge the support received from the SINOBIOWAY fund (No. 33020128038), Clinical Big Data System (No. 33020125039).

Footnote

Conflicts of Interest: The authors have no conflicts of interest to declare.

Ethical Statement: The authors are accountable for all aspects of the work in ensuring that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved. This study was approved by the Institute of Hypertension, Xiangya Hospital, Central South University.

Open Access Statement: This is an Open Access article distributed in accordance with the Creative Commons Attribution-NonCommercial-NoDerivs 4.0 International License (CC BY-NC-ND 4.0), which permits the non-commercial replication and distribution of the article with the strict proviso that no changes or edits are made and the original work is properly cited (including links to both the formal publication through the relevant DOI and the license). See: https://creativecommons.org/licenses/by-nc-nd/4.0/.

References

1. Isselbacher EM. Dissection of the Descending Thoracic Aorta: Looking Into the Future. JACC 2007;50:805-7.
2. Mészáros I, Mórocz J, Szlávi J, et al. Epidemiology and Clinicopathology of Aortic Dissection. Chest 2000;117:1271-8.
3. Crawford ES. The Diagnosis and Management of Aortic Dissection. JAMA 1990;264:2537-41.
4. Hagan PG, Nienaber CA, Isselbacher EM, et al. The International Registry of Acute Aortic Dissection (IRAD): New Insights Into an Old Disease. JAMA 2000;283:897-903.
5. Pape LA, Awais M, Woznicki EM, et al. Presentation, Diagnosis, and Outcomes of Acute Aortic Dissection: 17-Year Trends From the International Registry of Acute Aortic Dissection. JACC 2015;66:350-8.
6. Pourafkari L, Tajlil A, Ghaffari S, et al. The frequency

© Journal of Thoracic Disease. All rights reserved.

of initial misdiagnosis of acute aortic dissection in the emergency department and its impact on outcome. Intern Emerg Med 2017;12:1185-95.
7. Chen XF, Xiao-Min LI, Chen XB, et al. Analysis of Emergency Misdiagnosis of 22 Cases of Aortic Dissection. Clinical Misdiagnosis & Mistherapy 2016.
8. Teng Y, Gao Y, Feng S, et al. Diagnosis and Misdiagnosis Analysis of 131 Cases of Aortic Dissection. Chinese Journal of Misdiagnostics 2012;12:1873-3.
9. Wang H, Zhu Z. Analysis on clinical features and misdiagnosis of 58 patients with acute aortic dissection. Hainan Medical Journal 2016;27:800-2.
10. Vardhanabhuti V, Nicol E, Morgan-Hughes G, et al. Recommendations for accurate CT diagnosis of suspected Acute Aortic Syndrome (AAS) - On behalf of the British Society of Cardiovascular Imaging (BSCI)/British Society of Cardiovascular CT (BSCCT). Br J Radiol 2016;89:20150705.
11. Kukar M, Kononenko I, Grošelj C, et al. Analysing and improving the diagnosis of ischaemic heart disease with machine learning. Artif Intell Med 1999;16:25-50.
12. Hilario M, Kalousis A, Müller M, et al. Machine learning approaches to lung cancer prediction from mass spectra. Proteomics 2003;3:1716-9.
13. Weng SF, Reps J, Kai J, et al. Can machine-learning improve cardiovascular risk prediction using routine clinical data. PLoS One 2017;12:e0174944.
14. Huo D, Kou B, Zhou Z, et al. Machine learning model to classify aortic dissection patients in the early diagnosis phase. Sci Rep 2019;9:2701.
15. Erbel R, Aboyans V, Boileau C, et al. 2014 ESC guidelines on the diagnosis and treatment of aortic diseases: document covering acute and chronic aortic diseases of the thoracic and abdominal aorta of the adult. The task force for the diagnosis and treatment of aortic diseases of the European Society of Cardiology (ESC). Eur Heart J 2014;35:2873-926.
16. Chawla NV, Bowyer KW, Hall LO, et al. SMOTE: synthetic minority over-sampling technique. J Artif Intell Res 2002;16:321-57.
17. Wilson DL. Asymptotic Properties of Nearest Neighbor Rules Using Edited Data. IEEE Trans Syst Man Cybern 1972;SMC-2:408-21.
18. Tomek I. An Experiment with the Edited Nearest-Neighbor Rule. IEEE Trans Syst Man Cybern 2007;SMC-6:448-52.
19. Chawla NV, Lazarevic A, Hall LO, et al. SMOTEBoost: Improving Prediction of the Minority Class in Boosting. Knowledge Discovery in Databases: Pkdd 2003, European Conference on Principles and Practice of Knowledge Discovery in Databases, Cavtat-Dubrovnik, Croatia, September 22-26, 2003, Proceedings. DBLP, 2003:107-19.
20. Wang S, Yao X. Diversity analysis on imbalanced data sets by using ensemble models. Computational Intelligence and Data Mining, Nashville, TN, USA, 2009:324-31.
21. Seiffert C, Khoshgoftaar TM, Hulse JV, et al. RUSBoost: Improving Classification Performance when Training Data is Skewed. 19th International Conference on Pattern Recognition, Tampa, Florida, USA, 2008:1-4.
22. Liu XY, Wu J, Zhou ZH. Exploratory undersampling for class-imbalance learning. IEEE Trans Syst Man Cybern B Cybern 2009;39:539-50.
23. Domingos P. MetaCost: a general method for making classifiers cost-sensitive. ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM, 1999:155-64.
24. Elkan C. The Foundations of Cost-Sensitive Learning. Seventeenth International Joint Conference on Artificial Intelligence. 1991:973-8.
25. Núñez M. The use of background knowledge in decision tree induction. Mach Learn 1991;6:231-50.
26. Bahnsen AC, Aouada D, Ottersten B. Example-dependent cost-sensitive decision trees. Expert Syst Appl 2015;42:6609-19.
27. Freund Y, Schapire RE. A Decision-Theoretic Generalization of On-Line Learning and an Application to Boosting. J Comput Syst Sci 1997;55:119-39.
28. Ali A, Shamsuddin SM, Ralescu AL. Classification with class imbalance problem: a review. Int J Adv Soft Comput 2015;7:176-204.

Cite this article as: Liu L, Zhang C, Zhang G, Gao Y, Luo J, Zhang W, Li Y, Mu Y. A study of aortic dissection screening method based on multiple machine learning models. J Thorac Dis 2020;12(3):605-614. doi: 10.21037/jtd.2019.12.119

© Journal of Thoracic Disease. All rights reserved.