An accurate diagnosis of coronary heart disease by Catboost, with easily accessible data

Xiaoyu Zhang\textsuperscript{1,2}, Mu Wang\textsuperscript{1,2}, Wei Wei\textsuperscript{3}, Yang Xu\textsuperscript{1}, Lisheng Gao\textsuperscript{1}, Yining Sun\textsuperscript{1}, Zuchang Ma\textsuperscript{1*}, Shijun Wang\textsuperscript{4}

\textsuperscript{1} Institute of Intelligent Machines, Hefei Institutes of Physical Science, Chinese Academy of Sciences, Hefei, China
\textsuperscript{2} Department of Automation, University of Science and Technology of China, Hefei, China
\textsuperscript{3} Shenzhen Ramaxel Co., Ltd., Shanghai, PR China
\textsuperscript{4} Dalian medical university
\* Corresponding author’s e-mail: zcma@iim.ac.cn

Abstract. Coronary heart diseases (CHD) have become the leading cause of death worldwide. Coronary angiography is the “gold standard” for diagnosing this disease. However, the invasive risk and expensive price make it difficult to promote on a large scale. This study was using Catboost to diagnose CHD through simple indicators. 2642 samples, including 717 patients, were collected from 2018 to 2019. 33 features were collected, including demography, anthropometry, questionnaire and laboratory examination indicators. The diagnosis model of CHD was established by using Catboost, random forest and logistic regression. Accuracy and area under ROC (AUROC) were used to evaluate the classification performance of the diagnosis models. In order to facilitate the application, we also set up a simplified model merely based on non-laboratory dataset. Catboost showed the best performance in identifying patients with CHD. The accuracy of Catboost, random forest and logistic regression was 82.5%, 75.1%, 75.8%, respectively, and the AUROC of them was 0.881, 0.837, 0.832, respectively. Age, total cholesterol and family history of coronary heart disease were the three most important risk factors for diagnosing CHD. Catboost also worked best in simplified models with 77.9% accuracy and 0.857 AUROC. The models can contribute to early screening and diagnosis for CHD, which would facilitate the prevention and timely treatment of the diseases.

1. Introduction
Coronary heart disease (CHD) is the most common type of cardiovascular disease, which is the leading cause of death worldwide [1]. According to WHO, CHD killed more than 17 million in 2015 and 25% of them had no symptoms before sudden death [2]. Early screening can reduce the risk of coronary heart disease morbidity and mortality. Coronary angiography is used as the gold standard for CHD clinical diagnosis [3]. However, it is intrusive, expensive and required professional operation, which limit their application in large-scale population screening and monitoring. Other methods, such as electrocardiogram and treadmill exercise test, have low diagnostic sensitivity [4-5]. Therefore, researchers have been looking for an accurate, non-invasive, and convenient method of CHD diagnosis and early screening.
Machine learning, which intersects statistics and computer science, can detect complex nonlinear relationship among variables, through minimizing the errors between predictions and observations [6-7]. It has been widely used in predicting diseases, such as Parkinson, liver disease, skin disease, diabetes, some cancers [8-10], and as well as CHD [11-12]. Amin MS et al. utilized seven algorithms, including 13 features to predict heart diseases and Vote algorithms performed the best accuracy, 87.4% [7]. Marateb et al. proposed a fuzzy rule-based system based on a neuro-fuzzy classifier and it has 84% accuracy in predicting CHD [13]. Alizadehsani also predicted the left anterior descending coronary artery, left circumflex branch and right coronary artery stenosis, through a combination of SVM and kernel fusion algorithm. The results showed that the accuracy of these three parts were 86.14%, 83.17%, and 83.50% [14]. Most previous studies used clinical parameters, such as echocardiography and exercise tests, to predict CHD. Although these features are relatively simpler to collect than coronary angiography, the inspection process takes a lot of time and requires the participation of professional physicians. As a result, it is a challenge to intelligently diagnose CHD in large scale, especially in poor areas.

Therefore, the purpose of this study was to establish a new diagnostic method of CHD by using low-cost and easy to collect features.

2. Methods

2.1. Dataset
From 2018 to 2019, 2642 participants were randomly recruited, including 717 CHD. The diagnostic criteria of CHD are: medical records have a history of CHD, or coronary angiography shows coronary artery stenosis ≥50% (Gensini method). This study was approved by local Ethics Committee and written informed consent was obtained from each participant.

| Type                  | Features          | Description                                      | Range          |
|-----------------------|-------------------|--------------------------------------------------|----------------|
| **Demography**        | Age               | Age (in years)                                   | (21-71) years |
|                       | Sex               | 0: female; 1: male                               | (0,1)          |
|                       | Education         | 0: primary school or below; 1: Middle school; 2: College degree or above | (0-2) |
| **Anthropometry**     | BMI               | Body mass index (kg/m2)                           | (16.5-39.5) kg/m2 |
|                       | WC                | Waist circumference (cm)                          | (61-125) cm    |
|                       | SBP               | Systolic blood pressure (mmHg)                    | (80-185) mmHg  |
|                       | DBP               | Diastolic blood pressure (mmHg)                   | (51-116) mmHg  |
| **Laboratory examination** | FBS               | Fasting Blood Sugar (mmol/L)                      | (3.72-25.37) mmol/L |
|                       | TC                | Total cholesterol (mmol/L)                        | (2.18-9.77) mmol/L |
|                       | LDLC              | Low density lipoprotein (mmol/L)                  | (0.22-6.36) mmol/L |
|                       | HDLC              | High density lipoprotein (mmol/L)                 | (0.26-5.04) mmol/L |
|                       | TG                | Triglyceride (mmol/L)                             | (0.3-26.3) mmol/L |
| **Questionnaire**     | Hypertension      | 0: No; 1: Yes                                    | (0,1)          |
|                       | Diabetes          | 0: No; 1: Yes                                    | (0,1)          |
|                       | Family history of CHD | 0: No; 1: Yes                                    | (0,1)          |
|                       | Smoking           | 0: No; 1: Smoking cessation; 2: Yes               | (0-2)          |
|                       | Alcohol drinking  | 0: No; 1: Alcohol cessation; 2: Yes               | (0-2)          |
|                       | Physical inactivity| 0: No; 1: Yes                                    | (0,1)          |
|                       | Heavy work pressure| 0: No; 1: Yes                                    | (0,2)          |

| Dependent Variable    | Coronary heart disease | 0: No; 1: Yes | (0,1) |

2.2. Algorithms
This study used CatBoost, Random Forest Classifier (RFC) and Logistic Regression (LR) to predict CHD and to identify its risk factors.
CatBoost is an algorithm utilized gradient boosting technique on decision trees. It can handle categorical features and outperform other gradient boosting methods [15]. CatBoost utilizes one-hot-max-size encoding and permutation in converting labels to numbers, greedy methods at new split of trees and target-based statistics. The algorithm has the following steps:

1. Permutating the dataset in a random order;
2. Transforming the labels to integer numbers;
3. Converting the categorical values to the numerical.

RFC combines multiple decision trees and it uses bagging technique, bootstrap feature selection and features randomness, when building trees. The result of RFC comes from the vote of each tree and the result with more votes will become the final result [16-17].

LR, in this study, was the benchmark, which is a classification algorithm. It was developed by Hosmer and Lemeshow [18]. LR uses sigmoid function (logistic function) and presents the value between 0 and 1.

\[
sigmoid \text{ function: } f(x) = \frac{1}{1+e^{-(x)}}.
\]

2.3. Data processing

80% of the data was used as the training set and the remaining 20% as the test set. Ten-fold cross validation was used to test the generalization ability of the model. This study adopted maximum and minimum function to normalize features, \((current-min)/(max-min)\). Easy Ensemble Sampling technique which was proposed by Zhou et al. was used to optimize the results [13].

2.4. Evaluation

This study used confusion matrix and AUROC as evaluation indexes. There were four combination of positive and negative of CHD, true negative (TN): both observed and predicted results were negative; true positive (TP): both observed and predicted results were positive, false negative (FN): negative predicted result and positive observed result; and false positive (FP): positive predicted result and negative observed result. Accuracy, sensitivity, specificity and f-score were utilized to evaluate the results of three algorithms.

Accuracy = \(\frac{TP+TN}{TP+TN+FP+FN}\)
Sensitivity = \(\frac{TP}{TP+FN}\)
Specificity = \(\frac{TN}{TN+FP}\)
f-score = \(\frac{2*precision*recall}{precision+recall}\)

3. Results and Discussions

This study used CatBoost, RFC and LR to build six models, based on total features set and non-laboratory features, respectively. Table 2 and Table 3 illustrated the confusion matrix of six models on the test set.

| Models            | Observed results | Predicted results |
|-------------------|------------------|-------------------|
|                   |                  | Negative | Positive |
| CatBoost          | Negative        | 120      | 26       |
|                   | Positive        | 25       | 118      |
| Logistic regression | Negative   | 114      | 32       |
|                   | Positive        | 40       | 103      |
| Random Forest     | Negative        | 114      | 32       |
|                   | Positive        | 38       | 105      |

Table 2. Confusion matrix of total features.

| Models            | Observed results | Predicted results |
|-------------------|------------------|-------------------|
|                   |                  | Negative | Positive |
| CatBoost          |                  |          |          |
| Logistic regression |                |          |          |
| Random Forest     |                  |          |          |

Table 3. Confusion matrix of non-laboratory features.
Based on these two confusion matrixes, the accuracy of Catboost, LR and RFC of total features set was 82.5%, 75.1% and 75.8%, respectively and that of non-laboratory set was 77.9%, 75.4%, 74.7%, respectively.

|                | Negative | Positive |
|----------------|----------|----------|
| Catboost       | 108      | 38       |
| Logistic       | 110      | 36       |
| Random Forest  | 109      | 34       |

Figure 1. Comparison of different model performance.

Figure 2. The AUROC of different model.

Figure 1 showed the sensitivity, specificity and f-score of the three algorithms in two data sets. Figure 2 demonstrated the ROC of the three algorithms in two data sets. The results illustrated that Catboost had strong diagnostic ability for CHD whether using all features or non-laboratory features. There were few researches applied Catboost in CHD prediction. This study was the first to apply the Catboost algorithm to the diagnosis of CHD worldwide and it performed better than random forest and logistic regression. Moreover, this study was the first to develop a diagnostic model for CHD based on a Chinese population sample database.

CHD related databases have been developed and researchers can evaluate their study methods based on the standardized databases, such as UCI Repository of Machine Learning Database, The Long-Term ST Database and Z-Alizadeh Sani dataset. Alizadehsani et al. utilized Naïve Bayes, C4.5 AdaBoost and Sequence minimum optimization algorithm, with 16 features, including laboratory data and echocardiography, to predict CHD. The results showed that Sequence minimum optimization algorithm had the best accuracy of 82%, and ejection fraction and segmental wall motion abnormalities are the most important features for the diagnosis of CHD [19]. Tsipouras proposed an automatic diagnosis system, an improved decision tree model, of coronary artery disease, based on features of demographics,
questionnaires, blood tests and the evaluation index of arterial stiffness (PWV and ABI). The model had 73.4% accuracy, 80% sensitivity and 65% specificity [20]. Those studies proved that machine learning algorithms can perform well in CHD diagnosis. Different database, features, experimental designs and algorithms may have different results, and as a result, the results from different researches should not be simply compared.

The current study not only researched the applicability of the algorithms, but also considered the possibility of promotion of this intelligent CHD diagnosis. Most previous studies used complex features, such as Echocardiography, chest pain symptoms and exercise testing [21-22], which are high-cost, intrusive and need professional workers and equipment. Moreover, some of them have risks in testing process and some cannot be collected from the disabled. Thus, these features are not appropriate to promote in large population, especially in poor areas. The features used in current study are easily to collect and low cost, even free. In China, blood indicators, including fasting blood sugar and serum lipids, are routine inspection items and can be performed at the primary community health service centers. In addition, according to Chinese National Basic Public Health Service Specification, free blood indicators test is available for people over the age of 65, who happen to be at high risk of coronary heart disease. Rest features in this study can be collected from questionnaires. All these make it possible to apply the diagnostic model of CHD established in this study.

The importance of features in Catboost, which had the best performance in this study, was shown in Figure 3. In total features set, age contributed the most, followed by total cholesterol and family history of CHD. Aging has been proved to correlated with Various cardiovascular diseases. With aging, blood vessels are aging, intima thickening, elastin declining, arterial elasticity decreasing, atherosclerosis gradually forming, and blood vessels becoming blocked, eventually leading to CHD [23]. Abnormal lipid metabolism is one of the most important risk factors for CHD. There is a dose-response relationship between total cholesterol and the risk of coronary heart disease [24]. A large number of studies have confirmed that family history is an independent risk factor for CHD [25-26]. The family history of CHD not only increases the risk, but also advances the age of onset of coronary heart disease [27]. Among all non-laboratory features, age, BMI and systolic blood pressure were the three most important features for the diagnosis of CHD. The role of blood pressure in coronary heart disease has been demonstrated in numerous studies. BMI is not only an independent risk factor for CHD, but also can mediate CHD by increasing blood pressure, worsening blood glucose and inducing metabolic syndrome [28]. The order of importance of variables in the model is consistent with the results of epidemiological studies related to CHD, which also indicates that the diagnostic model of CHD constructed in this study is reasonable.

4. Conclusion

Based on the results and discussions presented above, the conclusions are obtained as below:

(1) It is shown that Catboost algorithm has the best performance of CHD diagnosis, compared to random forest and logistic regression.

(2) The diagnostic model of CHD based on Catboost algorithm can reach high accuracy, sensitivity and specificity.

(3) It is concluded that age, total cholesterol and family history of CHD were the three most important risk factors for identification of coronary heart disease.

Acknowledgments

The authors thank all the participants for their important contributions to this study. Zuchang Ma put forward the concept of the study; Xiaoyu Zhang designed the study, analyzed the data and prepared this manuscript; Wei Wei, Shijun Wang, Yang Xu, Lisheng Gao participated in data collection; Mu Wang modified the grammar of the manuscript. All authors commented on all the drafts and approved the final version.

This research was supported by National Key Research and Development Program (grant 2017YFB1002200).
References

[1] Naghavi, M., Wang, HD., Lozano, R., Davis, A., Liang, XF., Zhou, MG. et al. (2015) Global, regional, and national age-sex specific all-cause and cause-specific mortality for 240 causes of death, 1990-2013: a systematic analysis for the Global Burden of Disease Study. Lancet, 385:117-71.

[2] Bonow, R.O., Mann, D.L., Zipes D.P., Libby, P. (2012) Heart Disease: A Textbook of Cardiovascular Medicine (9th ed.), Saunders, New York.

[3] Levin, DC., Gardiner, GA. (1987) COMPLEX AND SIMPLE CORONARY-ARTERY STENOSES - A NEW WAY TO INTERPRET CORONARY ANGIOGRAMS BASED ON MORPHOLOGICAL FEATURES OF LESIONS. Radiology, 164: 675-80.

[4] Guedes, A., Keller, PF., L'Allier, PL., Lesperance, J., Gregoire, J., Tardif, JC. (2005) Long-term safety of intravascular ultrasound in nontransplant, nonintervened, atherosclerotic coronary arteries. J Am Coll Cardiol, 45:559-64.

[5] Mehta, SK., McCrary, JR., Frutkin, AD., Dolla, WJS., Marso, SP. (2007) Intravascular ultrasound radiofrequency analysis of coronary atherosclerosis: an emerging technology for the assessment of vulnerable plaque. Eur Heart J, 28:1283-8.

[6] Weng, SF., Reps, J., Kai, J., Garibaldi, JM., Qureshi, N. (2017) Can machine-learning improve cardiovascular risk prediction using routine clinical data? PLoS One, 12:14.

[7] Amin, MS., Chiam, YK., Varathan, KD. (2019) Identification of significant features and data mining techniques in predicting heart disease. Telemat Inform, 36:82-93.

[8] Abdar, M., Yen, NY ., Hung, JCS. (2018) Improving the Diagnosis of Liver Disease Using Multilayer Perceptron Neural Network and Boosted Decision Trees. J Med Biol Eng, 38:953-65.

[9] Heydari, M., Teimouri, M., Heshmati, Z., Alavinia, SM. (2016) Comparison of various classification algorithms in the diagnosis of type 2 diabetes in Iran. Int Diabetes Dev Ctries, 36:167-73.

[10] Amato, F., Lopez, A., Pena-Mendez, EM., Vanhara, P., Hampl, A., Havel, J. (2013) Artificial neural networks in medical diagnosis. J Appl Biomed, 11:47-58.

[11] Rajesh, K., Uhuli, R. (2017) Classification of ECG heartbeats using nonlinear decomposition methods and support vector machine. Comput Biol Med, 87:271-84.

[12] Krittanawong, C., Zhang, HJ., Wang, Z., Aydar, M., Kitai, T. (2017) Artificial Intelligence in Precision Cardiovascular Medicine. J Am Coll Cardiol, 69:2657-64.

[13] Marateb, HR., Goudarzi, S. (2015) A noninvasive method for coronary artery diseases diagnosis using a clinically-interpretable fuzzy rule-based system. Journal of Research in Medical Sciences the Official Journal of Isfahan University of Medical Sciences, 20:214-24.

[14] Alizadehsani, R., Zangooee, MH., Hosseini, MJ., Habibi, J., Khosravi, A., Roshanzamir, M., et al. (2016) Coronary artery disease detection using computational intelligence methods. Knowledge-Based Syst, 109:187-97.

[15] Prokorenkova, L., Gusev, G., Vorobev, A., Dorogush, AV., Gulin, A. (2017) CatBoost: unbiased boosting with categorical features.

[16] Ho, Tin Kam. (1995). Random Decision Forests. In: Proceedings of the 3rd International Conference on Document Analysis and Recognition. Montreal, QC. pp. 278–282.

[17] Amit, Y., Geman, D. (1997) "Shape quantization and recognition with randomized trees". Neural Computation, 9 (7): 1545–1588.

[18] David, W., Hosmer, Jr. (2000) Applied Logistic Regression (2nd ed.). Wiley.

[19] Alizadehsani, R., Habibi, J., Sani, ZA., Mashayekhi, H., Bahadorian, B. (2012) Diagnosis of Coronary Artery Disease Using Data Mining Based on Lab Data and Echo Features.

[20] Tsipouras, MG., Exarchos, TP., Fotiadis, DI., Kotsia, AP., Vakalis, KV., Naka, KK., et al. (2008) Automated diagnosis of coronary artery disease based on data mining and fuzzy modeling. IEEE T Inf Technol Biomed, 12:447-58.

[21] Verma, L., Srivastava, S., Negi, PC. (2016) A Hybrid Data Mining Model to Predict Coronary Artery Disease Cases Using Non-Invasive Clinical Data. Journal of Medical Systems, 40:178.

[22] Acharya, UR., Sudarshan, VK., Koh, JEW., Martis, RJ., Tan, JH., Oh, SL., et al. (2017) Application of higher-order spectra for the characterization of Coronary artery disease using
electrocardiogram signals. Biomed Signal Process Control, 31:31-43.

[23] Lakatta, EG. (2002) Age-associated cardiovascular changes in health: impact on cardiovascular disease in older persons. Heart failure reviews, 7:29-49.

[24] Peters, SAE., Singhateh, Y., Mackay, D., Huxley, RR., Woodward, M. (2016) Total cholesterol as a risk factor for coronary heart disease and stroke in women compared with men: A systematic review and meta-analysis. Atherosclerosis, 248:123-31.

[25] Pandey, AK., Pandey, S., Blaha, MJ., Agatston, A., Feldman, T., Ozner, M., et al. (2013) Family history of coronary heart disease and markers of subclinical cardiovascular disease: Where do we stand? Atherosclerosis, 228:285-94.

[26] Scheuner, MT. (2001) Genetic predisposition to coronary artery disease. Curr Opin Cardiol, 16:251-60.

[27] Kral, BG., Becker, DM., Vaidya, D., Yanek, L.R., Becker, LC. (2012) Severity of inducible myocardial ischemia predicts incident acute coronary syndromes in asymptomatic individuals with a family history of premature coronary artery disease. Journal of Nuclear Cardiology, 19:28-36.

[28] Park, J., Kim, SH., Cho, GY., Baik, I., Kim, NH., Lim, HE., et al. (2011) Obesity phenotype and cardiovascular changes. Journal of Hypertension, 29:1765-72.