Improving the Detection of Acute Coronary Syndrome Using Machine Learning of Blood Biomarkers

Khaled Rjoob[1], Victoria McGilligan[2], Raymond Bond[1], Steven Watterson[2], Melody Chemaly[2], Roisin McAlister[2], Tiago De Melo Malaquias[2], Stephen J Leslie[3], Charles Knoery[3], Aleeha Iftikhar[1], Anne McShane[4], Anthony Bjourson[2], Aaron Peace[5]

[1]Faculty of Computing, Engineering & Built Environment, Ulster University, Northern Ireland, UK.
[2] Centre for Personalised Medicine/Northern Ireland Centre for Stratified Medicine, Ulster University, Northern Ireland, UK
[3]Department of Diabetes & Cardiovascular Science, University of the Highlands and Islands, Centre for Health Science, Inverness, UK
[4]Emergency Department, Letterkenny University Hospital, Donegal, Ireland
[5]Western Health and Social Care Trust, C-TRIC, Ulster University, Northern Ireland, UK

Abstract

Background: Acute coronary syndrome (ACS) is one of the main causes of death worldwide. The 12-lead electrocardiogram (ECG) is used to help diagnose ACS, along with clinical risk factors (smoking, diabetes mellitus, hypertension, hscTn and positive family history of ACS. These methods however are associated with many limitations resulting in variable sensitivity/specificity. The aim of this study was to use a machine learning approach to develop an optimum panel of blood protein biomarkers capable of independently diagnosing ACS. Methods: A hybrid feature selection and ML prediction algorithms including two classifiers: 1) decision tree (DT) and 2) logistic regression were applied to protein biomarkers (327 proteins) collected from patients with ACS n=91 or non-ACS n=97. Results: Using this approach, 20 proteins out of 327 proteins were able to accurately distinguish between ACS and non-ACS (logistic regression ROC-AUC=0.8 and accuracy=82.5% and DT ROC-AUC=0.6 and accuracy=64.9%. Conclusion: Logistic regression obtained a higher performance compared to DT and showed promising results uncovering a panel of 20 protein biomarkers which included those associated with progressive atherosclerotic plaques, myocardial injury and inflammation. This approach was able to accurately discriminate between patients with ACS and non-ACS.

1. Introduction

Acute coronary syndrome (ACS) refers to a group of conditions associated with acute interruption in coronary blood flow. Either partial or complete interruption results in the development of either unstable angina (UA), or a non—ST-segment elevation myocardial infarction (NSTEMI), or a ST-segment elevation myocardial infarction (STEMI). Plaque rupture within the coronary artery in the setting of coronary atherosclerosis is generally the mechanism behind ACS and remains to be the primary cause of death globally [1]. ACS is a huge burden to emergency services costing around the £19 billion annually in the NHS [2]. Detection of ACS is highly subjective, and its accuracy depends on clinical skills of health care professionals. ACS diagnosis relies on the clinical history, the 12 lead ECG and elevation in the troponin the most widely used serum biomarker. ECG remains however to be the cornerstone of diagnosis however changes to aid with early diagnosis are frequently nonspecific or even absent in around 40% of the patients [2].

The inability of the ECG to identify these patients may lead to diversion to the wrong clinical pathway and can lead to delay in definitive management. The latter can ultimately culminate in avoidable mortality. ECG mimics of ACS where ST-segment changes are observed occur in other cardiac conditions such as pericarditis, myocarditis left ventricular hypertrophy, cardiomyopathies and channelopathies, adding to the diagnostic dilemma [3-4].

Machine learning (ML) encompasses a wide variety of methods whereby artificial intelligence learns to perform tasks when exposed to large amounts of data. The application of a ML algorithm may expedite management and identification of ACS for more accurate diagnosis, appropriate signposting to the correct clinical pathway and early initiation of appropriate ACS pharmacotherapy [5-10].

The aim of this study was therefore to apply ML on a

retrospective dataset from 198 patients who had 367 plasma protein biomarkers available for analysis. We aimed to discover biomarkers that could improve the performance of discriminating between ACS and non-ACS.

2. Method

2.1. Data collection

The dataset was collected at the cardiac catheterization laboratory at Altnagelvin hospital, Northern Ireland. In the dataset, 327 protein biomarkers were acquired from two different groups: group 1) acute coronary syndrome (ACS) patients (n=91) with ST-elevation myocardial infarction (STEMI), non-ST-elevation myocardial infarction (NSTEMI) or unstable angina (UA), and group 2) non-ACS patients (n=97).

2.2. Feature selection

A hybrid feature selection approach was applied which combines the filter method and the wrapper method to find an optimal set of proteins that provide good classification results. In the filter method, the 327 proteins were ranked from the most important to the least important using joint mutual information (JMI) filter in equation 1. Then in the wrapper method, the backward elimination method was applied on the ranked features to find the final optimal set of features.

$$f_t = \arg\max I(x_i; y) - \left[\alpha \sum_{k=1}^{t-1} I(x_{fk}; x_i) - \beta \sum_{k=1}^{t-1} I(x_{fk}; x_i|y) \right] \quad (1)$$

where x represents features and y represents labels

In JMI: $\alpha = \frac{1}{t-1}$ and $\beta = \frac{1}{t-1}$

2.3. Classification

Two machine learning (ML) classifiers were applied including 1) decision tree (DT) and 2) logistic regression (LOG).

3. Results

Using feature selection, 20 proteins (including those associated with progressive atherosclerotic plaques, myocardial injury and inflammation) out of 327 proteins were selected to distinguish between ACS and non-ACS using logistic regression and DT using 5-fold cross-validation. While the other 307 proteins were excluded. DT obtained 64.9% [95%CI=0.52,0.77] accuracy, 0.59 [95%CI=0.47,0.72] area under the curve (AUC), 70.4% sensitivity and 60.0% specificity. While LOG obtained 82.5% [95%CI=0.72,0.92] accuracy, 0.89 [95%CI=0.69,0.9] AUC, 85.2% sensitivity and 80.0%

specificity. Table 1 below shows the performance of the two ML classifiers (DT, LOG).

Figure 1 presents the confusion matrix of each ML classifier.

Table 1. Machine learning classifiers accuracy.

ML	ACCURACY	AUC	SE	SP
DT	64.9%	0.59	70.4%	60.0%
LOG	82.5%	0.89	85.2%	80.0%

ML: Machine learning, DT: Decision tree, LOG: Logistic regression, AUC: Area under the curve, SE: Sensitivity, SP: Specificity.
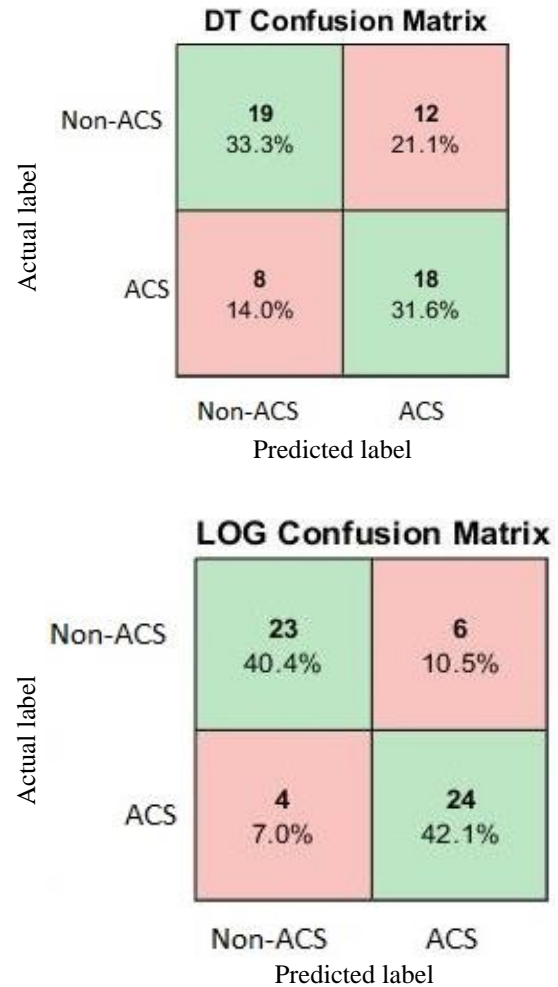

Figure 1. The confusion matrix for DT and LOG. Accuracy= (True Positive + True Negative)/(True Positive + True Negative+ False Positive + False Negative).

DT obtained 0.59 AUC while LOG obtained 0.89 AUC as shown in Figure 2. According to the results, LOG obtained the best performance compared to DT with respect to accuracy, area under the curve (AUC), sensitivity and specificity. And there is a significant difference between DT and LOG (p<0.03) in term of performance.
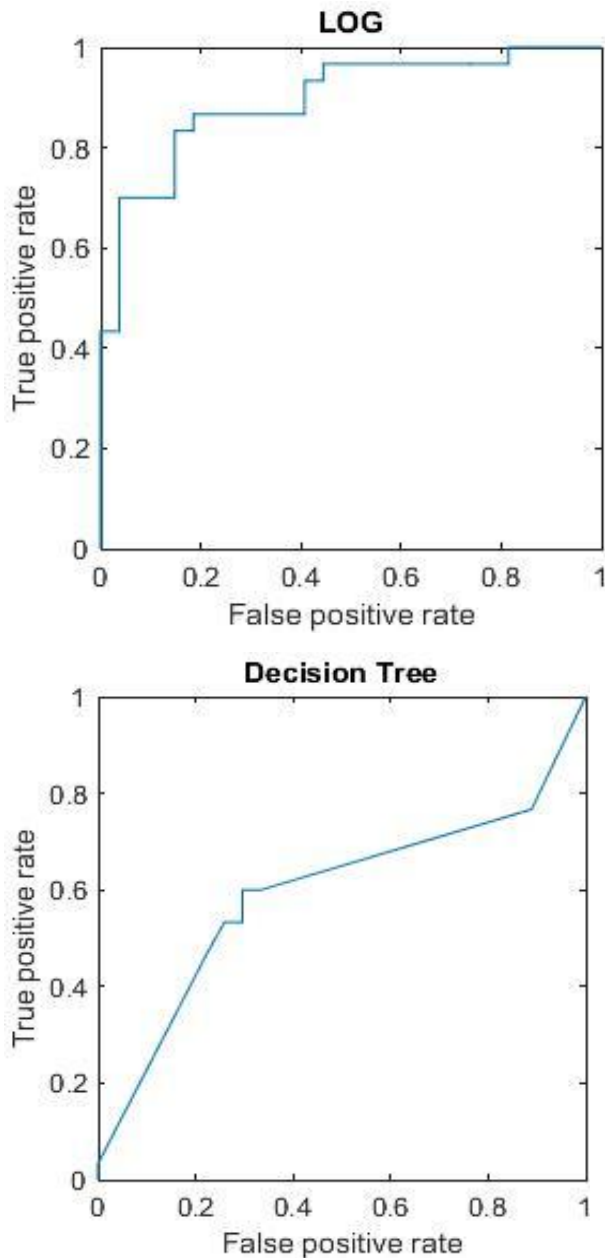

Figure 2. ROC curve for each ML classifier. DT AUC=0.59, LOG AUC=0.89

4.   Discussion

This research presents a ML approach that effectively distinguishes between ACS and non-ACS patients utilizing a panel of 20 proteins, with the potential to improve clinical diagnostics of this patient population: 1) ACS patients (STEMI, NSTEMI and UA) and 2) non-ACS by detecting a group of novel proteins that could improve clinical diagnosis. Using the feature selection method JMI, we prioritized a group of proteins (n=20) as the most important proteins. These proteins have established roles in progressive atherosclerotic plaques, myocardial injury and inflammation. In ML classifiers, the Logistic Regression classifier obtained the best performance to discriminate between the aforementioned patients as those two groups of patients are overlapped according to symptoms using 20 proteins. While previous studies obtained 0.81 AUC using LOG to detect ACS without feature selection [13], suggesting that the feature selection algorithm could improve the ML performance. Furthermore, we analyzed a large proteomic dataset, that provided the opportunity to assess a wider dataspace and to identity additional novel protein biomarkers. However, new ML classifiers and feature selection methods could be included in future work to improve the ML performance and to improve clinical decision making in cardiac care. Furthermore, combing those proteins with ECGs biomarkers could also improve the performance.

4.   Conclusion

Here we report that logistic regression shows promising results for discriminating between ACS and non-ACS. JMI improves LOG performance by reducing the number of features to the optimum number which improves accuracy, reduces training and testing time and avoids overfitting. Based on the ML and feature selection results, future work should be carried out to improve the performance by increasing the dataset size, including new feature selection approaches and using new machine learning classifies such as long-short term memory network (LSTM) and deep artificial neural networks (ANNs).

References

[1] WHO, Cardiovascular diseases facts sheet. World Health Organization, 2017. https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds).

[2] BHF, factsheet. british heart foundation, 2020.https://www.bhf.org.uk/what-we-do/our research/heart-statistics.

[3] EXpert Group on Biomarkers (2014). Biomarkers in cardiology--part 1--in heart failure and specific cardiomyopathies. Arquivos brasileiros de cardiologia, 103(6), 451–459. https://doi.org/10.5935/abc.20140184

[4] Gaetano Thiene, et al. Revisiting definition and classification of cardiomyopathies in the era of molecular medicine, European Heart Journal, Volume 29, Issue 2, January 2008, Pages 144–146, https://doi.org/10.1093/eurheartj/ehm585

[5] Kumar A, Cannon CP. Acute coronary syndromes: diagnosis and management, part I. Mayo Clin Proc. 2009;84(10):917–938. doi:10.1016/S0025-6196(11)60509-0

[6] Boersma E, Pieper KS, Steyerberg EW, Wilcox RG, Chang W-C, Lee KL, Akkerhuis KM, Harrington RA, Deckers JW, Armstrong PW, Lincoff AM, Califf RM, Topol EJ, Simoons ML. Predictors of outcome in patients with acute coronary syndromes without persistent STsegment elevation. Results from an International Trial of 9461 patients. The PURSUIT Investigators. Circulation 2000;101:2557e2567.

[7] Sakamoto JT, Liu N, Koh ZX, Fung NX, Heldeweg ML, Ng JC, Ong ME. Comparing HEART, TIMI, and GRACE scores for prediction of 30-day major adverse cardiac events in high acuity chest pain patientsin the emergency department. Int J Cardiol 2016;221:759e764.

[8] Awad M., Khanna R. (2015) Machine Learning. In: Efficient Learning Machines. Apress, Berkeley, CA, https://doi.org/10.1007/978-1-4302-5990-9_1

[9] McDonald L, Ramagopalan SV, Cox AP, Oguz M. Unintended consequences of machine learning in medicine?. F1000Res. 2017;6:1707. Published 2017 Sep 19. doi:10.12688/f1000research.12693.1.

[10] McCarthy CP, et al. A clinical and proteomics approach to predict the presence of obstructive peripheral arterial disease: From the Catheter Sampled Blood Archive in Cardiovascular Diseases (CASABLANCA) Study. Clin Cardiol. 2018;41(7):903–909. doi:10.1002/clc.22939.

[11] Carles Forné, Serafi Cambray, Marcelino Bermudez-Lopez, Elvira Fernandez, Milica Bozic, Jose M Valdivielso, NEFRONA investigators, Machine learning analysis of serum biomarkers for cardiovascular risk assessment in chronic kidney disease, Clinical Kidney Journal, , sfz094, https://doi.org/10.1093/ckj/sfz094.

[12] Bharath Ambale-Venkatesh, et al. Cardiovascular event prediction by machine learning. Circulation Research, Vol.121, pp.1092-1101, 13 October 2017. https://doi.org/10.1161/CIRCRESAHA.117.311312

[13] VanHouten JP, et al. Machine learning for risk prediction of acute coronary syndrome. AMIA Annu Symp Proc. 2014;2014:1940-1949. Published 2014 Nov 14.

Address for correspondence:
Khaled Rjoob
Newtownabbey, BT37 0QB, Northern Ireland, UK.
E-mail: Rjoob-k@ulster.ac.uk