Improving the detection of acute coronary syndrome using machine learning of blood biomarkers

Khaled Rjoob, MSc, Victoria McGilligan, PhD, Raymond Bond, PhD, Steven Watterson PhD, Melody Chemaly PhD, Roisin McAlister PhD, Tiago De Melo Malaquias PhD, Stephen J Leslie, PhD, Charles Knoery, MBChB, Aleeha Iftikhar, MSc, Anne McShane, MSc, Anthony Bjourson, PhD, Aaron Peace, PhD

1Faculty of Computing, Engineering & Built Environment, Ulster University, Northern Ireland, UK.
2Centre for Personalised Medicine/Northern Ireland Centre for Stratified Medicine, Ulster University, Northern Ireland, UK
3Department of Diabetes & Cardiovascular Science, University of the Highlands and Islands, Centre for Health Science, Inverness, UK
4Emergency Department, Letterkenny University Hospital, Donegal, Ireland
5Western 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, hsTn and positive family history of ACS). These methods however are associated with many limitations resulting in variable sensitivity/specificity. There is a clear need for the development of new methods to improve ACS diagnosis. 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 (wrapper method and filter method) and ML prediction algorithms including two classifiers: 1) decision tree (DT) and 2) logistic regression were applied to protein biomarker (367 proteins) collected from patients with ACS n=91 or stable angina n=97. Results: Using this approach, 20 proteins out of 367 proteins were able to accurately distinguish between ACS and stable angina (logistic regression ROC-AUC=0.8 [95%CI=0.69,0.9] and accuracy=82.5% [95%CI=0.72,0.92] and DT ROC-AUC=0.6 [95%CI=0.47,0.72] and accuracy=64.9% [95%CI=0.52,0.77]) using 5-fold cross-validation. 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 stable angina. Method validation is now required utilising other ML algorithms and new feature selection methods to confirm performance.

1. Introduction

Acute Coronary Syndrome (ACS) refers to any group of clinical symptoms associated with acute myocardial ischemia and includes unstable angina (UA), non—ST-segment elevation myocardial infarction (NSTEMI), and ST-segment elevation myocardial infarction (STEMI). These high-risk manifestations of coronary atherosclerosis, a major cardiovascular disease (CVD) is the primary cause of death globally [1]. It also results in high usage of emergency services and is reported to cost the UK healthcare service £19 billion annually [2]. Detection of ACS is highly subjective, and its accuracy remains very dependent on clinical skills of the health care professionals. ACS diagnosis has traditionally relied upon the 12 lead ECG in combination with ischemic symptoms and elevation in serum biomarkers. However, symptoms are often atypical or absent, and around 33% of patients that present with ACS may not have chest pains [1]. Similarly, ECG changes that aid with early diagnosis may be nonspecific or even absent in around 40% of the patients [2].

The inability of the ECG to identify these patients may lead to delay in their proper management. Moreover, ST-segment changes are also observed in other cardiac conditions like pericarditis, left ventricular hypertrophy, cardiomyopathies and channelopathies, which can add 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 of ACS for either early discharge or early initiation of ACS pharmacotherapy [5-10].

The aim of this study was therefore to apply ML on a retrospective dataset with information on 198 patients who had 367 serum protein biomarker features available for analysis. We aimed to discover biomarkers that could improve the performance of discriminating between (STEMI, NSTEMI and UA) and stable angina.

2. Method

2.1. Data collection

The dataset was collected at the cardiac catheterization laboratory at Altnagelvin hospital, Northern Ireland from 2015 to 2017. In the dataset, 367 biomarkers were acquired from two different groups: group 1) acute coronary syndrome (ACS) patients (n=91) who have ST-elevation myocardial infarction (STEMI), non-ST-elevation myocardial infarction (NSTEMI) or unstable angina (UA), and group 2) stable angina (SA) 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 367 proteins were ranked from the most important to the least important using joint mutual information (JMI) 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.

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

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 (included those associated with progressive atherosclerotic plaques, myocardial injury and inflammation) out of 367 proteins were selected to distinguish between ACS and stable angina using logistic regression and DT using 5-fold cross-validation. while the other 347 proteins were excluded. Table 1 below shows the performance of two ML classifiers (DT, LOG).

Figure 1 below shows the confusion matrix of each ML classifier. Figure 2 shows ROC-curve for each classifier. 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.

Table 1. Machine learning classifiers accuracy.

<table>
<thead>
<tr>
<th>ML</th>
<th>ACCURACY</th>
<th>AUC</th>
<th>SE</th>
<th>SP</th>
</tr>
</thead>
<tbody>
<tr>
<td>DT</td>
<td>64.9%</td>
<td>0.59</td>
<td>70.4%</td>
<td>60.0%</td>
</tr>
<tr>
<td>LOG</td>
<td>82.5%</td>
<td>0.89</td>
<td>85.2%</td>
<td>80.0%</td>
</tr>
</tbody>
</table>

ACS patients (STEMI, NSTEMI and UA) and 2) stable angina by detecting a group of novel proteins that could improve the clinical diagnosis in the cardiac care unit. 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 study obtained 0.81 AUC using LOG to detect ACS without feature selection [13], which means 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 identify more novel 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.

4. Conclusion

Here we report that logistic regression shows promising results for discriminating between ACS and SA. 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).

Acknowledgments

This work is supported by the European Union’s INTERREG VA programme, managed by the Special EU Programmes Body (SEUPB). The work is associated with the project – ‘Centre for Personalised Medicine – Clinical Decision Making and Patient Safety’. The views and opinions expressed in this study do not necessarily reflect those of the European Commission or the Special EU Programmes Body (SEUPB).

References


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