Informatics in Medicine Unlocked 13 (2018) 26–33

Contents lists available at ScienceDirect

Informatics in Medicine Unlocked

journal homepage: www.elsevier.com/locate/imu

Multiclass classification of myocardial infarction with convolutional and recurrent neural networks for portable ECG devices

Hin Wai Lui[a,b,*], King Lau Chow[a,c]

[a] Interdisciplinary Programs Office, Hong Kong University of Science and Technology, Hong Kong
[b] Technology, Leadership and Entrepreneurship Program, School of Engineering, Hong Kong University of Science and Technology, Hong Kong
[c] Division of Life Science, School of Science, Hong Kong University of Science and Technology, Hong Kong

ARTICLE INFO

Keywords:
Myocardial infarction
Electrocardiogram
Deep learning
Convolutional neural network
Recurrent neural network
Classification

ABSTRACT

Myocardial infarction (MI) is a medical emergency for which the early detection of symptoms is desirable. The prevalence of portable electrocardiogram (ECG) devices makes frequent screening for MI possible. In this study, we develop an MI classifier that combines both convolutional and recurrent neural networks, and is suitable for wearable ECG devices with only a single lead recording. It performs multiclass classification to discriminate the ECG records of MI from those of healthy individuals and patients with existing chronic heart conditions, as well as ECG records contaminated with noise. The method was tested on a dataset with MI ECG records and compared with a pure convolutional neural network and classifier with hand-crafted features. It was found that the addition of a recurrent layer improved the classification sensitivity by 28.0% compared to the convolutional neural network alone. Overall, it achieved 92.4% sensitivity, 97.7% specificity, a 97.2% positive predictive value, and a 94.6% F1 score.

1. Introduction

Myocardial infarction (MI), more commonly known as a heart attack, is a medical emergency requiring immediate attention [1]. During an MI episode, blood flow to the heart tissues is disrupted owing to full or partial blockage of the coronary arteries. Without a full supply of oxygen and nutrients (ischemia), the heart tissues surrounding the blocked coronary artery die, and such a process is often irreversible [2]. With heart tissue death, the normal heart conduction pathways are altered, causing fatal arrhythmias such as ventricular fibrillation, thereby leading to sudden cardiac arrest and cardiac sudden death [3].

Numerous severe symptoms may be experienced by the patient during an MI episode, such as loss of consciousness, chest pain, and shortness of breath [4]. However, many patients only experience mild symptoms or none at all, which is often described as a silent heart attack. It is estimated that between 22% and 64% of all MIs are silent [5].

It is important that silent MIs be discovered and treated at the earliest onset. MI may be diagnosed by examining the electrocardiogram (ECG) [6]. The ECG records the electrical signal generated by the heart sinus node to coordinate the contraction and relaxation of different heart chambers in a sequential and synchronized manner. In general, a healthy individual has an ECG with five distinctive waveforms for each cardiac cycle: P, Q, R, S and T waves [7]. With damaged heart tissues during an MI episode, distinctive ECG morphology can be observed. Observations of a long ST interval, ST elevation and other changes in T-wave morphology are several diagnostic indicators of MI [8].

Traditionally, ECG devices are bulky machines that are only available at hospitals and clinics. In recent years, portable ECG devices have become available in various forms, such as chest patches, phone cases and smart watches [9], which make regular ECG recording possible. With the increasing volume of data generated by portable ECG devices, automatic detection of MI by a computer system, without relying solely on trained physicians, is preferable.

Various methods have been proposed and developed to achieve automatic MI detection from ECG records [10–24]. The majority of these methods extract the relevant features from the ECG morphology that are indicative of MI, such as the ST interval, ST amplitude and RR interval [12,14,16,17,20,24]. Following feature extraction, classifiers are developed to discriminate between MI and healthy ECG records. These classifiers are either based on simple thresholds [22,25], or implemented by training a machine learning model from a database of MI and healthy ECGs, such as support vector machines [13,17,23,25], shallow neural networks [11,15,22], the random forest [17], and K-nearest neighbours [20]. Numerous methods use different signal processing and transformation techniques to enhance the classification

[*] Corresponding author. Interdisciplinary Programs Office, Hong Kong University of Science and Technology, Hong Kong.
E-mail address: hwluiaa@connect.ust.hk (H.W. Lui).

https://doi.org/10.1016/j.imu.2018.08.002
Received 25 June 2018; Received in revised form 9 August 2018; Accepted 12 August 2018
Available online 24 August 2018

performance, such as wavelet transform [14,22–24], principle component analysis [16,17], ordinary differential equations [27], entropy analysis [17] and polynomial approximation [12].

The majority of these machine learning-based methods have trained their models with the PTB diagnostic database [28], available on Physionet [29]. This database comprises ECG records of MI, healthy individuals and seven other cardiovascular diseases (CVDs). Many of these methods have trained a binary classifier to discriminate between MI and healthy ECGs, and certain approaches have achieved very high performance, with 99.97% sensitivity and 99.90% specificity [20].

While some patients who experience MI could be assumed to be healthy prior to the MI episode, many had pre-existing chronic heart conditions [30]. Therefore, an effective MI classifier must be capable of discriminating MI from other forms of CVD, in addition to a healthy ECG. Otherwise, it is likely that a patient with another form of CVD will be classified as MI by a binary classifier between MI and healthy ECG.

With portable ECG devices, it is often not possible to obtain the full 12-lead ECGs used in hospitals or clinics. Most ECG chest patches only provide two to three leads, and most hand-held ECG devices only provide lead I [9]. Therefore, an effective MI classifier based on such devices must also function with a limited number of ECG leads. Moreover, for ease of regular use and comfort, portable ECG devices often do not use wet electrodes, resulting in a weak and noisy ECG signal owing to the high impedance of dry skin. As these devices are designed to be integrated into the everyday life of the user, the ECG may not be recorded in a stationary condition, rendering the recording unusable owing to motion artefacts. Therefore, a robust MI classifier must be capable of discriminating between valid and noisy ECG records.

In recent years, deep learning techniques have demonstrated far superior classification accuracy compared to traditional shallow machine learning techniques for numerous tasks [31]. One benefit of the deep learning technique is that features are learned by deep neural networks automatically, without the need for extracting hand-crafted features by human experts [32]. Convolutional neural networks (CNNs) have been used extensively for image classification, as the convolutional layers can effectively encode spatial information [31]. Several authors have experimented on applying CNNs to ECG classification, treating the ECG signal as a 1D image [10,16,33–35]. These methods divide an ECG record into short segments of a few seconds, or by individual heartbeats based on the position of the QRS complex that constitutes the Q, R, and S waves.

It has been found that certain CNN-based methods for MI classification do not achieve satisfactory results, and are sometimes inferior to traditional shallow machine learning methods with appropriate hand-crafted features and signal transformation [12,26]. A possible explanation is that these CNN methods only individually classify a short segment or single heartbeat of ECG, while optimal shallow machine learning methods extract features from the entire ECG record of many heartbeats, including those derived from beat-to-beat variations, such as heart rate variability [11]. Therefore, the approach taken by many CNN classifiers may not be able to utilise the beat-to-beat heart rate variations and ECG waveform morphology caused by MI fully.

In this study, we have developed a CNN-based MI classifier designed to overcome the limitations of portable ECG devices and shortcomings of previous CNN methods. Firstly, as opposed to solely classifying individual ECG heartbeat segments, recurrent layers were added to analyse the beat-to-beat variations of the ECG morphology after the convolutional layers. Secondly, instead of performing a binary classification of MI or healthy ECGs, a multiclass classifier was trained with the addition of the “other” class of other CVD types and the “noisy” class for noisy signals.

2. Materials and methods

2.1. Materials

Two separate databases from Physionet [29] were used for training and testing our classifiers: the Physikalisch-Technische Bundesanstalt (PTB) diagnostic ECG database [28] and the AF classification from a short single lead ECG recording: Physionet/computing in cardiology challenge 2017 database (AF-Challenge) [36]. The PTB database contained 549 records from 290 subjects, with 209 males and 81 females. The mean age was 57.2 for males and 61.6 for females. Of the 290 subjects, 148 were labelled as MI with 368 records, 52 were labelled as healthy with 80 records, and the remainder had labels of heart failure, bundle branch block, dysrhythmia, myocardial hypertrophy, valvular heart disease and myocarditis. We divided the PTB database into three classes: “MI”, “healthy” and “other” for all other classes. The AF-Challenge database contained 8528 records from an unknown number of subjects. We included 278 noisy signal records from the AF-Challenge database as the “noisy” class.

The PTB database contained 15 simultaneously measured ECG leads (I, II, III, avr, avl, avf, v1, v2, v3, v4, v5 and v6), together with the three Frank lead ECGs (vx, vy and vz). The sampling frequency of the PTB database was set to 1000 Hz. The AF-Challenge database contained only single-lead measurements by a handheld ECG device in a lead I configuration, and the sampling frequency was set to 300 Hz. As the goal of this study was to develop an MI classifier for portable ECG devices, only lead I from the PTB database was used, and the signal was down-sampled to 300 Hz in order to match the AF-Challenge database sampling frequency.

All data samples were collected from the relevant classes of the above two databases. No data was collected from human subjects and hence ethics committee approval was not required.

2.2. Signal pre-processing

As the ECG records were noisy, denoising was performed using a Savitzky Golay filter with a frame size of 13 and polynomial order of 3 [37], equivalent to a cut-off frequency at 82Hz [38]. Following denoising, the baseline wander was extracted using a larger Savitzky Golay filter with a frame size of 1201 and polynomial order of 4, equivalent to a cut-off frequency at 0.78 Hz. Thereafter, the baseline wander was subtracted from the signal. All filter parameters were selected by a grid search to yield the best classification performance. Denoising and baseline wander removal were performed on all samples including the noisy class to ensure a fair comparison for the testing set.

After noise and baseline wander removal, to effectively utilise the ability of CNN to extract meaningful features from ECG morphology, a long ECG record was segmented into its constituent heartbeats at the QRS complexes using an algorithm with a new nonlinear transform and first-order Gaussian [39]. As CNN requires fixed-length inputs, but each heartbeat has a variable length depending on the heart rate, zero-padding was performed so that all heartbeat samples had a fixed length of 512 samples. This is equivalent to 1707 ms, and heartbeats with a rate as low as 35 beats per minute could still fit into this window.

Input normalisation is a technique that normalises the varying scales and dynamic ranges of different samples, and has been demonstrated to improve classification performance [40]. For all records, we shifted and scaled the signal from between -2500 and 2500 to 0 and 1.

All the above pre-processing steps were applied to all selected samples in both databases. Fig. 1 presents selected samples of segmented ECG heartbeat following signal pre-processing and scaling for the four classes. As seen in Fig. 1, the T wave in the “MI” class is distinctively different from the “healthy” class. However, as the T wave in the “other” class also has similar morphology, a binary MI classifier might incorrectly classify it as MI, resulting in unnecessary emergency medical attention. The “noisy” class on the other hand has no

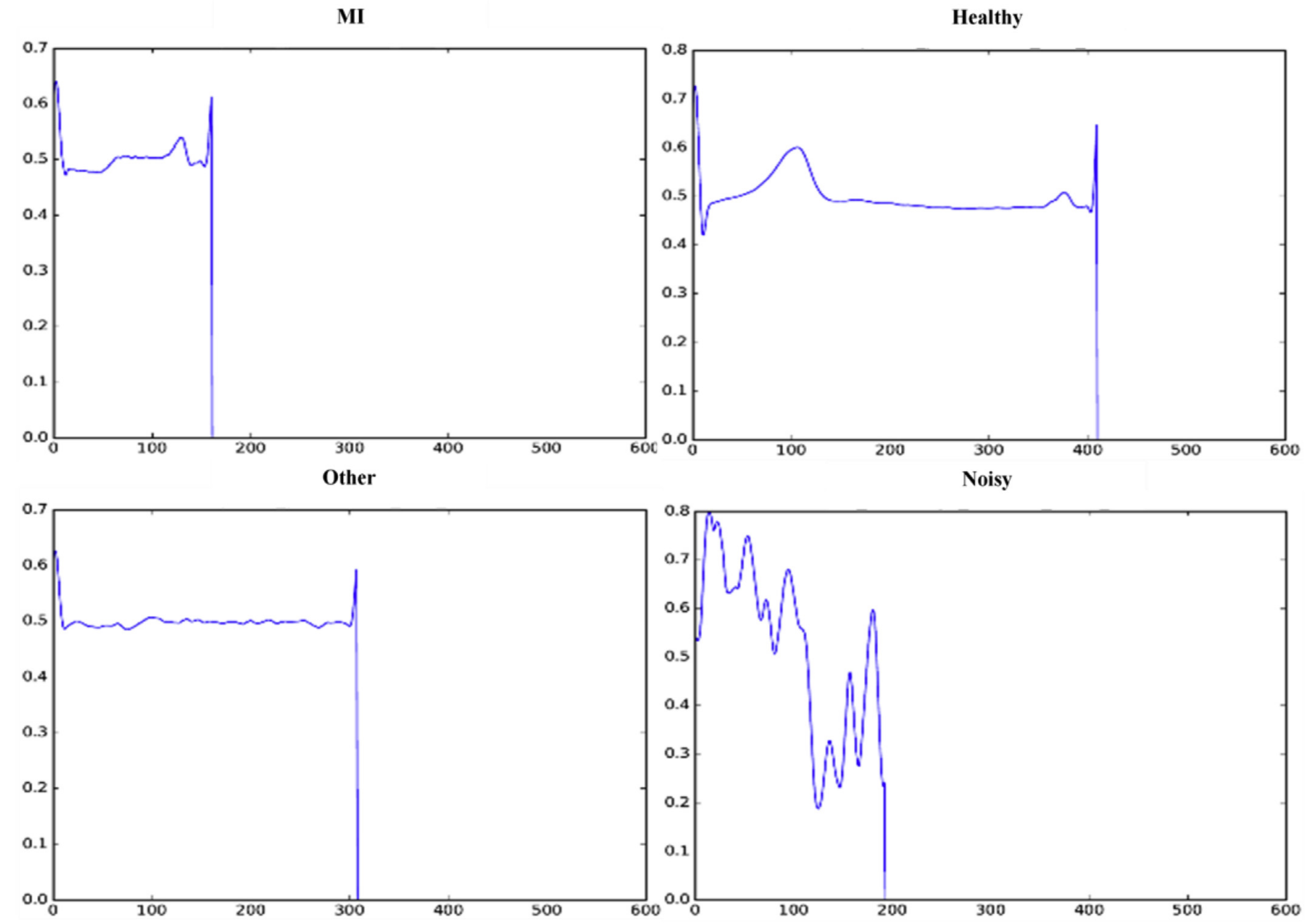

Fig. 1. ECG heartbeat after signal pre-processing, heartbeat segmentation, and scaling.

recognisable cardiac signal.

2.3. Classifier models

2.3.1. CNNs

The state-of-the-art CNN consists of a series of convolutional layers that apply convolutional operations to its inputs in order to detect different spatial patterns [41]: max-pooling layers to provide translational and orientational invariance [41]; dropout layers to provide regularisation and prevent over-fitting [42]; batch normalisation to prevent internal covariate shifting and allow for higher training rates [43]; and fully connected layers to perform the final classification [44]. Table 2 presents the architecture of the CNN classifier used in this study, which was constructed from four convolutional blocks and three fully connected layers. Table 1 presents the architecture of each convolutional block, containing two convolutional 1D layers, two batch normalisation layers, one max-pooling 1D layer and one dropout layer.

Table 1
Architecture of convolutional block.

Type	Filter size	Kernel/pool size
Convolutional 1D	32	3
Batch normalisation	–	–
Convolutional 1D	32	3
Batch normalisation	–	–
Max-pooling 1D	-	2
Dropout 50%	–	–

Table 2
Architecture of CNN classifier.

Layers	Type	Output shape
0	Inputs	512
1–6	Convolutional block	256 × 32
7–12	Convolutional block	128 × 32
13–18	Convolutional block	64 × 32
19–24	Convolutional block	32 × 32
25	Flattened	1024
26	Fully connected	32
27	Batch normalisation	32
28	Dropout 50%	32
29	Fully connected	32
30	Batch normalisation	32
31	Dropout 50%	32
32	Fully connected	16
33	Batch normalisation	16
34	Dropout 50%	16
35	Outputs	4

All convolutional layers contained 32 filters with a kernel size of 3, and a rectified linear activation function was applied [45]. The max-pooling layers all had a pool size of 2, and all fully connected layers also used the rectified linear activation function, while the output layer used the softmax activation function. The classifier was trained and evaluated with heartbeat ECG segments as input. As multiple heartbeats existed per ECG record, the final record classification was decided by majority voting, according to the classifications of its constituent heartbeat segments.

Table 3
Architecture of CNN-LSTM classifier.

Layers	Type	Output shape
0	Inputs	8 × 512
1–6	Time-distributed convolutional block	8 × 256 × 32
7–12	Time-distributed convolutional block	8 × 128 × 32
13–18	Time-distributed convolutional block	8 × 64 × 32
19–24	Time-distributed convolutional block	8 × 32 × 32
25	Time-distributed flattened	8 × 1024
26	Time-distributed fully connected	8 × 32
27	Time-distributed batch normalisation	8 × 32
28	Time-distributed dropout 50%	8 × 32
29	LSTM	32
30	Batch normalisation	32
31	Dropout 50%	32
32	Fully connected	16
33	Batch normalisation	16
34	Dropout 50%	16
35	Outputs	4

2.3.2. *Recurrent neural networks*

The recurrent neural network (RNN) is a class of the artificial neural network with recurrent connections. The activation of a recurrent node consists of feedback to itself from one time step to the next. In this study, we have utilised a family of RNN known as long short-term memory (LSTM). Each node in LSTM is a cell that comprises input, output and forget gates. These gates control the manner in which internal states are retained or discarded [46]. Owing to the LSTM ability to retain and process information over multiple time steps, it is an excellent candidate for processing heartbeat sequences, with each heartbeat acting as a separate time step for the LSTM layer.

We replaced layer 29 of the CNN classifier with a LSTM layer to create the CNN-LSTM classifier, the architecture of which is presented in Table 3. The LSTM layer had 32 nodes, the same number as the fully connected layer it replaced. All CNN classifier layers up to 29 were wrapped in a time-distributed function, aggregating the output vector for each heartbeat time step, prior to passing it to the LSTM layer.

Each record in the selected dataset had different recording durations and contained a varying number of heartbeats. However, to achieve fair comparisons, the number of heartbeats in each LSTM sequence had to be fixed. Moreover, owing to computational and memory constraints, we fixed the number of heartbeats in each LSTM sequence to 8. Therefore, each ECG record containing more than 8 heartbeats was divided into multiple 8-beat sequences. The final record classification was also decided by majority voting, based on the classifications of its constituent heartbeat sequences.

2.3.3. *Stacking decoding*

The CNN and CNN-LSTM classifier both contained an output layer of four units to represent the “MI”, “healthy”, “other” and “noisy” classes. This is the simplest means of performing multiclass classification with deep neural networks. Another technique that has exhibited improved performance is stacking decoding [47,48]. Firstly, binary classifiers are trained in a one-versus-one manner for each pair of available classes. Then, a multiclass meta-classifier is trained, with its inputs taken from the outputs of each binary classifier.

We trained an additional two classifiers with stacking decoding for both the CNN and CNN-LSTM. For the four-class classification problem, six binary classifiers were trained in a one-versus-one manner. Each of the two outputs of the six binary classifiers were used as inputs to the multiclass meta-classifier. A simple multi-layer perceptron (MLP) was used for the meta-classifier, as detailed in Table 4. All fully connected layers used the tanh activation function, while the output layer used the softmax activation function.

2.3.4. *Classifier with hand-crafted features*

To compare the performance of our deep learning classifiers, the traditional method of using an MLP classifier with hand-crafted features was also tested. Similar to one previous method [11], heart rate variability (HRV) features presented in Table 5 were used owing to their simplicity. As the duration of records in the selected dataset was short, and the desired application of portable ECG devices, it was not possible to compute valid frequency domain HRV features, and only time domain HRV features were included. The MLP classifier has the same architecture as presented in Table 4.

Table 4
Architecture of meta classifier.

Layers	Type	Output shape
0	Inputs	12
1	Fully connected	32
2	Batch normalisation	32
3	Dropout 20%	32
4	Fully connected	32
5	Batch normalisation	32
6	Dropout 20%	32
7	Outputs	4

Table 5
List of hand-crafted features.

Feature	Definition
MHR	Mean heart rate
MRRI	Mean beat-to-beat interval
NN50	Number of pairs of successive beat-to-beat intervals that differ by more than 50 ms
PNN50	Proportion of NN50 out of total number of heartbeats
RMSSD	Square root of mean of squares of successive differences between adjacent heartbeat intervals
SDNN	Standard deviation of heartbeat intervals
MAD	Mean absolute difference of ECG signal
AD_10	10^{th} percentile of absolute difference of ECG signal
AD_25	25^{th} percentile of absolute difference of ECG signal
AD_50	50^{th} percentile of absolute difference of ECG signal
AD_75	75^{th} percentile of absolute difference of ECG signal
AD_90	90^{th} percentile of absolute difference of ECG signal
AD_IR_75	Inter-range between 75th and 25th percentiles of absolute difference of ECG signal
AD_IR_90	Inter-range between 90^{th} and 10th percentiles of absolute difference of ECG signal

Furthermore, as HRV could not be effectively defined for noisy signals, the mean absolute difference of the ECG signal and the distribution of the absolute differences were used as additional features. Table 5 presents the features used.

Finally, for comparison purposes, a CNN-LSTM classifier with stacking decoding was trained with the addition of the hand-crafted features listed in Table 5. This classifier included the 12 inputs from the six binary classifiers, as well as a further 14 inputs from the hand-crafted features, resulting in 26 inputs in total.

2.4. *Study design*

2.4.1. *Cross-validation*

Ten-fold cross-validation was performed on the various classifiers trained in this study. For each fold, 10% of the samples were reserved for testing, and not used for training and validation. A separate testing set was used for each fold so that the entire dataset was employed for both training and testing. The remaining 90% of the samples were further divided into a 90% training set and 10% validation set. The training set was used for back-propagation and weight updating of the neural networks, while the validation set was used for evaluating early stopping and model saving.

2.4.2. *Over-sampling minority class*

The number of samples from each class was not balanced for the selected dataset, which would result in poor sensitivity of the minority class, as most optimisation algorithms would simply classify the majority class more frequently [49]. The dataset could be balanced by either under-sampling the majority class or over-sampling the minority class. Owing to the small sample size of the selected dataset, we opted for the over-sampling approach. Random samples from the minority class were added to the dataset, until all minority classes matched the sample size of the majority MI class with 368 samples. Over-sampling was only performed on the training set and validation set, and not the testing set.

2.4.3. Sample shuffling

Sample shuffling was performed on the training set for each training epoch, in order to ensure the optimisation was stochastic and prevent convergence to the local minimum [50]. Sample shuffling was performed once the training, validation and testing sets were divided, to ensure that heartbeat samples from the same record did not appear in both the training and testing. Without such a procedure, the classifier could optimise record-specific features with poor generalisation, such as the noise signature of the particular ECG device used for the record. This could result in artificially high performance of the testing set, but poor performance of a new dataset. One author demonstrated that the performance of a CNN classifier actually decreased after the noise was removed [10].

2.4.4. Hyperparameters

The RMS-propagation optimisation method was used in this study, owing to its excellent performance in training recurrent layers [51]. Each classifier was trained for at least 50 epochs, with a batch size of 32, and early stopping was performed when the validation loss was not improved over 10 epochs. Models were only saved and used for testing when the validation loss was minimised. All classifiers were trained on a workstation with a 4.0 GHz Core i7 6700K CPU, GTX 1080 GPU and 32 GB of memory. The TensorFlow numerical library was used in this study [52]. The duration of each epoch for the most complex classifier was 33 s.

2.5. Statistical analysis

In order to assess the relative performance of each classifier, the sensitivity, specificity, positive predictive value (PPV) and F1 score for the MI class were computed. Given the confusion matrix in Table 6, the evaluation metrics are defined in equations (1)–(4). Friedman test was performed as it is the most suitable method for statistical significance testing of multiple machine learning algorithms [53]. The distribution of each performance metric from the 10-fold cross-validation for the 6 classifiers were used to evaluate the corresponding chi-square statistic and p-value. The distributions of performance metric are considered different between classifiers with statistical significance if the p-value is less than 0.05.

$$Sensitivity = \frac{Mm}{\sum M} \tag{1}$$

Table 6
Confusion matrix of classifier.

		Predicted classification				
		MI	Healthy	Other	Noisy	Total
Reference classification	MI	Mm	Mh	Mo	Mn	ΣM
	Healthy	Hm	Hh	Ho	Hn	ΣH
	Other	Om	Oh	Oo	On	ΣO
	Noisy	Nm	Nh	No	Nn	ΣN
	Total	Σm	Σh	Σo	Σn	

Table 7
Confusion matrix of hand-crafted features MLP classifier.

		Predicted classification				
		MI	Healthy	Other	Noisy	Total
Reference classification	MI	200	46	110	12	368
	Healthy	12	49	16	3	80
	Other	15	9	48	2	74
	Noisy	2	1	5	270	278
	Total	229	105	179	287	

$$Specificity = \frac{Hh + Ho + Hn + Oh + Oo + On + Nh + No + Nn}{\sum \mathrm{H} + \sum \mathrm{O} + \sum \mathrm{N}} \tag{2}$$

$$PPV = \frac{Mm}{\sum m} \tag{3}$$

$$F1 = \frac{2 \times Mm}{\sum M + \sum m} \tag{4}$$

3. Results

Tables 7–12 display the confusion matrices of the aggregated total of the testing set from the 10-fold cross-validation of the various classifiers. Table 13 presents the performance metrics of the classifiers and the corresponding chi-square statistics and p-values. All performance metrics were evaluated based on taking the corresponding metric average from each fold. Table 14 displays the performance comparison of the best classifier from the proposed method with existing methods by other researchers using the PTB database.

4. Discussion

Substantial and statistically significant performance variations were observed among the classifiers, as indicated in Table 13. A general trend was that the addition of the LSTM layer and stacking decoding improved the classification performance. Of the six tested classifiers, only the CNN-LSTM stacking decoding classifier could achieve over 90% in all performance metrics. Such a large performance variation among each classifier demonstrates the difficulty and validity of a multiclass approach to MI detection.

4.1. LSTM layer evaluation

As observed in Table 13, the inclusion of the LSTM layer significantly improved the classification performance. Comparing the sensitivity between the CNN and CNN-LSTM classifiers, there was an 18.3% improvement for the pair without stacking decoding, and a 28.0% improvement for the pair with stacking decoding. There was a slight decrease in specificity of 5.2% and PPV of 5.3% for the CNN-LSTM classifier without stacking decoding. However, the loss in specificity and PPV was offset by the gain in sensitivity; hence, an overall improvement of 13.3% in the F1 score was achieved.

The results support the hypothesis that the LSTM layer can

Table 8
Confusion matrix of CNN classifier.

		Predicted classification				
		MI	Healthy	Other	Noisy	Total
Reference classification	MI	184	73	68	43	368
	Healthy	9	58	6	7	80
	Other	22	12	35	5	74
	Noisy	4	2	2	270	278
	Total	219	145	111	325	

Table 9
Confusion matrix of CNN-LSTM classifier.

		Predicted classification				
		MI	Healthy	Other	Noisy	Total
Reference classification	MI	251	43	56	18	368
	Healthy	17	54	4	5	80
	Other	35	5	30	4	74
	Noisy	5	1	6	266	278
	Total	308	103	96	293	

Table 10
Confusion matrix of CNN stacking decoding classifier.

		Predicted classification				
		MI	Healthy	Other	Noisy	Total
Reference classification	MI	237	54	66	11	368
	Healthy	7	66	5	2	80
	Other	7	6	60	1	74
	Noisy	2	0	2	274	278
	Total	253	126	133	288	

Table 11
Confusion matrix of CNN-LSTM stacking decoding classifier.

		Predicted classification				
		MI	Healthy	Other	Noisy	Total
Reference classification	MI	340	16	9	3	368
	Healthy	3	76	1	0	80
	Other	5	2	67	0	74
	Noisy	2	1	1	274	278
	Total	350	95	78	277	

Table 12
Confusion matrix of CNN-LSTM stacking decoding classifier with hand-crafted features.

		Predicted classification				
		MI	Healthy	Other	Noisy	Total
Reference classification	MI	294	22	49	3	368
	Healthy	2	73	4	1	80
	Other	3	2	67	2	74
	Noisy	0	0	3	275	278
	Total	299	97	123	281	

Table 13
Performance metrics of different classifiers.

Classifier	Sensitivity	Specificity	PPV	F1
Hand-crafted features MLP	54.4%	93.3%	87.2%	66.3%
CNN	49.8%	92.0%	86.5%	59.7%
CNN-LSTM	68.1%	86.8%	81.2%	73.0%
CNN stacking decoding	64.4%	96.3%	93.9%	75.9%
CNN-LSTM stacking decoding	**92.4%**	97.7%	97.2%	**94.6%**
CNN-LSTM stacking decoding with hand-crafted features	79.9%	**98.8%**	**98.3%**	87.2%
Chi-square statistic	32.7	29.4	35.9	34.6
P-value	4.0e-6	1.9e-5	1.0e-6	2.0e-6

effectively extract information on the beat-to-beat variations of the ECG morphology and improve classification performance.

4.2. Stacking decoding evaluation

Comparing the CNN classifiers with and without stacking decoding, we observed a performance gain of 14.6% in sensitivity, a more modest gain of 4.3% in specificity, and a 7.4% gain in PPV. The overall F1 score thus improved by 16.2%. The improvements in the CNN-LSTM classifiers with and without stacking decoding were even higher, with 24.3%, 10.9%, 16.0% and 21.6% improvements in the sensitivity, specificity, PPV and F1 score, respectively.

The results are consistent with previous findings that stacking decoding achieves superior performance to a single multiclass classifier [48]. Stacking decoding is similar to an ensemble learner, in which multiple weak learners are trained with a specialisation of a different subset of samples, thereby forming a strong learner [54]. In this study, each weak learner was specialised for pairwise binary classification, and together they formed a strong learner for the multiclass classification. Despite the improved performance, the disadvantage of stacking decoding is that the computational cost is scaled quadratically with the number of classes. From our experiment, the training and testing time for the entire 10-fold cross-validation was 44353 s for the most complex CNN-LSTM with stacking decoding, compared to 8822 s without.

4.3. Hand-crafted features evaluation

It was observed that the CNN classifier, as a deep learning method, did not perform better than the more traditional method of using hand-crafted features. The sensitivity, specificity, PPV and F1 score were in fact decreased by 4.6%, 1.3%, 0.7% and 6.6% respectively, for the CNN classifier. This is consistent with our literature review in which certain traditional classifiers with a well-selected set of features could perform better than a deep learning method [20,26]. However, with the addition of the LSTM layer and stacking decoding, all of our other deep learning classifiers performed better than the hand-crafted features MLP classifier.

A surprising result was that of the CNN-LSTM stacking decoding classifier with hand-crafted features, for which the sensitivity decreased by 12.5%. While the specificity and PPV both improved by 1.1%, the F1 score was decreased by 7.4% overall. It appears that the HRV and signal absolute difference features provided conflicting information against the CNN and LSTM layers, resulting in a performance decrease.

4.4. Sensitivity of individual classes

When evaluating the confusion matrices of all classifiers from Tables 7–12, a common trend emerged whereby most of the MI misclassifications were assigned to the “other” class. For the majority of classifiers, the number of misclassifications from “MI” to “other” was higher than that from “MI” to “healthy”. The reverse was also true, with most of the misclassifications from the “other” class being assigned to the “MI” instead of the “healthy” class for all classifiers.

These results demonstrate the importance of MI classification being a multiclass problem. Most previous works on MI detection only considered binary classification of MI and healthy ECGs, and the sensitivity could reach 99.6% [20]. One author used three very simple HRV features, namely MHR, SDNN and RMSSD, to perform binary classification, and achieved 100% sensitivity [11]. It is likely that such a classifier would incorrectly classify patients with an existing chronic heart condition such as MI, causing the patient to receive unnecessary emergency medical attention.

Surprisingly, the noisy data classification performance was generally high across all classifiers, with very few noisy data misclassified. However, the addition of noisy data caused several MI misclassifications, as indicated in Table 8. This further supports the case that noisy

Table 14
Performance comparisons with existing methods.

Author	Method	No. of leads	Classes	Sensitivity	Specificity
Banerjee et al. [22]	Predefined threshold of morphological features extracted by DWT	V1-V4	MI, healthy	92.0%	100%
Chang et al. [12]	SVM on polynomial approximation coefficients of ST segment following PCA	12 leads	MI, non-MI	98.7%	96.6%
Arif et al. [20]	KNN on morphological features	12 leads	MI, healthy	99.6%	99.1%
Sun et al. [21]	KNN on morphological features with multiple instance learning	12 leads	MI, healthy	92.3%	88.1%
Sharma et al. [24]	SVM on principal component multivariate multi-scale sample entropy features	12 leads	MI, healthy, cardiomyopathy, hypertrophy, dysrhythmia	94.0%	89.5%
Huang et al. [26]	SDA on morphological features following PCA	II	healthy, non-healthy	89.7%	84.6%
Kora et al. [23]	Levenberg–Marquardt neural network on morphological features selected by improved bat algorithm	II, III, aVF	MI, healthy	93.3%	92.2%
Zewdie et al. [14]	SVM on coefficients of second-order ordinary differential equation fitting	I	MI, healthy, hypertrophy, valvular heart disease, myocarditis and miscellaneous	99.8%	72.7%
Kumar et al. [17]	LS-SVM on features extracted by sample entropy in flexible analytic wavelet transform	II	MI, healthy	98.2%	99.2%
Uddin et al. [11]	LDA on HRV features	I	MI, healthy	100%	75.0%
Acharya et al. [16]	11-layer CNN on single beat waveform	II	MI, healthy	95.5%	94.2%
Proposed method	**CNN-LSTM stacking decoding**	**I**	**MI, healthy, other CVD, noisy**	**92.4%**	**97.7%**

data should be incorporated as an additional data class in order to create a robust MI detection system.

Finally, our method is also an excellent candidate for detecting CVDs in general. As demonstrated in Table 11, for the most accurate CNN-LSTM stacking decoding classifier, the sensitivity of the “healthy” and “other” classes was also high, at 95.0% and 90.5%, respectively. Further investigations could be carried out on a generic classifier for healthy and CVDs with a larger dataset.

4.5. Performance comparison with existing methods

As indicated in Table 14, the majority of existing methods did not perform multiclass MI classification against other forms of CVDs, with the exception of [24] and [14]. In general, methods that only perform binary classification of MI and healthy subjects achieved higher sensitivity and specificity, with [20] [12], and [17] demonstrating nearly perfect results. This is expected, as with binary classification only, there is less opportunity for misclassification; hence, higher classification performance is achieved. However, as these methods did not include ECG records of other CVD types, they might not be able to accurately classify MI and other CVD types. An MI detection system that frequently misclassify other CVD types as MI is less useful for determining whether emergency treatment is necessary. Indeed, misclassification of other CVD types as MI was found to be common in our less accurate classifiers, as discussed in section 4.4.

While [24] and [14] both achieved higher sensitivity than our proposed method, at 94.0% and 99.8%, these approaches demonstrated considerably lower specificity, at 89.5% and 72.7%, respectively. As the ECG records of other CVD forms exhibit numerous similarities to those of MI, it is more difficult to achieve high specificity. It should be noted that, while both methods performed multiclass classification, they did not include noisy ECG records, which is an important criterion for portable ECG devices. This is especially problematic in Ref. [24], which relied on entropy features for performing classification. ECG records with MI exhibit higher entropy than those of healthy and other CVDs, as the heart is in a more chaotic state during an MI episode. This method may be less capable of differentiating between MI and noisy ECG records, as the entropy of noisy ECG records is also high.

4.6. Computational time

The computational time of interference for a single ECG record was approximately 350 ms for the most accurate CNN + LSTM stacking decoding classifier. While this is a reasonable timeframe for an MI detection system, a portable ECG device is unlikely to possess the same level of computational resources as our system, owing to power consumption constraints. However, our classifier is still suitable for cloud-based implementation, in which ECG records are sent to a cloud server with more computational resources through a mobile network.

4.7. Limitations and future work

It is noted that the localisation of the infarction should determine which ECG lead is capable of revealing ST elevation and reciprocal ST depression. In standard ECG diagnostic guidelines, lead 1 ECG should only be capable of revealing ST changes of lateral, inferior, and arterial MI [55]. In the present and some prior works [11,14], the sensitivity with only lead 1 is higher than the 50–60% expected. This could be attributed to the fact that CNN and RNN are deep learning techniques which search for all possible regularities indicative of MI, any subtle changes of any waveform morphology as well as beat-to-beat variations could be used as features by the deep neural networks. In addition, hand-crafted features MLP and [11] can achieve reasonably good performance and high specificity despite only using HRV features, which should only have prognosis values for MI. While a 10-fold cross-validation was performed in the present work, owing to the small sample size of the selected dataset there is still a possibility of overfitting. Future work can explore testing the presented techniques on a larger dataset.

5. Conclusion

In this study, we have developed an MI classifier that combines both the CNN and RNN to achieve higher performance than the CNN alone, with 92.4% sensitivity, 97.7% specificity, 97.2 PPV and 94.6% F1 score. The classifier uses the stacking decoding technique to achieve higher performance for the multiclass classification of “MI”, “healthy”, “other” and “noisy” ECG recordings. It was compared with and demonstrated to achieve higher performance than machine learning method with hand-crafted features in our study. Moreover, we compared our proposed method with other existing methods and it was found to achieved higher performance than existing multiclass classifiers. While some existing binary classifiers between MI and healthy ECGs achieved higher performance, they are less useful for an MI detection system that requires accurate classification between MI and other CVD types to determine the need of emergency treatment. Further work could be done for these methods to perform similar multiclass classification as this paper to compare performance. While the developed MI classifier was trained on a dataset of MI ECG records, the same method can be applied to detecting other CVD types and heart conditions, by training the classifiers on the relevant ECG datasets. Further work could explore the training of different ECG classifiers to detect various heart conditions.

Acknowledgement

This research was supported by University Grants Committee of HKSAR government for the funding of research studentship.

Appendix A. Supplementary data

Supplementary data related to this article can be found at https://doi.org/10.1016/j.imu.2018.08.002.

References

[1] Maxwell S. Emergency management of acute myocardial infarction. Br J Clin Pharmacol 1999;48:284–98. https://doi.org/10.1046/j.1365-2125.1999.00998.x.
[2] Reed GW, Rossi JE, Cannon CP. Acute myocardial infarction. Lancet 2017;389:197–210. https://doi.org/10.1016/S0140-6736(16)30677-8.
[3] Kutty RS, Jones N, Moorjani N. Mechanical complications of acute myocardial infarction. Cardiol Clin 2013;31:519–31. https://doi.org/10.1016/j.ccl.2013.07.004.
[4] Kasper D, Fauci A, Hauser S, Longo D, Jameson J, Loscalzo J. Harrison's principles of internal medicine. 2015. https://doi.org/10.1036/007149619X.
[5] Valensi P, Lorgis L, Cottin Y. Prevalence, incidence, predictive factors and prognosis of silent myocardial infarction: a review of the literature. Arch Cardiovasc Dis 2011;104:178–88. https://doi.org/10.1016/j.acvd.2010.11.013.
[6] Thygesen K, Alpert. Third universal definition of myocardial infarction. Circulation 2012;126:2020–35. https://doi.org/10.1161/CIR.0b013e31826e1058.
[7] Mowad EM. Electrocardiogram interpretation. Compr. Pediatr. Hosp. Med. 2007:522–35. https://doi.org/10.1016/B978-032303004-5.50086-7.
[8] Tung R, Zimetbaum P. Use of the electrocardiogram in acute myocardial infarction. Card. Intensive Care; 2010. p. 106. https://doi.org/10.1016/B978-1-4160-3773-6.10011-4.
[9] Baig MM, Gholamhosseini H, Connolly MJ. A comprehensive survey of wearable and wireless ECG monitoring systems for older adults. Med Biol Eng Comput 2013;51:485–95. https://doi.org/10.1007/s11517-012-1021-6.
[10] Wu JF, Bao YL, Chan SC, Wu HC, Zhang L, Wei XG. Myocardial infarction detection and classification-A new multi-scale deep feature learning approach. Int. Conf. Digit. Signal Process. DSP 2017:309–13. https://doi.org/10.1109/ICDSP.2016.7868568.
[11] Uddin SA, Rahman A. Myocardial infraction classification by HRV analysis using single lead ECG. AIUB J Sci Eng 2017;41–6.
[12] Chang P-C, Lin Y-CW J-J. Myocardial infarction classification using polynomial aproximation and principle component analysis. Natl Digit Libr Thesis Diss Taiwan; 2011.
[13] Sopic D, Aminifar A, Aminifar A, Atienza Alonso D. Real-time classification technique for early detection and prevention of myocardial infarction on wearable devices. 13th IEEE biomed. Circuits syst. Conf. 2017. p. 2–5.
[14] Zewdie G, Xiong M. Wearable computing for fully automated myocardial infarction classification. Proc. 8th Int. Conf. Bioinforma. Comput. Biol. 2016:17–22.
[15] Banerjee S, Mitra M. A classification approach for myocardial infarction using voltage features extracted from four standard ECG leads. 2011 int conf recent trends inf syst ReTIS 2011 - proc 2011. p. 325–30. https://doi.org/10.1109/ReTIS.2011.6146890.
[16] Acharya UR, Fujita H, Oh SL, Hagiwara Y, Tan JH, Adam M. Application of deep convolutional neural network for automated detection of myocardial infarction using ECG signals. Inf Sci 2017;415–416:190–8. https://doi.org/10.1016/j.ins.2017.06.027.
[17] Kumar M, Pachori R, Acharya U. Automated diagnosis of myocardial infarction ECG signals using sample entropy in flexible analytic wavelet transform framework. Entropy 2017;19:488. https://doi.org/10.3390/e19090488.
[18] Thatipelli T, Kora P. Classification of myocardial infarction using discrete wavelet transform and support vector machine. Int Res J Eng Technol 2017;4:429–32.
[19] Remya RS, Indiradevi KP, Babu KKA. Classification of myocardial infarction using multi resolution wavelet analysis of ECG. Procedia Technol 2016;24:949–56. https://doi.org/10.1016/j.protcy.2016.05.195.
[20] Arif M, Malagore IA, Afsar FA. Detection and localization of myocardial infarction using K-nearest neighbor classifier. J Med Syst 2012;36:279–89. https://doi.org/10.1007/s10916-010-9474-3.
[21] Sun L, Lu Y, Yang K, Li S. ECG analysis using multiple instance learning for myocardial infarction detection. IEEE Trans Biomed Eng 2012;59:3348–56. https://doi.org/10.1109/TBME.2012.2213597.
[22] Banerjee S. ECG feature extraction and classification of anteroseptal myocardial infarction and normal subjects using discrete wavelet transform. Syst Med Biol 2010:55–60.
[23] Kora P, Kalva SR. Improved Bat algorithm for the detection of myocardial infarction. SpringerPlus 2015;4:1–18. https://doi.org/10.1186/s40064-015-1379-7.
[24] Sharma LN, Dandapat S, Tripathy RK. A new way of quantifying diagnostic information from multilead electrocardiogram for cardiac disease classification. Healthc Technol Lett 2014;1:98–103. https://doi.org/10.1049/htl.2014.0080.
[25] Menown IBA, MacKenzie G, Adgey AAJ. Optimizing the initial 12-lead electrocardiographic diagnosis of acute myocardial infarction. Eur Heart J 2000. https://doi.org/10.1053/euhj.1999.1748.
[26] Huang R, Zhou Y. Disease classification and biomarker discovery using ECG data. BioMed Res Int 2015;2015. https://doi.org/10.1155/2015/680381.
[27] Zewdie G, Xiong M. Fully automated myocardial infarction using ordinary differential equations. Proc. 8th int. Conf. Bioinforma. Comput. Biol. BICOB. 2016. p. 17–22.
[28] Bousseljot R, Kreiseler D, Schnabel A. Nutzung der EKG-Signaldatenbank CARDIODAT der PTB über das Internet. Biomed Tech 1995;40:317–8. https://doi.org/10.1515/bmte.1995.40.s1.317.
[29] Goldberger AL, Amaral LAN, Glass L, Hausdorff JM, Ivanov PC, Mark RG, et al. PhysioBank, PhysioToolkit, and PhysioNet : components of a new research resource for complex physiologic signals. Circulation 2000;101:e215–20. https://doi.org/10.1161/01.CIR.101.23.e215.
[30] Colledge NR, Walker BR, Ralston S, Britton R, Davidson SS. Davidson's principles and practice of medicine. Princ. Pract. Med. 2010:588–99.
[31] Guo Y, Liu Y, Oerlemans A, Lao S, Wu S, Lew MS. Deep learning for visual understanding: a review. Neurocomputing 2016;187:27–48. https://doi.org/10.1016/j.neucom.2015.09.116.
[32] Wiatowski T, Bölcskei H. A mathematical theory of deep convolutional neural networks for feature extraction. ArXiv 2015:1–48. doi:10.1109/TIT.2017.2776228.
[33] Acharya UR, Oh SL, Hagiwara Y, Tan JH, Adam M, Gertych A, et al. A deep convolutional neural network model to classify heartbeats. Comput Biol Med 2017;89:389–96. https://doi.org/10.1016/j.compbiomed.2017.08.022.
[34] Acharya UR, Fujita H, Lih OS, Hagiwara Y, Tan JH, Adam M. Automated detection of arrhythmias using different intervals of tachycardia ECG segments with convolutional neural network. Inf Sci 2017;405:81–90. https://doi.org/10.1016/j.ins.2017.04.012.
[35] Acharya UR, Fujita H, Lih OS, Adam M, Tan JH, Chua CK. Automated detection of coronary artery disease using different durations of ECG segments with convolutional neural network. Knowl Base Syst 2017;132:62–71. https://doi.org/10.1016/j.knosys.2017.06.003.
[36] Clifford G, Liu C, Moody B, Silva I, Li Q, Johnson A, et al. AF classification from a short single lead ECG recording: the PhysioNet computing in cardiology challenge. Comput Cardiol 2017;44:1–4. (2010).
[37] Hargittai S. Savitzky-Golay least-squares polynomial filters in ECG signal processing. Comput Cardiol 2005;32:763–6. https://doi.org/10.1109/CIC.2005.1588216.
[38] Schafer RW. Savitzky-Golay filters. 2011 Digit. Signal process. Signal process. Educ. Meet. DSP/SPE 2011 - proc. 2011. https://doi.org/10.1109/DSP-SPE.2011.5739186.
[39] Kathirvel P, Manikandan MS, Prasanna SRM, Soman KP. An efficient R-peak detection based on new nonlinear transformation and first-order Gaussian differentiator. Cardiovasc Eng Technol 2011;2:408–25. https://doi.org/10.1007/s13239-011-0065-3.
[40] Jayalakshmi T, Santhakumaran A. Statistical normalization and backpropagation for classification. Int J Comput Theory Eng 2011;3:89–93.
[41] Gu J, Wang Z, Kuen J, Ma L, Shahroudy A, Shuai B, et al. Recent advances in convolutional neural networks. ArXiv 2015:1–14. doi:10.1007/s12274-015-0938-0.
[42] Srivastava N, Hinton G, Krizhevsky A, Sutskever I, Salakhutdinov R. Dropout: a simple way to prevent neural networks from overfitting. J Mach Learn Res 2014;15:1929–58. https://doi.org/10.1214/12-AOS1000.
[43] Windows M, Os. Batch normalization: accelerating deep network training by reducing internal covariate shift. Uma Ética Para Quantos? 2014;XXXIII:81–7. https://doi.org/10.1007/s13398-014-0173-7.2.
[44] Lecun Y, Bengio Y, Hinton G. Deep learning. Nature 2015;521:436–44. https://doi.org/10.1038/nature14539.
[45] Nair V, Hinton GE. Rectified linear units improve restricted Boltzmann machines. Proc 27th Int Conf Mach Learn 2010:807–14. doi:10.1.1.165.6419.
[46] Hochreiter S, Urgen Schmidhuber J. Long short-term memory. Neural Comput 1997;9:1735–80. https://doi.org/10.1162/neco.1997.9.8.1735.
[47] Wolpert DH. Stacked generalization. Neural Network 1992;5:241–59. https://doi.org/10.1016/S0893-6080(05)80023-1.
[48] Lezoray O, Cardot H. Combining multiple pairwise neural networks classifiers: a comparative study. Proc. 1st int. Work. Artif. Neural networks intell. Inf. Process. ANNIIP 2005 - conjunction with ICINCO 2005. 2005. p. 52–61.
[49] Poolsawad N, Kambhampati C, Cleland JGF. Balancing class for performance of classification with a clinical dataset. World Congr Eng WCE 2014 2014;1:237–42.
[50] Lai TL. Stochastic approximation. Ann Stat 2003;31:391–406. https://doi.org/10.1214/aos/1051027873.
[51] Tieleman T, Hinton G. Lecture 6.5-rmsprop: divide the gradient by a running average of its recent magnitude. COURSERA neural networks mach learn. 2012.
[52] Abadi M, Barham P, Brain G. TensorFlow: a system for large-scale machine learning TensorFlow: a system for large-scale machine learning. 12th USENIX symp. Oper. Syst. Des. Implement. (OSDI '16) 2016. p. 265–84. https://doi.org/10.1038/nn.3331.
[53] Demšar J. Statistical comparisons of classifiers over multiple data sets. J Mach Learn Res 2006;7:1–30. https://doi.org/10.1016/j.jecp.2010.03.005.
[54] Opitz D, Maclin R. Popular ensemble methods: an empirical study. J Artif Intell Res 1999;11:169–98. https://doi.org/10.1613/jair.614.
[55] Wung SF, Kahn DY. A quantitative evaluation of ST-segment changes on the 18-lead electrocardiogram during acute coronary occlusions. J Electrocardiol 2006. https://doi.org/10.1016/j.jelectrocard.2005.10.007.