Mehrang, S., Jafari Tadi, M., Knuutila, T., Jaakkola, J., Jaakkola, S., Kiviniemi, T., Vasankari, T., Airaksinen, J., Koivisto, T., Pänkäälä, M. 2022. End-to-end sensor fusion and classification of atrial fibrillation using deep neural networks and smartphone mechanocardiography. Physiological Measurement 5/2022.

IOP Publishing Physiol. Meas. 43 (2022) 055004 https://doi.org/10.1088/1361-6579/ac66ba

Physiological Measurement

IPEM Institute of Physics and Engineering in Medicine

RECEIVED
15 November 2021

REVISED
9 April 2022

ACCEPTED FOR PUBLICATION
12 April 2022

PUBLISHED
25 May 2022

PAPER

End-to-end sensor fusion and classification of atrial fibrillation using deep neural networks and smartphone mechanocardiography

Saeed Mehrang[1,4], **Mojtaba Jafari Tadi**[1,2,4], **Timo Knuutila**[1], **Jussi Jaakkola**[3], **Samuli Jaakkola**[3], **Tuomas Kiviniemi**[3], **Tuija Vasankari**[3], **Juhani Airaksinen**[3], **Tero Koivisto**[1] **and Mikko Pänkäälä**[1]

[1] Department of Computing, University of Turku, Turku, FI, Finland
[2] School of ICT, Faculty of Engineering, Turku University of Applied Sciences, Turku, FI, Finland
[3] Heart Center, Turku University Central Hospital, Turku, FI, Finland
[4] These authors equally contributed to this work.

E-mail: saeed.mehrang@utu.fi

Keywords: seimocardiography, gyrocardiography, accelerometer, gyroscope, atrial fibrillation, deep learning, sensor fusion

Abstract

Objective. The purpose of this research is to develop a new deep learning framework for detecting atrial fibrillation (AFib), one of the most common heart arrhythmias, by analyzing the heart's mechanical functioning as reflected in seismocardiography (SCG) and gyrocardiography (GCG) signals. Jointly, SCG and GCG constitute the concept of mechanocardiography (MCG), a method used to measure precordial vibrations with the built-in inertial sensors of smartphones. *Approach.* We present a modified deep residual neural network model for the classification of sinus rhythm, AFib, and Noise categories from tri-axial SCG and GCG data derived from smartphones. In the model presented, pre-processing including automated early sensor fusion and spatial feature extraction are carried out using attention-based convolutional and residual blocks. Additionally, we use bidirectional long short-term memory layers on top of fully-connected layers to extract both spatial and spatiotemporal features of the multidimensional SCG and GCG signals. The dataset consisted of 728 short measurements recorded from 300 patients. Further, the measurements were divided into disjoint training, validation, and test sets, respectively, of 481 measurements, 140 measurements, and 107 measurements. Prior to ingestion by the model, measurements were split into 10 s segments with 75 percent overlap, pre-processed, and augmented. *Main results.* On the unseen test set, the model delivered average micro- and macro-F1-score of 0.88 (0.87–0.89; 95% CI) and 0.83 (0.83–0.84; 95% CI) for the segment-wise classification as well as 0.95 (0.94–0.96; 95% CI) and 0.95 (0.94–0.96; 95% CI) for the measurement-wise classification, respectively. *Significance.* Our method not only can effectively fuse SCG and GCG signals but also can identify heart rhythms and abnormalities in the MCG signals with remarkable accuracy.

1. Introduction

Atrial fibrillation (AFib) is a widespread chronic and relapsing heart arrhythmia, present in approximately 2% of individuals, accounting for 20%–45% of all ischemic strokes worldwide (Kirchhof *et al* 2016). AFib increases the risk of heart failure, which lowers the quality of life, especially in symptomatic patients (Gregory and Antonio 2006, Elisa *et al* 2010), and heightens morbidity and mortality rates (Valentin *et al* 2001, Camm *et al* 2010).

Today, various measurement techniques are available to detect heart arrhythmia, of which electrocardiography (ECG) is the most widely validated and guideline-recommended gold standard. In addition, a variety of clinically validated wearable devices offer ECG-based monitoring, including smartphones and smartwatches (Lau *et al* 2013, Tieleman *et al* 2014, Hendrikx *et al* 2014, Barrett *et al* 2014, Perez *et al* 2019). The other measurement technique which has been effective in detecting AFib is mechanocardiography (MCG) (Jaakkola *et al* 2018) which refers to the joint measurement of tri-axial seismocardiogram (SCG) (Zanetti and Tavakolian 2013) and tri-axial gyrocardiogram

(GCG) (Tadi *et al* 2017) signals. Tri-axial SCG refers to the chest movement acceleration resulting from heart functions recorded by a 3-dimensional accelerometer. Similarly, tri-axial GCG refers to the chest movement angular velocity resulted from heart functions recorded by a 3-dimensional gyroscope. Nowadays, almost all of the smartphones and wearable sensors are equipped with 3-dimensional accelerometer and/or gyroscope sensors which can be used for ambulatory MCG recording.

In ECG signals, AFib is characterized by two attributes: (i) absence of regular sinus node originated P-waves and (ii) presence of irregularly irregular inter-beat timing and amplitude variations (Hindricks *et al* 2020). In MCG signals, since the signals are originated from mechanical movements rather than the electrophysiological activity of the heart, we may not necessarily observe the same AFib characteristics as in ECG signals. On the other hand, unlike ECG signals, finding reliable and robust AFib characteristics in MCG signals is challenging and requires extensive domain knowledge and substantial exploratory analysis. In this case, data-driven learning approaches such as deep learning can be useful as they provide us with fully automated feature extraction and classification (Zhang *et al* 2020).

Deep learning has been widely applied to biomedical data (Baldi 2018, Park *et al* 2018) including ECG signals (Pyakillya *et al* 2017). The introduction of deep learning to ECG analysis has opened new avenues for improving the detection and classification of pathological heart conditions (Somani *et al* 2021). The key to the success of deep neural networks (DNN) is learning representative features through iterative optimization of model weights according to the model output compared to the ground truth or expected output values. The deep automated feature learning applicability becomes even more pronounced when we deal with multidimensional heterogeneous time series data, especially if the data are difficult to interpret by visual inspection or conventional signal processing-based feature extraction (Miotto *et al* 2017). Consequently, automated feature learning becomes relevant when we deal with multidimensional MCG data (Suresh *et al* 2020).

In our previous contributions (Tadi *et al* 2018, Mehrang *et al* 2019, 2020), we have addressed the classification of AFib and SR classes utilizing MCG signals via feature engineering and injecting domain knowledge into the solution. In the absence of sufficient domain knowledge and/or the presence of a huge target population, DNNs are legitimate alternatives that are highly scalable in terms of generalization and predictive power. With this motivation, in this paper, we present a deep convolutional-recurrent neural network (CRNN) architecture that consists of attention-based convolutional and residual blocks (He *et al* 2016) as well as stage-level dense connections (Huang *et al* 2017) to perform automated early sensor fusion (Münzner *et al* 2017) and spatial feature extraction. In particular, Squeeze-and-Excitation (SE) blocks (Hu *et al* 2018) are used for implementing attention mechanism, stage-level shortcut connections for alleviating the vanishing-gradient problem and facilitating feature reuse (Huang *et al* 2017), bidirectional Long Short-term Memory (LSTM) layers (Hochreiter and Schmidhuber 1997) for temporal feature extraction, and fully-connected layers on top for the classification.

In this paper, in addition to the AFib and sinus rhythm (SR) classification, we aim to detect noisy measurements as well. One of the major issues in the analysis of cardiac signals, be it ECG or MCG, is the proper detection of the noise level in the collected signals (Kumar and Sharma 2020). An excess amount of noise can mislead the algorithms and consequently lead to the wrong disease classification (Kumar and Sharma 2020). Detecting the noisy episodes of a measurement is, therefore, a crucial step toward improving the reliability of the MCG signal analysis. In the case of MCG signals, noise class is defined as a condition where the underlying physiological properties cannot be seen or extracted from the signals. This condition can be caused by sensor placement failure or the presence of unwanted/undesired data generation sources. Hereafter, we denote the noise class with *Noise* throughout the rest of this document.

The main contribution of this study is the adoption of attention-based residual CRNN model architecture for performing an automated end-to-end sensor fusion, spatiotemporal feature extraction, and classification on the multidimensional MCG signals that are collected solely by smartphones. The adopted model performs a three-class classification to discriminate AFib, SR, and Noise classes.

2. Methods

2.1. Data acquisition and measurement protocol

Our dataset included retrospective (de-identified) data from 300 age and gender matched elderly patients, including 150 patients with AFib as the prevalent heart rhythm during the recording (Jaakkola *et al* 2018). The demographics of the patients can be found in Tadi *et al* (2018). An android smartphone with a running custom-designed application for research was placed on the subject's chest longitudinally while the screen was facing upwards and the bottom edge of the phone at the level of the lower edge of the body of the sternum (Tadi *et al* 2018). We gathered two sets of measurement scenarios, including physician-applied and patient-applied. In a patient-applied measurement, the subject was instructed to place the sensor on the chest and initiate the recording (Tadi *et al* 2018). Among all the subjects, 182 patients (86 AFib) proceeded with two recordings, one physician-applied and one patient-applied. The remaining patients ($n = 118$) were either nervous, physically in

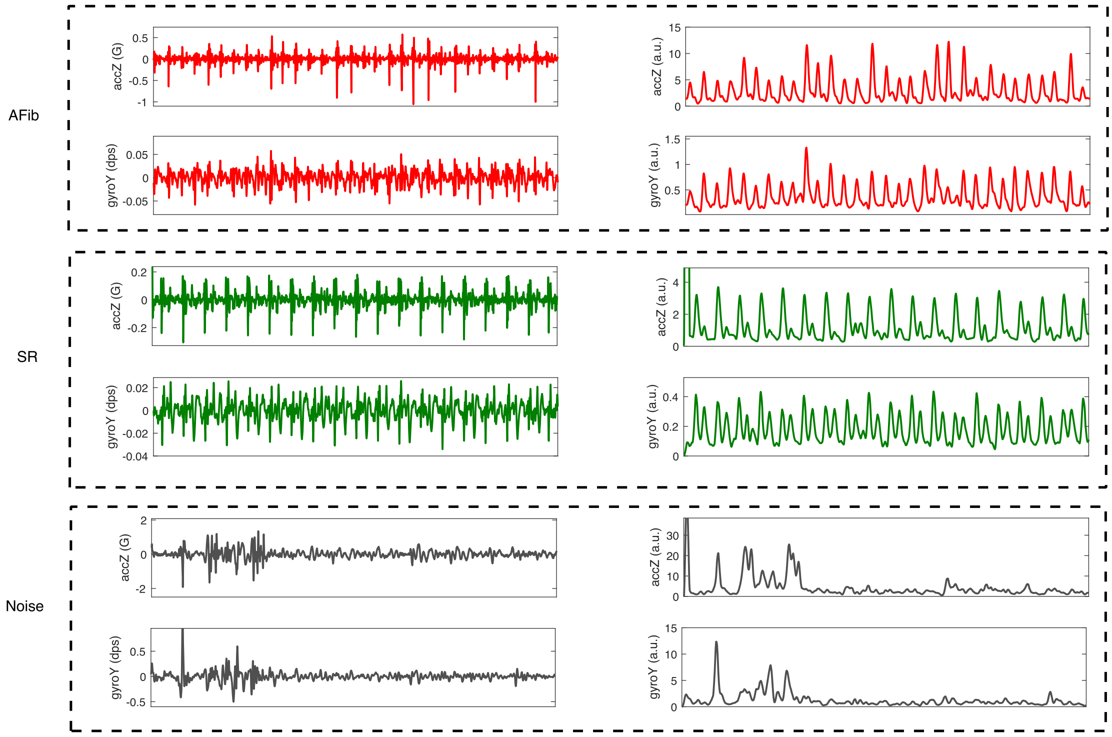


Figure 1. Smartphone-derived accelerometer (SCG) and gyroscope (GCG) signals (left side panels) with their corresponding envelope signals (right side panels) from an AFib, SR, and Noise recording.

poor condition, or not interested in performing the patient-applied measurement. During the recording, the subjects were advised to stay calm, silent, and motionless. The data logger application automatically terminated the recording after three minutes from the manual initiation. Those measurements in which either the patient or physician failed to obtain a valid recording were regarded as Noise class, for example, due to excessive movements, lack of concentration, delayed placement, phone drop, and poor placement. In total, we collected 827 sMCG measurements, of which 345 recordings were annotated as Noise category. Helsinki ethical declaration was strictly followed during all phases of the data collection. The study has also been reviewed and accepted by the Ethical Committee of the Hospital District of South-Western Finland.

A continuous 5-lead telemetry ECG (Philips IntelliVue MX40) was acquired simultaneously with the sMCG recordings and was used as the comparison method to assess the cardiac rhythm. The rhythm of each telemetry ECG was labeled as SR, AFib, or other by two independent cardiologists and the study investigator. In cases of inconsistency in the labeling of the two cardiologists, a third independent cardiologist made the final decision. The medical history of the subjects was collected from the electronic patient records. Following consent collection, background data were gathered. Afterward, a recording of three minutes was obtained using a Sony Xperia Z1 or Z5 smartphone. Detail descriptions of the measurement protocol and the demographics of the participants are available in Jaakkola *et al* (2018).

2.2. Pre-processing

All the recordings were acquired simultaneously with a 200 Hz sampling frequency. The signal processing starts with filtering each of the six data axes—corresponding to tri-axial SCG and tri-axial GCG—separately by a bandpass filter. A 4th order Butterworth filter with passband frequencies of 3–20 Hz was respectively applied on the GCG and SCG, allowing the removal of white noise and signals offset. We applied the filter forward and backward to every signal. In addition to the band-pass filtering, we obtained the pulse amplitude signal by computing the envelope of the SCG and GCG signals from all six channels and used them as six additional input dimensions. This envelope detection algorithm operates based on moving-average filtering as was described in (Tadi *et al* 2018). Figure 1 shows sample AFib, SR, and Noise class measurements together with their corresponding pulse-wave envelopes.

2.3. Dataset and sampling

Dataset preparation started by dividing individual MCG channels of each sensor/modality into a sequence of 10 s segments, each with 75% overlap. Next, we split the entire dataset into three disjoint subsets, train,

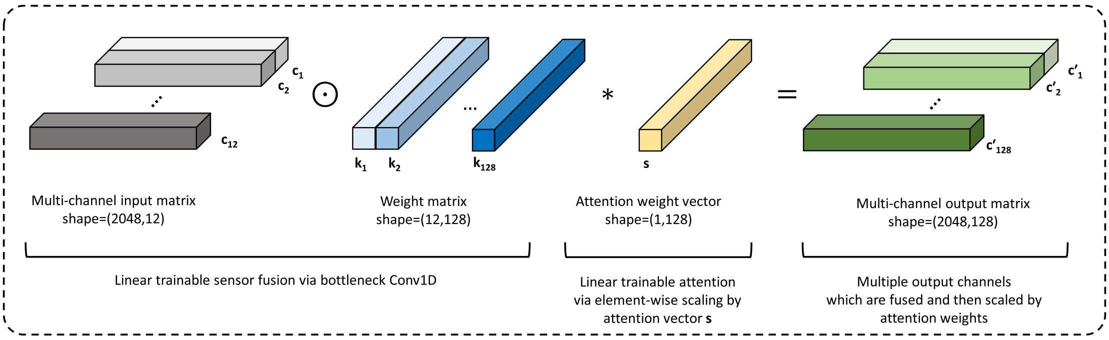

c1
c2
c12
⊙
k1
k2
k128
*
s
=
c'1
c'2
c'128
Multi-channel input matrix shape=(2048,12)
Weight matrix shape=(12,128)
Attention weight vector shape=(1,128)
Multi-channel output matrix shape=(2048,128)
Linear trainable sensor fusion via bottleneck Conv1D
Linear trainable attention via element-wise scaling by attention vector s
Multiple output channels which are fused and then scaled by attention weights

Figure 2. Schematic illustration of the implemented linear trainable sensor fusion mechanism. c_m for m between 1 and 12 are the input channels, k_n for n between 1 and 128 are the bottleneck convolutional kernels, s is the attention vector, and c'_n are the output channels.

Table 1. Number of measurements and segment counts used in this study.

	Train	Validation	Test	Total
AFib (patient-recorded)	149 (56)	45 (16)	42 (14)	236
SR (patient-recorded)	149 (62)	57(20)	40 (14)	246
Noise	183	38	25	246
Total (split percentage)	481 (66%)	140 (19%)	107 (15%)	728
10 s segments w/o augmentation	21 612	6993	2798	31 403
10 s segments with augmentation	43 224	13 986	5596	62 806

validation, and test. Finally, we randomly sampled measurements exhibiting each category; from the patient data, corresponding physician- and patient-recorded samples of individual subjects were selected to be included in only one of the subsets (train/valid/test) to avoid any data leakage. Table 1 shows the total number of sampled measurements and the corresponding number of windows (10 s segments) obtained from the incorporated measurements in each subset. It is worth mentioning that the segment duration was chosen based on our previous research studies (Tadi *et al* 2018, Mehrang *et al* 2019). The 75% overlap was chosen to enable the creation of a larger dataset.

2.4. Data augmentation

Since the availability of labeled signals for training the DNN models in this study was limited to the retrospective measurements, a data augmentation approach was utilized to expand the dataset size in training and validation sets. We considered the rotational data augmentation method proposed in Um *et al* (2017). Using rotational transformation on geometric vector data, we can generate synthetic data which can correspond to real-life observations. In detail, considering that acceleration and angular velocity are both geometric vectors, we can rotate the measured tri-axial SCG and tri-axial GCG signals around an arbitrary axis to resemble the rotation or placement variations of the measurement device. With this approach, we aimed to create synthetic data that closely correspond with real-life measurement scenarios that might have been absent in our original dataset.

2.5. Channel recalibration

Convolutional kernels are naturally designed for efficient transformation or filtering of the input data by sweeping along and transforming local receptive fields independently. In the case of the 1-dimensional convolutional (Conv1D) layer, the kernels sweep along the time axis. As a result, convolutional kernels are unable to learn the global channel-wise information (Hu *et al* 2018). To get a view of the global channel-wise information, SE blocks (Hu *et al* 2018) were introduced to our classification model which implemented an adaptive recalibration of channel-wise feature maps by modeling interdependencies between channels.

2.6. Sensor fusion

An efficient and learnable channel fusion technique can be implemented using the Conv1D layer. A bottleneck Conv1D layer contains kernels of length one, which can be used as a sample-by-sample channel fusion (aggregation) (Sandler *et al* 2018) when applied to multi-channel data with m number of channels. The fusion operation is implemented via dot product of a kernel and every multi-channel time sample of the data, which are both of shape $(1, m)$. We can have n number of these kernels, each computing and learning a different channel fusion. See figure 2 for a schematic illustration of the implemented sensor fusion mechanism and

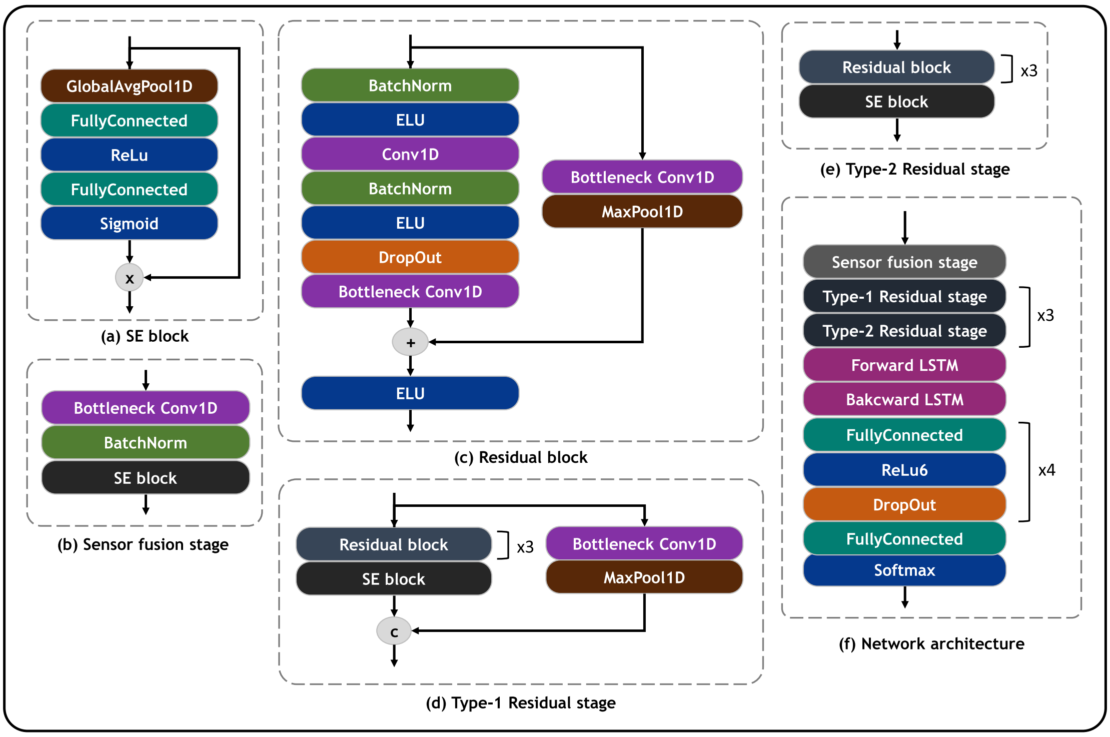


Figure 3. Building blocks of the proposed CRNN architecture. (a) SE block, (b) linear sensor fusion stage, (c) residual block, (d) type-1 residual stage which contains a dense connection, (e) type-2 residual stage without dense connection, (f) the full CRNN architecture.

figures 3(a) and (b) for the layer-by-layer illustration. In this figure, every input channel is denoted by c_m for m between 1 and 12. There are 128 kernels in the filter matrix, each denoted by k_n for n between 1 and 128. The channel recalibration is done via scaling the obtained feature maps from linear sensor fusion by vector s. In other words, every channel n in the feature map gets scaled by the scalar s_n at index n of vector s that is computed by an SE block. The output channels c_n' are the fused input channels that are each scaled by the attention coefficient s_n.

2.7. Deep neural network classifier

We considered a deep CRNN (Zihlmann *et al* 2017) to unveil arrhythmia pattern from the MCG signals, which takes as input the filtered and envelopes of tri-axial SCG and GCG signals and as output the expected class labels. Every input sample (segment) is of shape (2048, 12); while, the output is a one-hot encoded array of length 3. The one-hot encoded array is a binary vector that contains zeros everywhere except for the expected target class index. In our case, AFib, SR, and Noise were denoted by the class indices 0, 1, and 2, respectively. Accordingly, categorical cross-entropy was used for the loss function (Goodfellow *et al* 2016).

Our CRNN has been inspired by Hannun *et al* (2019) study in which a deep residual network (ResNet) architecture (He *et al* 2016) was adopted for the classification of ECG signals. The differences of the presented architecture with that of Hannun *et al* architecture are the use (1) CNN-based linear sensor fusion, (2) channel-attention by SE blocks, (3) stage-level dense connections, and (4) long-short term memory (LSTM) layers for temporal aggregation of features. See figure 3 for all the different building blocks of the proposed architecture.

We used linear bottleneck Conv1D layer, a batch normalization (BatchNorm) layer, and a SE block at the very beginning of the network to implement an end-to-end learnable early sensor fusion (Münzner *et al* 2017) which altogether constitute the sensor fusion stage as shown in figure 3(b). Furthermore, We opted to efficiently integrate such a sample-by-sample channel fusion into all residual blocks in our network by placing a linear bottleneck Conv1D layer followed by a 1-dimensional max-pooling (MaxPool1D) layer into the dense connection of the residual blocks as illustrated in figure 3(c). The residual blocks are grouped into two types of residual stages, type-1 and type-2, as depicted in figures 3(d) and(e). There is a dense connection in the type-1 residual stage, with Conv1D and MaxPool1D layers, which concatenates the input of the stage to its output. This dense connection helps to overcome the vanishing gradient (He *et al* 2016). All residual stages contain three residual blocks stacked on top of each other, plus an SE block placed at the end of the stage utilized for channel recalibration. For regularization, we used batch normalization and dropout layers extensively throughout the whole network architecture, as illustrated in figure 3. For training, Adam optimizer (Kingma and Ba 2014) with a learning rate of 0.001 was used.

Table 2. Notational representation of the confusion matrix for the classification problem in this study.

		Predicted labels			
		AFib	SR	Noise	Total
True labels	AFib	Aa	As	An	$\sum A$
	SR	Sa	Ss	Sn	$\sum S$
	Noise	Na	Ns	Nn	$\sum N$
	Total	$\sum a$	$\sum s$	$\sum n$	

Starting from the input, the network contains a sensor fusion stage, three pairs of type-1 and type-2 residual stages, a stack of forward and backward LSTM layers, four fully-connected layers that each is followed by Relu6 activation layer (Krizhevsky and Hinton 2010) and a dropout layer (Srivastava *et al* 2014), and ultimately a fully-connected layer followed by softmax activation for getting the classification probabilities. The whole network architecture is summarized and sketched in figure 3(f).

The hyper-parameters of the network were chosen as follows:

- For all the dropout layers, the dropout ratio was set to 0.15.
- For all the convolutional layers in type-1 and type-2 residual stages, the number of kernels was set to 32 except for the last residual block in the last type-2 residual stage, where there are 128 kernels.
- For all the dense connections in the type-1 residual stages, the convolutional kernels were set to 64. The number of units in both LSTM layers was set to 32.
- The number of units in fully-connected layers on top of LSTM layers was set to 64, 32, 16, 8, and 3, respectively.

Altogether there was 237 363 trainable and 3456 non-trainable parameters in the model. It is worth mentioning that no automatic hyper-parameter tuning was used due to insufficient computing power.

For the software tools, mainly Scikit-learn (Pedregosa *et al* 2011) and Tensorflow 2.4 (Abadi *et al* 2015) were used. The experiments were done on a desktop machine with an Nvidia RTX-2070 graphics card and 64 Gb RAM.

2.8. Experiments

In order to get an unbiased evaluation of the created CRNN model and the whole data processing pipeline, we repeated the training and testing processes for 10 fully random iterations initialized with different random seeds. Hereafter, we call these 10 random iterations with *evaluation iterations*. In every iteration, we trained the model for a maximum of 80 epochs, used categorical cross-entropy for the loss function, computed macro-averaged F1-score for measuring the goodness of fit, validated the trained model at the end of each epoch, and subsequently checkpointed the models at the end of each epoch based on the validation set macro-averaged F1-score. The model that provided the highest macro-averaged F1-score on the validation set was then automatically pulled from the model registry and used in the testing process. We stored the test set predictions for further in-depth statistical and performance analysis in each of the 10 evaluation iterations. Subsequently, various micro- and macro-averaged metrics were calculated using the obtained test set predictions. Hereafter, we refer to macro-averaged F1-score with *macro-F1-score* and similarly micro-averaged F1-score with *micro-F1-score*. We use the same shorthand for micro- and macro-averaged recall and precision.

2.9. Ablation study

The presented model architecture has been built by combining neural network components which were separately shown effective in a wide body of literature. To determine the utility of the proposed blocks and stages for the problem at hand, we performed an ablation study. All components that ended up in the presented model architecture were the ones that contributed to improving the predictive power, improving the convergence speed, and/or stability of model predictions when weights are initialized randomly. To limit the search space for optimal components, the performance of the goodness of fit was observed with and without each and every component separately. To measure the goodness of fit or the predictive power, models were trained for 10 randomly initialized iterations each 80 epochs, and the validation set macro-F1-score graph was plotted. The convergence speed, i.e. how fast the peak of predictive power is touched, was examined by checking the

Table 3. Performance of the CRNN classifier averaged over all three classes.

Resolution	segment-level		measurement-level	
Scores	Mean	95% CI	Mean	95% CI
Macro-F1-score	0.83	0.83–0.84	0.95	0.94–0.96
Macro-recall	0.89	0.88–0.90	0.95	0.94–0.96
Macro-precision	0.80	0.79–0.81	0.95	0.94–0.96
Micro-F1-score	0.88	0.87–0.89	0.95	0.94–0.96

validation set macro-F1-score graph and the epoch number at which the best model was checkpointed. The stability of the model was examined by inspecting the standard deviation of the validation set macro-F1-score values across the 10 random training rounds at each epoch. The best model architecture was expected to achieve the highest validation set macro-F1-score in as fewest epochs as possible and show lowest variations across the 10 random training rounds.

2.10. Performance metrics

Given the three-class classification problem at hand, a notational confusion matrix can be represented via table 2 (Clifford *et al* 2017). Macro-F1-score was calculated according to equation (1) considering the given notations in table 2. Similarly, macro-recall and macro-precision were calculated according to equations (2) and (3), respectively. In the case of multi-class classification, micro-F1-score, micro-recall, micro-precision, and accuracy are all equal and can be computed using equation (4).

$$Macro - F1 - score = \frac{\frac{2 * Aa}{\sum A + \sum a} + \frac{2 * Ss}{\sum S + \sum s} + \frac{2 * Nn}{\sum N + \sum n}}{3.}, \tag{1}$$

$$Macro - recall = \frac{\frac{Aa}{\sum A} + \frac{Ss}{\sum S} + \frac{Nn}{\sum N}}{3}, \tag{2}$$

$$Macro - precision = \frac{\frac{Aa}{\sum a} + \frac{Ss}{\sum s} + \frac{Nn}{\sum n}}{3}, \tag{3}$$

$$Micro - F1 - score = \frac{Aa + Ss + Nn}{\sum A + \sum S + \sum N}, \tag{4}$$

3. Results

We compared the performance of the presented CRNN classifier against the ground truth annotations of the test dataset by calculating statistical performance metrics, including micro- and macro-F1-score, precision, recall, and area under the receiver operating characteristic curve (ROC-AUC). In addition, we report the detection performance for *segment-level*, which we define as one rhythm class per segment, as well as *measurement-level*, which we define as one rhythm class per measurement. The measurement-level results were obtained by first gathering all the segment-level predictions of each unique measurement and then calculating the statistical mode of the rounded predictions. Such an averaging approach plays the role of a voting system.

Table 3 shows the performance metrics for both segment-level and measurement-level classification averaged over all the classes using micro- and macro-averaging. The obtained metrics show an acceptable macro-recall over all the classes for the segment-level classification. A more reliable performance was obtained for the measurement-level predictions as shown in the right-most column of table 3.

For a more in-depth view on the class-specific goodness of fit, table 4 shows the segment-level one-versus-all F1-score and ROC-AUC scores. The two classes, AFib and SR, were classified reliably, as shown by the high values of both F1-score and ROC-AUC. However, the Noise class classification suffered from low precision. Computing the same metrics for measurement-level predictions resulted in improved performance, in particular for the Noise class precision, as shown in table 5.

A pair of cumulative confusion matrices are created out of the test set predictions. Cumulative confusion matrices were obtained by performing element-wise summation on the 10 confusion matrices, each corresponding to one of the 10 evaluation iterations. Table 6 presents the segment-level classification and table 7 presents the measurement-level classification cumulative confusion matrices.

Table 4. Segment-level performance of the CRNN classifier per each class.

Class	AFib		SR		Noise	
Scores	Mean	95% CI	Mean	95% CI	Mean	95% CI
F1-score	0.89	0.88–0.90	0.89	0.88–0.90	0.72	0.71–0.74
ROC-AUC	0.96	0.95–0.97	0.97	0.96–0.97	0.98	0.98–0.98
Precision	0.92	0.91–0.94	0.89	0.86–0.91	0.60	0.57–0.62
Recall	0.86	0.84–0.89	0.90	0.88–0.91	0.91	0.89–0.94

Table 5. Measurement-level performance of the CRNN classifier per each class.

Class	AFib		SR		Noise	
Scores	Mean	95% CI	Mean	95% CI	Mean	95% CI
F1-score	0.94	0.93–0.96	0.94	0.92–0.96	0.96	0.95–0.97
ROC-AUC	0.98	0.98–0.99	0.99	0.99–1.00	0.99	0.99–1.00
Precision	0.96	0.94–0.98	0.95	0.92–0.98	0.93	0.92–0.94
Recall	0.93	0.90–0.96	0.94	0.92–0.96	0.99	0.98–1.00

Table 6. Segment-level cumulative confusion matrix obtained by element-wise summation of confusion matrices overall evaluation iterations.

		Predicted		
		AFib	SR	Noise
True	AFib	23 597	2814	889
	SR	1891	23 194	825
	Noise	85	152	2513

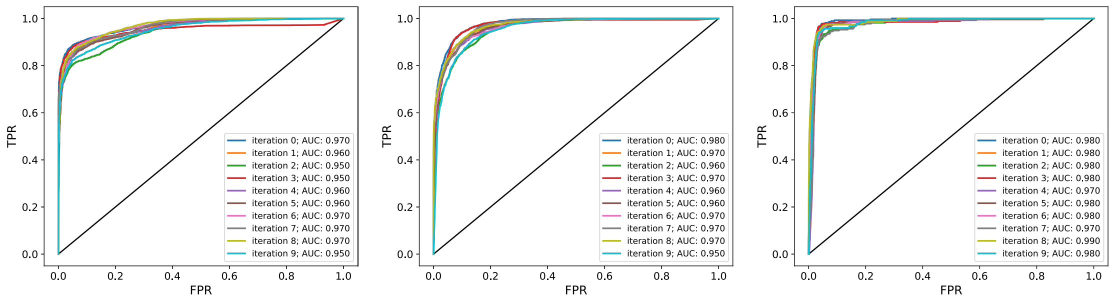
Table 7. Measurement-level cumulative confusion matrix obtained by element-wise summation of confusion matrices over all evaluation iterations.

		Predicted		
		AFib	SR	Noise
True	AFib	390	21	9
	SR	15	376	9
	Noise	2	0	248

The per-class ROC curves obtained from the evaluation iterations can be seen in figures 4(a)–(c) for the classes AFib, SR, and Noise, respectively. According to the obtained ROC-AUC values, the presented CRNN classifier captured meaningful patterns for all three classes across all the evaluation iterations without significant variation in the performance.

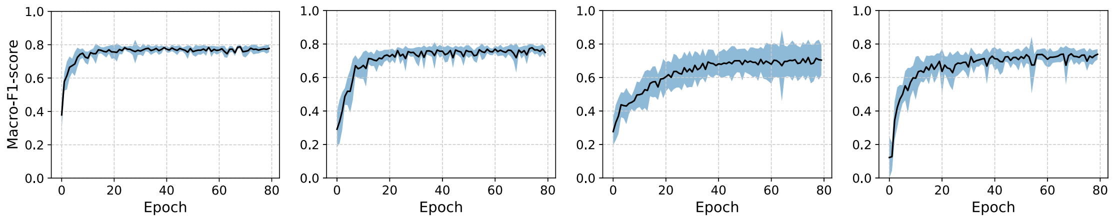
3.1. Ablation study results

Figure 5 concisely presents the ablation study results for a few of the model architecture building blocks, namely SE block, sensor fusion stage, and stage-level dense connections. We excluded residual blocks and LSTM layers from the presented ablation study results as they have been investigated sufficiently in the literature (Somani *et al* 2021). In figure 5, the thick line in the middle of the shadows represents the arithmetic mean of the validation set macro-F1-score across the 10 evaluation experiments performed for each ablated model. The shadows around the graphs represent the standard deviation of the metric. These shadows provide us with a qualitative view of the stability of the model architecture across different training rounds. To quantify the predictive power and the speed of convergence of each of the four model architectures in the ablation study, we calculated the arithmetic mean and 95% confidence interval of the maximum macro-F1-score across the 10 evaluation experiments.



(a) AFib ROC curve. (b) SR ROC curve. (c) Noise ROC curve.

Figure 4. Class-wise ROC curves obtained from the evaluation iterations.



(a) (b) (c) (d)

Figure 5. Predictive power and stability of (a) the full model versus models without (b) sensor fusion stage, (c) SE blocks, or (d) stage-level dense connections. The predictive power of the models is quantified by arithmetic mean of the validation set macro-F1-score. The stability of the models is quantified by the shadows around the graphs which represent the standard deviation of the metric across the 10 randomly initialized training experiments.

Similarly, we provide the same statistics of the epoch number where the maximum macro-F1-score has occurred. These statistics are shown in table 8.

4. Discussion

Our study is the first demonstration of an end-to-end DNN-based classification approach to AFib detection using smartphone MCG. In this study, we implemented a SE mechanism for channel recalibration, a trainable convolutional-based sensor fusion, a deep CRNN for spatiotemporal feature extraction, and fully-connected neural networks for the classification. The utility of the building blocks we have added to the widely adopted plain CRNN architecture (Somani *et al* 2021) was thoroughly tested via an ablation study and the effect of each block was separately shown in figure 5 and table 8. In particular, in comparison with the full model, a model without a sensor fusion stage was less stable and needed 16 more epochs on average to achieve its peak performance. The model without SE blocks, was hugely unstable, needed 20 more epochs on average to reach its peak performance, and was unable to achieve the peak performance level of the full model. Similarly, the model without stage-level dense connections was less stable, needed 22 more epochs on average to accomplish its peak performance, and was incapable of achieving the peak performance level of the full model.

The classification task was performed at two resolutions, segment-level and measurement-level. Even though we had only measurement-level annotations and each measurement was on average three minutes long, we had to perform segmentation on the data to limit the number of time samples fed to the CRNN classifier. Following our previous contributions, we segmented the data into 10 s segments (Tadi *et al* 2018, Mehrang *et al* 2019).

The reported measurement-level classification performances in this study are comparable to other screening modalities such as ECG and PPG (Ramkumar *et al* 2018, Zungsontiporn and Link 2018). Furthermore, performance levels obtained in this study were almost at the same level as with that of 2017 Physionet Challenge (Clifford *et al* 2017), where a total of 12 186 single lead ECG measurements were analyzed to classify AFib, SR, Noise, and Other rhythm classes. The overall macro-F1-score for the top 11 algorithms in the 2017 Physionet challenge was approximately equal to 0.83 (Clifford *et al* 2017). Thus, we see smartphone MCG as a complementary AFib detection technique in settings where ECG recording is not feasible.

Rhythm classification using mechanical functioning of the heart, including ballistocardiography (BCG) and MCG, has been previously done by combining feature engineering and machine learning (Bruser *et al* 2012,

Table 8. Statistics of the validation set maximum macro-F1-score and the corresponding epoch number of (a) the full model versus models without (b) sensor fusion stage, (c) SE blocks, or (d) stage-level dense connections. The arithmetic mean and the 95% CIs are computed across 10 randomly initialized training experiments.

Model architecture	Max. macro-F1-score		Epoch	
	Mean	95% CI	Mean	95% CI
Full model	0.81	0.80–0.81	44	30–59
W/O sensor fusion stage	0.8	0.80–0.80	60	48–71
W/O SE blocks	0.75	0.70–0.80	64	51–78
W/O stage-level dense connections	0.77	0.76–0.79	66	59–73

Lahdenoja *et al* 2017, Tadi *et al* 2018, Mehrang *et al* 2020). Naturally, designing, implementing, and experimenting with hand-crafted features were the first steps in the approaches proposed by previous studies. In contrast to those, DNNs provide us with automatic end-to-end feature extraction and classification (Andreotti *et al* 2017). As shown in this study, the capabilities of the DNNs can be extended beyond the ordinary automatic feature extraction by incorporating sensor fusion and channel recalibration mechanisms. These all facilitate knowledge discovery and pattern recognition with less human intervention and domain knowledge prerequisites (Somani *et al* 2021). One drawback of DNN based approaches is the computational load and the size of the datasets they need to be trained with. DNNs are purely data-driven; therefore, it is the content of the datasets that mainly derives and constrains the learning process (Somani *et al* 2021). Similar to the other DNN based classification use-cases, the key challenge for MCG analysis is not necessarily the computational load, but the availability of sufficiently sized datasets with high-quality and high-resolution annotations.

When CNNs are adopted for feature extraction, feature learning is done in an end-to-end fashion together with the classification (Goodfellow *et al* 2016). When training by plain stochastic gradient descent or its variants, as the most popular optimization algorithm for training DNNs (Le *et al* 2011), we need precise and high-resolution annotations. If the proportion of imprecise annotations increases, the optimizer gets confused and, as a result, cannot find the optimal latent space and decision boundaries (Nigam *et al* 2020). This, in turn, results in misclassifications for the inputs which are located close to the ground truth class boundaries. In our study, the two classes, AFib and SR, were annotated by a team of cardiologists, while a single senior researcher only annotated the Noise class. In addition, the size of the Noise class in our dataset was quite limited compared with the other two classes. Moreover, by definition, the Noise class covers a wide variety of measurement failure conditions of which some might be under-represented in our dataset. The size and the potentially inconsistent annotations of the Noise class were most likely the root causes of the low precision score of the segment-level Noise class classification. Despite the potentially inconsistent annotations, the measurement-level predictions were sufficiently accurate and comparable to our previous contributions (Mehrang *et al* 2019).

To improve the model's generalization, we increased the size of our training and validation sets with a data augmentation scheme tailored to the MCG data at hand. Specifically, we deployed segmentation with overlap (Tadi *et al* 2018, Mehrang *et al* 2019) and rotation augmentation technique (Um *et al* 2017) which is suitable for the data that hold a geometric description. Our observations showed that data augmentation helped to prevent overfitting and also improved the overall classification performance. However, further inclusion of augmented data did not improve the segment-level Noise classification results, primarily because of the inconsistent ground truth labels. We postpone the relabeling and addition of more consistently labeled Noise class data to the future contributions. Besides, more advanced unsupervised (Nurmaini *et al* 2019), generative adversarial networks (Yoon *et al* 2019), weakly supervised (Tong *et al* 2021), and self-supervised learning (Jawed *et al* 2020, Sarkar and Etemad 2020) techniques can be adapted to improve the process of feature learning without being constrained by the quality of the human-generated annotations.

Automated detection of cardiac rhythms is rapidly growing with the emergence of mobile and wearable devices that facilitate personalized monitoring and early detection of life-threatening conditions. Modern smartphone devices and mobile applications are profoundly enriching to serve the growing healthcare needs by being affordable, non-invasive, and easy to use. Sensor-rich smartphones are today accessible to most people worldwide, offering ubiquitous heart monitoring even without acquiring extra peripherals via MCG signals. As the number of users, amount of data, and complexity of the gathered data are growing, advanced data-driven knowledge discovery techniques are highly demanded.

5. Conclusion

Smartphone MCG devices may offer a practical and cost-efficient screening and monitoring alternative for AFib which can complement the other monitoring modalities. Analysis of multi-channel MCG data via deep learning facilitates the automatic and scalable extraction of the potential pathological conditions. The proposed CRNN architecture delivered promising AFib classification performance, proving the applicability of data-driven knowledge discovery techniques on MCG data. With the adoption of these data-driven techniques, we can improve the performance of AFib detection at scale, and as a result, increase the reliability of AFib screening and monitoring.

Acknowledgments

The authors would like to thank all the participants of the study. This work was supported in part by the Academy of Finland under Grant 290 930, in part by the Finnish Foundation for Technology Promotion under Grant 6756, and in part by the Nokia Foundation under Grant 202 010 415. Dr. Jafaritadi is thankful to Ulla Tuominen Foundation and Turku University Hospital State Research fund for supporting his postdoctoral fellowship at Stanford University.

Conflict of interests

Tero Koivisto and Mikko Pänkäälä are founders of Precordior which offers seismo- and gyrocardiography based cardiac monitoring products. The rest of the authors declare no conflict of interest.

ORCID iDs

Saeed Mehrang https://orcid.org/0000-0002-2092-2868
Mojtaba Jafari Tadi https://orcid.org/0000-0002-4085-4057
Timo Knuutila https://orcid.org/0000-0001-5390-3594
Jussi Jaakkola https://orcid.org/0000-0002-3398-4283
Samuli Jaakkola https://orcid.org/0000-0001-5944-6814
Tuomas Kiviniemi https://orcid.org/0000-0002-0908-3741
Tuija Vasankari https://orcid.org/0000-0002-1862-8133
Juhani Airaksinen https://orcid.org/0000-0002-0193-568X
Tero Koivisto https://orcid.org/0000-0003-3792-8999
Mikko Pänkäälä https://orcid.org/0000-0001-6108-9975

References

Abadi M *et al* 2015 TensorFlow: large-scale machine learning on heterogeneous systems *A System for {Large-Scale} Machine Learning In 12th USENIX symposium on operating systems design and implementation (OSDI 16)* pp. 265–83 (https://tensorflow.org/)

Andreotti F, Carr O, Pimentel M A F, Mahdi A and De Vos M 2017 Comparing feature-based classifiers and convolutional neural networks to detect arrhythmia from short segments of ecg *2017 Computing in Cardiology (CinC) IEEE* pp 1–4

Baldi P 2018 Deep learning in biomedical data science *Annu. Rev. Biomed. Data Sci.* **1** 181–205

Barrett P M, Komatireddy R, Haaser S, Topol S, Sheard J, Encinas J, Fought A J and Topol E J 2014 Comparison of 24 h holter monitoring with 14 d novel adhesive patch electrocardiographic monitoring *Am. J. Med.* **127** 95–e11

Bruser C, Diesel J, Zink M D H, Winter S, Schauerte P and Leonhardt S 2012 Automatic detection of atrial fibrillation in cardiac vibration signals *IEEE J. Biomed. Health Inform.* **17** 162–71

Camm A J *et al* 2010 Guidelines for the management of atrial fibrillationthe task force for the management of atrial fibrillation of the european society of cardiology (ESC) *Eur. Heart J.* **31** 2369–429

Clifford G D, Liu C, Moody B, Li-Wei H L, Silva I, Li Q, Johnson A E and Mark R G 2017 Af classification from a short single lead ecg recording: the physionet/computing in cardiology challenge *2017 Computing in Cardiology (CinC) IEEE* pp 1–4

Elisa C-G, Angel O and Jaume R 2010 Heart failure in acute ischemic stroke *Curr. Cardiol. Rev.* **6** 202–13

Goodfellow I, Bengio Y and Courville A 2016 *Deep Learning* (Cambridge, MA: MIT Press)

Gregory Y H L and Antonio T-M 2006 Management of atrial fibrillation *Heart* **92** 1177–82

Hannun A Y, Rajpurkar P, Haghpanahi M, Tison G H, Bourn C, Turakhia M P and Ng A Y 2019 Cardiologist-level arrhythmia detection and classification in ambulatory electrocardiograms using a deep neural network *Nat. Med.* **25** 65–9

He K *et al* 2016 Deep residual learning for image recognition *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition* pp 770–8

Hendrikx T, Rosenqvist M, Wester P, Sandström H and Hörnsten R 2014 Intermittent short ECG recording is more effective than 24 h holter ecg in detection of arrhythmias *BMC Cardiovasc. Disorders* **14** 1–8

Hindricks G *et al* (ESC Scientific Document Group 2020 2020 ESC Guidelines for the diagnosis and management of atrial fibrillation developed in collaboration with the European Association for Cardio-Thoracic Surgery (EACTS): the Task Force for the diagnosis and management of atrial fibrillation of the European Society of Cardiology (ESC) Developed with the special contribution of the European Heart Rhythm Association (EHRA) of the ESC *Eur. Heart J.* **42** 373–98

Hochreiter S and Schmidhuber J 1997 Long short-term memory *Neural Comput.* **9** 1735–80

Hu J, Shen L and Sun G 2018 Squeeze-and-excitation networks *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition* pp 7132–41

Huang G, Liu Z, Van Der Maaten L and Weinberger K Q 2017 Densely connected convolutional networks *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition* pp 4700–8

Jaakkola J, Jaakkola S, Lahdenoja O, Hurnanen T, Koivisto T, Pänkäälä M, Knuutila T, Kiviniemi T O, Vasankari T and Airaksinen K E J 2018 Mobile phone detection of atrial fibrillation with mechanocardiography: the MODE-AF study (mobile phone detection of atrial fibrillation) *Circulation* **22** 108–18

Jawed S, Grabocka J and Schmidt-Thieme L 2020 Self-supervised learning for semi-supervised time series classification *Pacific-Asia Conference on Knowledge Discovery and Data Mining* (Berlin: Springer) pp 499–511

Kingma D P and Ba J 2014 Adam: a method for stochastic optimization arXiv:1412.6980 (https://doi.org/10.48550/arXiv.1412.6980)

Kirchhof P *et al* 2016 2016 ESC Guidelines for the management of atrial fibrillation developed in collaboration with EACTSy *Eur. Heart J.* **37** 2893–962

Krizhevsky A and Hinton G 2010 Convolutional deep belief networks on cifar-10 *Unpublished Manuscript* **40** 1–9

Kumar P and Sharma V K 2020 Detection and classification of ecg noises using decomposition on mixed codebook for quality analysis *Healthcare Technol. Lett.* **7** 18–24

Lahdenoja O *et al* 2017 Atrial fibrillation detection via accelerometer and gyroscope of a smartphone *IEEE J. Biomed. Health Inform.*

Lau J K, Lowres N, Neubeck L, Brieger D B, Sy R W, Galloway C D, Albert D E and Freedman S B 2013 iphone ECG application for community screening to detect silent atrial fibrillation: a novel technology to prevent stroke *Int. J. Cardiol.* **165** 193-194

Le Q V, Ngiam J, Coates A, Lahiri A, Prochnow B and Ng A Y 2011 On optimization methods for deep learning *ICML* 265–72

Mehrang S *et al* 2019 Reliability of self-applied smartphone mechanocardiography for atrial fibrillation detection *IEEE Access* **7** 146801–12

Mehrang S *et al* 2020 Classification of atrial fibrillation and acute decompensated heart failure using smartphone mechanocardiography: a multi-label learning approach *IEEE Sensors J.* **20** 7957–68

Miotto R, Wang F, Wang S, Jiang X and Dudley J T 2017 Deep learning for healthcare: review, opportunities and challenges *Briefings Bioinform.* **19** 1236–46

Münzner S, Schmidt P, Reiss A, Hanselmann M, Stiefelhagen R and Dürichen R 2017 Cnn-based sensor fusion techniques for multimodal human activity recognition *Proc. of the 2017 ACM International Symp. on Wearable Computers* pp 158–65

Nigam N, Dutta T and Gupta H P 2020 Impact of noisy labels in learning techniques: a survey *Advances in Data and Information Sciences* (Berlin: Springer) pp 403–11

Nurmaini S, Umi Partan R, Caesarendra W, Dewi T, Naufal Rahmatullah M, Darmawahyuni A, Bhayyu V and Firdaus F 2019 An automated ecg beat classification system using deep neural networks with an unsupervised feature extraction technique *Appl. Sci.* **9** 2921

Park C, Took C C and Seong J-K 2018 *Mach. Learn. Biomed. Eng.* **137** 1524–27

Pedregosa F *et al* 2011 Scikit-learn: machine learning in Python *J. Mach. Learn. Res.* **12** 2825–30

Perez M V *et al* 2019 Large-scale assessment of a smartwatch to identify atrial fibrillation *New Engl. J. Med.* **381** 1909–17

Pyakillya B, Kazachenko N and Mikhailovsky N 2017 Deep learning for ecg classification *Journal of Physics : Conference Series* 913 (Bristol: IOP Publishing) p 012004

Ramkumar S, Nerlekar N, D'Souza D, Pol D J, Kalman J M and Marwick T H 2018 Atrial fibrillation detection using single lead portable electrocardiographic monitoring: a systematic review and meta-analysis *BMJ Open* **8** e024178

Sandler M, Howard A, Zhu M, Zhmoginov A and Chen L-C 2018 Mobilenetv2: inverted residuals and linear bottlenecks *Proc. of the IEEE Conf. on Computer Vision and Pattern recognition* pp 4510–20

Sarkar P and Etemad A 2020 Self-supervised learning for ecg-based emotion recognition *ICASSP 2020-2020 IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP). IEEE* pp 3217–21

Somani S *et al* 2021 Deep learning and the electrocardiogram: review of the current state-of-the-art *EP Europace* **23** 1179–91

Srivastava N, Hinton G, Krizhevsky A, Sutskever I and Salakhutdinov R 2014 Dropout: a simple way to prevent neural networks from overfitting *J. Mach. Learn. Res.* **15** 1929–58

Suresh P, Narayanan N, Pranav C V and Vijayaraghavan V 2020 End-to-end deep learning for reliable cardiac activity monitoring using seismocardiograms arXiv:2010.05662

Tadi M J, Lehtonen E, Saraste A, Tuominen J, Koskinen J, Teräs M, Airaksinen J, Pänkäälä M and Koivisto T 2017 Gyrocardiography: a new non-invasive monitoring method for the assessment of cardiac mechanics and the estimation of hemodynamic variables *Sci. Rep.* **7** 6823

Tadi M J *et al* 2018 Comprehensive analysis of cardiogenic vibrations for automated detection of atrial fibrillation using smartphone mechanocardiograms *IEEE Sensors J.* **19** 2230–42

Tieleman R G, Plantinga Y, Rinkes D, Bartels G L, Posma J L, Cator R, Hofman C and Houben R P 2014 Validation and clinical use of a novel diagnostic device for screening of atrial fibrillation *Europace* **16** 1291–95

Tong Y, Sun Y, Zhou P, Shen Y, Jiang H, Sha X and Chang S 2021 Locating abnormal heartbeats in ecg segments based on deep weakly supervised learning *Biomed. Signal Process. Control* **68** 102674

Um T T, Pfister F M J, Pichler D, Endo S, Lang M, Hirche S, Fietzek U and Kulić D 2017 Data augmentation of wearable sensor data for parkinson's disease monitoring using convolutional neural networks *Proc. of the XIX ACM Int. Conf. on Multimodal Interaction* pp 216–20

Valentin F *et al* 2001 ACC/AHA/ESC guidelines for the management of patients with atrial fibrillation: executive summary: a report of the american college of cardiology/american heart association task force on practice guidelines and the european society of cardiology committee for practice guidelines and policy conferences *J. Am. Coll. Cardiol.* **38** 1231–65

Yoon J, Jarrett D and van der Schaar M 2019 Time-series generative adversarial networks *Advances in Neural Information Processing Systems* **32**

Zanetti J M and Tavakolian K 2013 Seismocardiography: past, present and future *2013 XXXV Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC) IEEE* pp 7004–7

Zhang J, Liu A, Gao M, Chen X, Zhang X and Chen X 2020 Ecg-based multi-class arrhythmia detection using spatio-temporal attention-based convolutional recurrent neural network *Artif. Intell. Med.* **106** 101856

Zihlmann M, Perekrestenko D and Tschannen M 2017 Convolutional recurrent neural networks for electrocardiogram classification *2017 Computing in Cardiology (CinC). IEEE* pp 1–4

Zungsontiporn N and Link M S 2018 Newer technologies for detection of atrial fibrillation *Bmj* **363** k3946