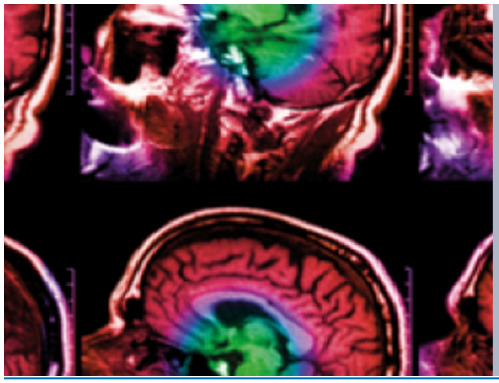
Analysis of an adaptive lead weighted ResNet for multiclass classification of 12-lead ECGs

PAPER • OPEN ACCESS

Analysis of an adaptive lead weighted ResNet for multiclass classification of 12-lead ECGs



IOP Publishing Physiol. Meas. 43 (2022) 034001 https://doi.org/10.1088/1361-6579/ac5b4a

Physiological Measurement

IPEM Institute of Physics and Engineering in Medicine

OPEN ACCESS

RECEIVED
8 December 2021

REVISED
23 February 2022

ACCEPTED FOR PUBLICATION
7 March 2022

PUBLISHED
4 April 2022

PAPER

Analysis of an adaptive lead weighted ResNet for multiclass classification of 12-lead ECGs

Z Zhao[1,2], D Murphy[1], H Gifford[3], S Williams[4], A Darlington[1], S D Relton[4], H Fang[5] and D C Wong[1]

[1] University of Manchester, United Kingdom
[2] Xi'an Jiaotong University, Xi'an, People's Republic of China
[3] University of Exeter, Exeter, United Kingdom
[4] University of Leeds, Leeds, United Kingdom
[5] Loughborough University, Loughborough, United Kingdom

E-mail: david.wong@manchester.ac.uk

Keywords: ECG, deep neural network, machine learning, electrocardiogram, signal processing, PhysioNet

Supplementary material for this article is available online

Abstract

Background. Twelve lead ECGs are a core diagnostic tool for cardiovascular diseases. Here, we describe and analyse an ensemble deep neural network architecture to classify 24 cardiac abnormalities from 12 lead ECGs. *Method*. We proposed a squeeze and excite ResNet to automatically learn deep features from 12-lead ECGs, in order to identify 24 cardiac conditions. The deep features were augmented with age and gender features in the final fully connected layers. Output thresholds for each class were set using a constrained grid search. To determine why the model made incorrect predictions, two expert clinicians independently interpreted a random set of 100 misclassified ECGs concerning left axis deviation. *Results*. Using the bespoke weighted accuracy metric, we achieved a 5-fold cross-validation score of 0.684, and sensitivity and specificity of 0.758 and 0.969, respectively. We scored 0.520 on the full test data, and ranked 2nd out of 41 in the official challenge rankings. On a random set of misclassified ECGs, agreement between two clinicians and training labels was poor (clinician 1: $\kappa = -0.057$, clinician 2: $\kappa = -0.159$). In contrast, agreement between the clinicians was very high ($\kappa = 0.92$). *Discussion*. The proposed prediction model performed well on the validation and hidden test data in comparison to models trained on the same data. We also discovered considerable inconsistency in training labels, which is likely to hinder development of more accurate models.

1. Introduction

The 12-lead electrocardiogram (ECG) provides critical information that assists in identifying cardiac abnormalities. The signal from each of the 12 leads corresponds to the hearts electrical activity from a distinct angle that can be mapped to the anatomy of the heart. A skilled interpreter can therefore use ECG signals from multiple leads to localise the source and nature of a cardiac abnormality. Expert cardiologists can identify abnormalities with high accuracy. A recent systematic review highlighted how the accuracy of human expert interpretation may be as high as 95% in a controlled setting in which the final diagnosis was known (Cook *et al* 2020).

However, expert-level human ECG interpretation is limited by the availability of a trained cardiologist and the time required to synthesize information from the 12-lead signal (and to document their findings). In clinical practice, the absence of cardiologists means that generalist and specialist clinicians are required to interpret ECGs at the bedside in order to direct acute management and consider referral to specialist cardiology services. However, non-specialist clinicians are demonstrably less accurate compared to trained cardiologists (Salerno *et al* 2003).

Computer-aided methods for ECG interpretation have been suggested as one approach for circumventing these resource constraints. Historically, the accuracy of these methods has been poorer than humans (Estes 2013). For instance, Anh *et al* (2006) reported how 19% of atrial fibrillation were considered to be false positives when reviewed by a cardiologist.

Traditional machine learning approaches, in which salient features of the ECG signal are first identified, have been successful for some use cases. As far back as 1991, the performance of some methods were almost as accurate as cardiologists, for a limited subset of clinical conditions (Willems *et al* 1991). However, such methods have frequently struggled to correctly interpret ECG with arrhythmias, conduction disorders and pacemaker rhythms (Schläpfer and Wellens 2017).

Modern deep learning methods may be able to improve interpretation accuracy. Until recently, the use of such techniques for 12-lead ECGs has been impractical due to the shortage of labelled training data. There remains room for improvement over initial promising results (Ribeiro *et al* 2020). The public release of a new large labelled data set presents a fresh opportunity to revisit this challenging problem (Alday *et al* 2020).

Here, we consider the task of cardiac abnormality classification from 12-lead ECG recordings of varying sampling frequency and duration. We have tackled this problem by developing a deep neural network architecture (Zhao *et al* 2020), which was submitted to the 2020 Physionet Challenge.

Our architecture acknowledges the importance of the spatial relationship between the ECG channels by using a squeeze-and-excitation (SE) block. In this extended analysis, we present a deeper investigation into the strengths and weaknesses of this approach, including the use of expert clinical knowledge to determine why examples may be misclassified.

2. Methods

Our objective was to create a model that could accurately classify 12-lead ECG recordings into one or more of 27 clinical diagnoses. In practice, we considered only the 24 clinical diagnoses shown in table 3 in the competition description paper (Alday *et al* 2020). Each class corresponds to a single ICD-10 code, with the exceptions of classes 'PVC' (Premature Ventricular Contraction), 'PAC' (Premature Atrial Contraction), and 'RBBB' (Right Bundle Branch Block). These classes corresponded to two clinical codes which we considered to be identical. The trained model and training code are available at: (https://github.com/ZhaoZhibin/Physionet2020model).

2.1. Dataset

We trained our model used publicly-available data released for the 2020 Physionet Challenge. Alday *et al* (2020) provide detailed information about the data set. This work preceded the 2021 Challenge and therefore excludes the additional data made available in 2021 (Reyna *et al* 2021).

In brief, the training data set contains 43 101 12-lead ECGs and labels. ECG recordings are of variable duration (6–1800 s) and sampling frequency (257–1000 Hz), corresponding to variations in real-life practice. Each ECG was linked to a gender and age. Data were sourced from four consituent databases (CSPC, INCART, PTB, G12EC); the number of recordings from each location varied between 74 and 10 344 examples. The validation set contained 6630 examples, but no labels, from the two of these databases (CPSC, G12EC).

We tested the model on a hidden test data set of 16 630 ECGs. The set comprised of 6630 (40%) ECGs collected from two of the same locations as the training set (CPSC, G12EC), and 10 000 (60%) from a fifth undisclosed location. The fifth location was an American institution, in which mean age, and ratio of male to female sex, were similar to the rest of the test set. However, the sampling frequency was lower (300 Hz) than the vast majority of the training data.

Age and gender information for the training data (split by their constituent locations) and test data sets are provided by Alday *et al* (2020). Training and test data were matched as closely as possible for age, sex and diagnosis.

2.2. Pre-processing

We resampled all ECGs to the minimum frequency of 257 Hz using linear interpolation. To allow a fixed input size in the deep learning model, each ECG was set to be 4096 points. During training, we ensured this by zero-padding any shorter duration signals and randomly clipping any longer duration signals.

We scaled age into the range [0,1]. Gender was encoded using one-hot encoding and two additional mask variables represented missing values for age and gender (figure 2). No other pre-processing was undertaken. In particular, we highlight that the signals were not filtered, as we did not want to accidentally remove pertinent information.

2.3. Model architecture

After obtaining the input signals, we designed an improved ResNet to assign the 12-lead ECG recordings into one or more of the 24 diagnostic classes.

The improved ResNet can be decomposed into *feature extraction*, *feature fusion*, and the *classifier*. *Feature extraction* consists of one convolutional layer followed by a batch normalization (BN) layer, a ReLU activation

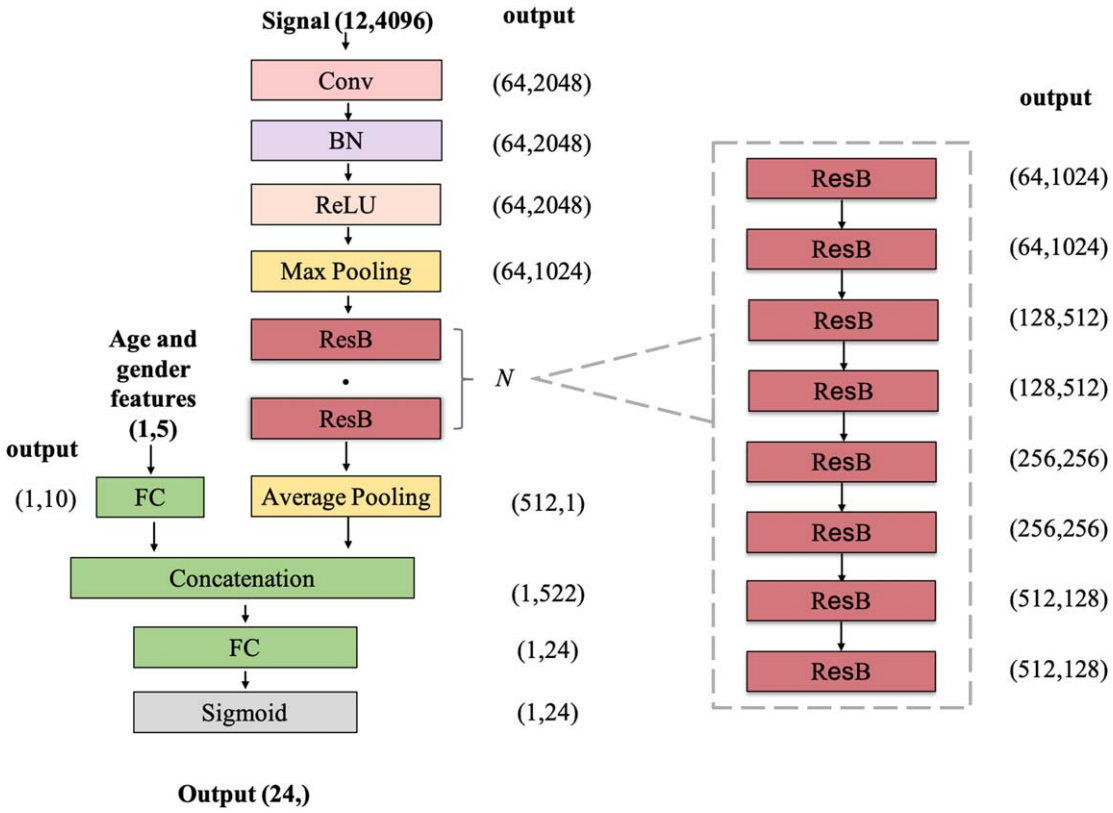


Figure 1. The proposed SE-Resnet network architecture. Raw ECG signals of length 4096 are input into the right branch of the network. The ResB blocks contain an SE block modification, as depicted in figure 2. The resulting features are concatenated with the output of a fully connected layer with age and gender feature inputs. The concatenated feature vector is the input to a fully connected layer that outputs the independent probability for each disease class.

function, a max pooling layer, $N = 8$ residual blocks (ResBs), each of which contains two convolutional layers and an SE block and an average pooling layer (figure 1). *Feature fusion* concatenates deep features from the feature extraction part and the additional age and gender information. These combined features are input to the *classifier*, which constitutes a fully connected (FC) layer and a Sigmoid layer, and outputs the probability of belonging to a disease class. An overview of our model is illustrated in figure 1.

Further details of the *feature extraction* portion are as follows. The first layer of the network and the initial two ResB units have 64 convolution filters. The number of filters increases by a factor of two for every second ResB unit. The feature dimension is halved after the max pooling layer, and then after the third, fifth, and seventh ResBs.

We used a relatively large kernel size of 15 in the first convolution layer, and a kernel size of 7 in subsequent layers. Previous work has shown, empirically, that large convolutional kernels lead to better performance for ECG classification (Hannun *et al* 2019).

We chose stacked ResBs to extract features from ECG data as they are easy to optimize and have previously been effective in generating discriminative features (He *et al* 2016). The structure of the modified ResB we use is illustrated in figure 2. It addresses the optimization problem by introducing a *deep residual learning* framework. Instead of directly learning the underlying mapping $H(\mathbf{X})$ (we assume that $\mathbf{X} \in \mathbb{R}^{C \times W}$ is the input of the modified ResB, where C is the lead dimension and W is the feature dimension), stacked layers in ResB approximate a residual function $F(\mathbf{X}) = H(\mathbf{X}) - \mathbf{X}$, where $\mathbf{X}$ is the input for ResB. Then the original function becomes $H(\mathbf{X}) = F(\mathbf{X}) + \mathbf{X}$. This is easier to optimize as the identity mappings $H(\mathbf{X}) = \mathbf{X}$ provides reasonable preconditioning.

We added a BN layer after each convolution layer to accelerate training. To reduce the likelihood of overfitting, we added a dropout layer with a drop out rate of 0.2 between the two convolutional kernels in each ResB.

In developing our model architecture, we considered that, for many cardiac conditions, human interpretation of 12-lead ECG involves reviewing a subset of leads. For instance expert clinical interpretation of LAD is typically dependent on information from only three leads (Hampton 1997). We therefore chose an architecture that provides flexibility to model the relative importance of each lead. One such approach, the SE

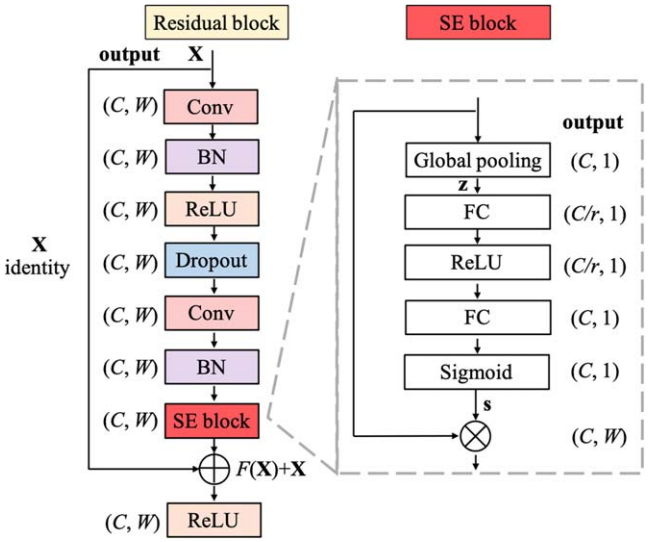


Figure 2. Residual block (ResB) and SE block, where W is the feature dimension, C is the lead dimension, and r is the reduction factor. © 2020 IEEE. Reprinted, with permission, from Zhao *et al* 2020. CC BY 4.0.

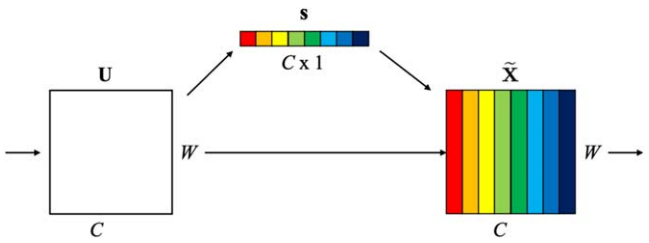


Figure 3. Visualisation of the working mechanism of the SE block, showing how it outputs a vector of weights, s, that correspond to each ECG lead.

block developed by Hu *et al* (2018), has been previously applied to image data. We integrated the SE block to our ResB, as depicted in figure 2, to model the spatial relationship between the ECG leads.

First, global spatial information is *squeezed* into a lead descriptor $\mathbf{z} \in \mathbb{R}^{C\times 1}$ by global average pooling, where C is the lead dimension. To make use of the information in $\mathbf{z}$, a gating mechanism with a Sigmoid activation is used to obtain the rescaling vector $\mathbf{s} \in \mathbb{R}^{C\times 1}$ that can be thought of as a weight for each ECG lead:

$$\mathbf{s} = \sigma(g(\mathbf{z}, \mathbf{W})) = \sigma(\mathbf{W}_2\delta(\mathbf{W}_1\mathbf{z}))), \tag{1}$$

where $\delta(\cdot)$ is the ReLU function, $\mathbf{W}_1 \in \mathbb{R}^{\frac{C}{r}\times C}$ refers to the weight parameters of the first FC layer, and $\mathbf{W}_2 \in \mathbb{R}^{C\times\frac{C}{r}}$ refers to the weight parameters of the second FC layer. The parameter $r = 16$ denotes the reduction factor, which controls the capacity and computational cost of the SE block. Specifically, the total number of additional weight parameters $\mathbf{W}_1$ and $\mathbf{W}_2$ is $\frac{2}{r}C^2$; without the reduction factor there would be $2C^2$ parameters. The final output of the SE block $\widetilde{\mathbf{X}} = [\tilde{\mathbf{x}}_1, \tilde{\mathbf{x}}_2, \ldots, \tilde{\mathbf{x}}_C]$ is calculated by the following lead-wise multiplication:

$$\tilde{\mathbf{x}}_c = s_c\mathbf{u}_c, \tag{2}$$

where $\mathbf{u}_c \in \mathbb{R}^{1\times W}$ denotes the feature map from lead c ($\mathbf{U}$ is the input of the SE block), and the scalar s_c is the rescaling parameter from lead c in $\mathbf{s}$. The working mechanism of the SE block is visualized in figure 3, showing how the rescaling parameter $\mathbf{s}$ acts as a set of lead weights.

The output of *feature extraction* results in 512 deep features. These are concatenated with 10 additional features generated from the output of an FC layer with 5 age and gender input features (1 × age, 2 × one-hot encoded gender, 2 × mask variables for missing data). We believed, *a priori*, that these were important features, given the clinical literature highlighting differing prevalence of certain heart conditions by age and gender (see, for instance, Feinberg *et al* 1995).

A final FC layer is used to complete classification from the 522 fused features. Finally, a Sigmoid function is used to scale classification results into [0, 1].

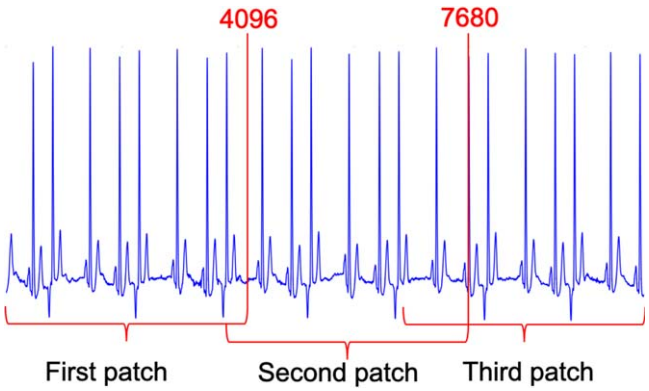


Figure 4. The input to our model consisted of 4096 sequential samples. Any signal greater than 4096 samples was segmented into epochs with an overlap of 256, as shown.

The training error for this multi-task problem was measured by the average binary cross-entropy loss:

$$\mathcal{L} = \frac{1}{n_{class}} \sum_{i=1}^{n_{class}} y_i \log(\hat{y}_i) + (1 - y_i)\log(1 - \hat{y}_i), \tag{3}$$

where $n_{class} = 24$ is the total number of classes, $\hat{y}_i$ denotes the output of the Sigmoid layer for the ith class, and y_i denotes the corresponding true class. All the classes were equally weighted, and problems of class-imbalance were addressed via setting individual classification thresholds for each class, as described below.

The loss was optimized using the Adam optimizer with an initial learning rate 0.003. The learning rate was reduced tenfold in the 20th and 40th epochs, and the model was trained for in total 50 epochs with a batch size of 64.

2.4. Decision threshold optimisation

The final layer of the proposed model is a (24, 1) Sigmoid layer that estimates the probability that the signal belongs to a disease class.

Heavy class imbalance means that a default threshold of $P = 0.5$ for each Sigmoid may be too insensitive. Rather than adjusting the model directly via the cost function, or by data resampling, we chose the pragmatic approach of adjusting the decision threshold. This assumes that the underlying representation, that is, the decision surface, is accurate. Kang *et al* (2020) have previously argued that accurate representation is possible in problems, like ours, with class imbalance.

For our multi-task problem, we determined thresholds by performing a constrained grid-search for each class individually, using $s_{normalized}$, a bespoke metric that was used to score the Physionet Challenge 2020 entries and is described by Alday *et al* (2020). This metric is a weighted accuracy that rewards incorrect classifications with similar risks or outcomes to the true class.

In detail, we first constrained thresholds for all classes to be equal. We set a temporary global threshold by searching in [0, 1] in steps of 0.1. We then adjusted this threshold individually for each class, by searching in steps of 0.01, with all other thresholds fixed.

The drawback of this approach is that it makes the simplifying assumption that each class is independent. However, we decided that searching for the joint set of optimal thresholds would be too consuming to determine in this case, as it would require searching in 24-dimensional space.

2.5. Model analysis

We used multi-label stratified five-fold cross-validation to assess the performance of the model.

For the validation and test signals, we continued to zero-pad any signals with fewer than 4096 samples. For signals longer than 4096 samples, we segmented the signals into multiple epochs with a fixed overlap, $O = 256$. An example with sample length 10 000 is depicted in figure 4. The number of epochs, P, for a single signal can be formulated as:

$$P = ceil\left(\frac{L - 4096}{4096 - O}\right) + 1, \tag{4}$$

where $ceil(\cdot)$ rounds a number upward to the nearest integer. We processed all P epochs, and used the mean of the output class probabilities to classify the recording.

Table 1. Model results with different thresholds using five-fold cross-validation. model-default: an SE-ResNet without any threshold optimization; model-final: an improved ResNet with thresholds optimized by constrained grid-search. The training set comprised of 43 101 examples from five locations. The validation set comprised of 6630 examples from 3 locations. The test set comprised of 16 630 examples from two locations; one of these locations was not used in the training or validation sets.

Method	Training set			Validation set	Test set
	Sens.	Spec.	$s_{normalized}$	$s_{normalized}$	$s_{normalized}$
model-default	0.599	0.986	0.630	0.607	
model-final	**0.758**	**0.969**	**0.684**	**0.672**	**0.520**

We reported sensitivity, specificity, and the challenge metric, $s_{normalized}$. Finally, we performed exploratory analysis to investigate the source of the classification errors. In post-hoc analysis, we selected a set of misclassified ECGs for a specific clinical condition, LAD. This condition was chosen as it is commonly determined using well-known heuristics, and we would therefore expect training labels for this condition to be reliable.

LAD occurs when the mean direction of the action potentials travelling through the ventricles at depolarisation (QRS axis) is less than $-30°$. Archetypal examples are recognised via a positive QRS complex in Lead I, and negative QRS complexes in Leads II and III (Hampton 1997).

A total of 100 examples (50 false positive and 50 false negatives) were selected at random. For each example, we asked two expert clinicians (HG,SW) with experience in ECG interpretation to determine whether LAD was present, not present, or whether it was unclear. The clinicians were Non-Cardiology specialists with core medical training including the MRCP postgraduate qualification.

We reported their agreement with each other and the training data labels using Cohen's κ.

3. Results

The final row of table 1 states the $s_{normalized}$ metric for the intermediate validation set and hidden test set for our final model. In addition, we report 5-fold cross-validation estimates of sensitivity, specificity, $f1$-score and $s_{normalized}$ based on the training data. The first row reports the model metrics in which examples are assigned a class if $P(\text{class}) > 0.5$ (i.e. no threshold adjustment). The second row reports metrics for the final model with full threshold adjustment. Only this final model was scored on the hidden test set. The final model was submitted by the team '*between a ROC and a heart place*', and was 2nd of 41 models officially entered (and 2nd of 70 models submitted to the challenge).

3.1. Comparison to other models using validation set

Per-class metrics for the validation set were provided by the challenge organisers. Table 2 summarises the average classification performance of the top 10 entries to the Physionet 2020 competition on the validation set, grouped by ECG class. All models had high $f1$-scores (>0.8) for Sinus Bradycardia (SB), Atrial Fibrillation (AF), Sinus Tachycardia (STach), and Complete Right Bundle Branch Block (CRBBB).

All models were poor ($f1 < 0.3$) at classifying Bradykinesia (Brady), Non-Specific Intra-Ventricular Conduction Delay (NSIVCD), Pacing Rhythm (PR), Pre-Ventricular Contraction (PVC), Right Axis Deviation (RAD), and T-wave Inversion (TInv). In most cases, poor performance can be explained by the relatively small number of training and validation examples. For instance, in the case of PR, there were 299 training and 2 validation examples.

The limited number of validation samples for Brady also explains the difference in performance between the classification of Brady ($f1 = 0.003$, $n = 1$) and Sinus Bradycardia (SB) ($f1 = 0.91$, $n = 860$). Given that SB is a common sub-type of Brady, we might otherwise expect classification metrics to be similar.

For others conditions, such as NSIVCD, where manual assessment is known to be challenging (Eschalier *et al* 2015), it is unsurprising that automated methods have poor performance. This will occur if either the key clinical features are unable to be represented by a deep neural network (e.g. if the features are too nuanced), or if the wrong features are encoded (e.g. if training data mislabelled). The exception to this is TInv ($f1 = 0.29$, $n = 438$). The uniform poor performance over all models is surprising, given that there many examples, and that the clinical feature is simple to recognise for human experts.

In comparison to the other top 10 models, our model had similar predictive power for all classes—the f-score was within 1 s.d. of the mean, for every class. More detailed per-class comparison of the models is not helpful, as minor differences may be due to model overfitting. Indeed, the model that generalised best to the test data had the lowest per-class metrics in the validation data (Natarajan *et al* 2020).

Table 2. Mean and standard deviation *f*-measure of each class, alongside number of training and validation examples, for the 10 top performing ($s_{normalized}$) entries to the 2020 Physionet challenge. Emboldened items indicate conditions that were poorly classified by all entries.

Class (training *n*, validation *n*)	SE-net	Top 10 Mean *f*-1	Top 10 std *f*-1
IAVB (2394, 552)	0.736	0.756	0.027
AF (3475, 552)	0.801	0.810	0.022
AFL (314, 109)	0.467	0.473	0.081
Brady (288, 1)	**0.000**	**0.003**	**0.008**
CRBBB (683, 18)	0.817	0.815	0.027
IRBBB (1611, 206)	0.495	0.463	0.060
LAnFB (1806, 110)	0.428	0.391	0.079
LAD (6086, 478)	0.672	0.636	0.053
LBBB (1041, 156)	0.752	0.754	0.031
LQRSV (556, 192)	0.573	0.393	0.177
NSIVCB (997, 96)	**0.133**	**0.13**	**0.057**
PR (299, 2)	**0.000**	**0.041**	**0.061**
PAC (1729, 459)	0.618	0.592	0.061
PVC (188, 178)	**0.301**	**0.286**	**0.048**
LQT (1513, 740)	0.585	0.570	0.059
QAb (1013, 239)	0.357	0.362	0.073
RAD (427, 38)	**0.208**	**0.243**	**0.056**
SA (1240, 236)	0.624	0.598	0.089
SB (2359, 860)	0.934	0.908	0.036
SNR (20846, 1100)	0.670	0.651	0.044
STach (2402, 648)	0.844	0.857	0.022
TAb (4673, 1119)	0.626	0.595	0.047
TInv(1112, 438)	**0.248**	**0.201**	**0.060**

We further note that our model architecture was very similar to the entry that placed third in the 2020 Physionet challenge (Zhu *et al* 2021). The primary difference between the two models, which may explain difference in performance, was our inclusion of age and gender features.

3.2. Model error examples

Having established that our model performed relatively well overall, we then investigated individual examples that our model misclassified. Figure 5(a) depicts an example that was labelled as having LAD in the training set, but was not identified by our model (i.e. a false negative). In this case, two independent clinical expert reviewers were both unable to determine whether LAD was present, due to low signal-to-noise ratio in the key I, II and III leads. Figure 5(b) depicts an example that was not labelled with LAD, but our model classified it is as LAD (a false positive). Additional clinical review (HG, SW) determined that this should have been labelled with LAD, due to the presence of a positive R peaks in Lead I and negative R peaks in Leads II and III.

From visual inspection of these, and other similar examples, we noted that discordance occurred because the neural network failed to recognise the key indicators of LAD (either due to noisy data, or inadequate representation) or that the neural network was correct, but the examples appear to be mislabelled. These two problems will place an upper-bound on the performance of any deep-learning model.

To explore this further and provide an initial attempt to quantify this problem, we randomly selected 50 examples of false-positive LAD and 50 examples of false-negative LAD. Two clinical experts classified the examples into {*LAD, no LAD, unsure*}. The raw results of these classifications are presented in supplementary material (available online at stacks.iop.org/PMEA/43/034001/mmedia) and summarised below. Table 3 shows the interrater agreement between the two clinicians. Of the 100 examples, at least one clinician was *unsure* in 18 cases. In the remaining 82 examples, the clinicians disagreed in only two instance—resulting in Cohen's $\kappa = 0.92$.

In contrast, both clinicians commonly disagreed with the training labels (table 4). HG disagreed in 47/87 cases (Cohen's $\kappa = -0.057$) and SW disagreed in 53/91 (Cohen's $\kappa = -0.159$), excluding cases in which the clinicians were unsure. For both clinicians, disagreements mainly occurred when the example was labelled with LAD, but the clinicians, and model, did not believe there was LAD (i.e. False Negatives).

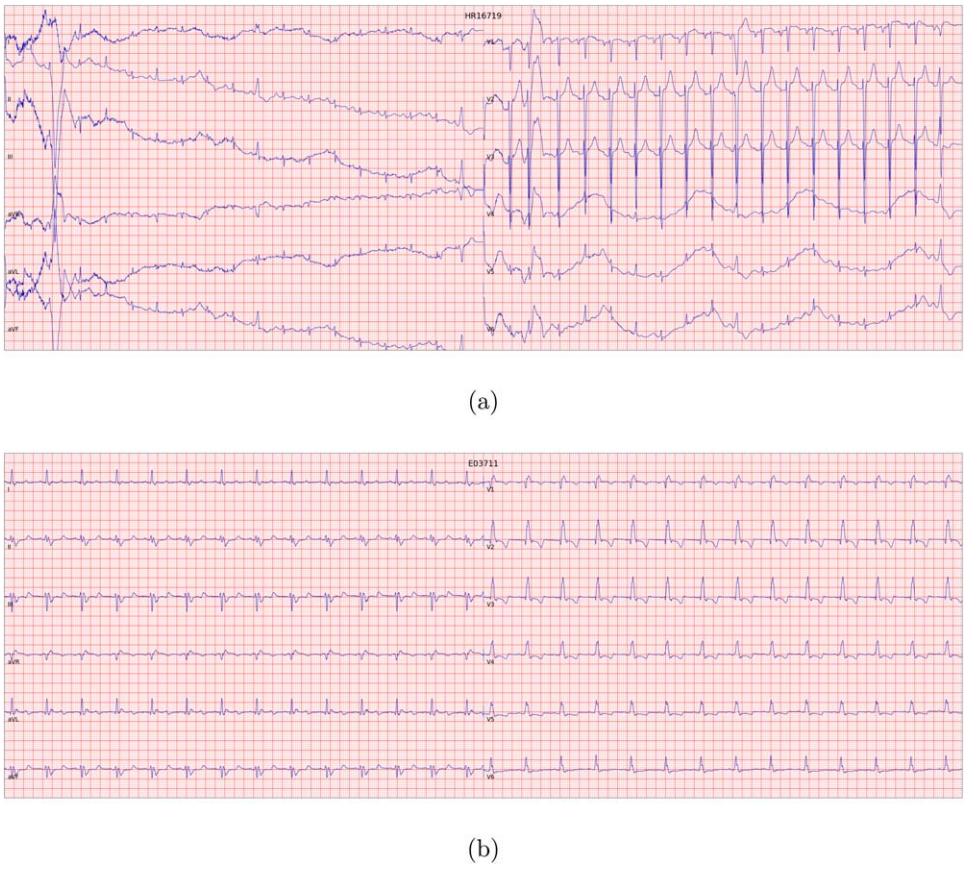


Figure 5. Examples of ECGs misclassified by the SE-Resnet model. (a) shows a false negative example in which the training label included LAD, but model prediction did not include LAD. In this case, movement artefact and heavy baseline wander on multiple leads are likely to have led to the misclassification. (b) shows a false positive example in which the training label did not include LAD, but model predicted LAD. We note that the QRS complex in lead II is not prominent, and are thus challenging for humans to label accurately.

Table 3. Comparison of clinican 1 (HG) versus clinician 2 (SW) for classification of 50 FP and 50 FN left axis deviation (LAD) examples. $\overline{\text{LAD}}$ represents a classification of *no* LAD.

		Clinician 1 (HG)		
		LAD	$\overline{\text{LAD}}$	Unsure
Clinician 2 (SW)	LAD	7	38	5
	$\overline{\text{LAD}}$	9	33	8
	Unsure	9	33	8

Table 4. Comparison of clinican experts HG, SW classification with training set labels for left axis Deviation (LAD). $\overline{\text{LAD}}$ represents a classification or label of *no* LAD.

		Clinician 1 (HG)			Clinician 2 (SW)		
		LAD	$\overline{\text{LAD}}$	Unsure	LAD	$\overline{\text{LAD}}$	Unsure
Model	LAD	7	38	5	8	38	4
	$\overline{\text{LAD}}$	9	33	8	15	30	5

4. Discussion

This paper proposes the use of an modified Resnet for multiclass ECG classification. The main modification is the addition of an SE layer, which allows better modelling of channel interdependencies.

The model performed well, placing 2nd out of 41 teams in the 2020 Physionet Challenge. However, like many other deep neural network architectures trained on the Physionet Challenge 2020 data, our model did not

generalise well to the hidden test data. This may be due to overfitting, and we noted that 2020's winning entry incorporated hand-crafted features that would increase model robustness.

By analysing a small convenience sample of training data examples of one cardiac condition, LAD, we observed instances that appeared to be mislabelled. To estimate the extent of possible mislabelling, we extracted a larger sample of training data that have been misclassified by our model. Two clinicians independently assessed the sample. We found that the clinician's had very high agreement with each other ($\kappa = 0.92$), but very poor agreement with the training labels (HG: $\kappa = -0.057$, SW: $\kappa = -0.159$). The implication of this for machine learning models is that there is an upper-bound on their accuracy which may limit the possibility of creating a generalisable model that performs at human expert level in all settings.

The two problems of model overfitting and unreliable training data might both be caused by the same underlying issue. Even though the Physionet ECG data has been sourced from many locations, the differences in size of the datasets means that any resulting deep learning model will tend to be dominated by the largest data sets—leading to overfitting. Similarly, differences in the reliability of training labels may not be random, but dataset-specific. Indeed, for our example of LAD, we are aware that the presence of QRS-complex inversion in Lead II is a requirement for diagnosis of LAD in UK text books, but not in some Chinese texts (Hampton 1997, Chen *et al* 2013).

One approach to dealing with this problem may be create models that account for differences between datasets. More generally, this approach is known as domain adaptation. One specific way to implement this, initially proposed by Alvi *et al* (2018), is so-called 'Joint Learning and Unlearning'. This uses an adversarial multi-task approach to simultaneously minimise domain (i.e. data set) prediction accuracy and maximise task accuracy. This approach has been successfully used for MRI segmentation problems, and we have begun investigating how it may be used for ECG data (Shang *et al* 2021).

In conclusion, we have trained an SE-net that automatically identifies 24 cardiac abnormalities from 12-lead ECG. In validation on external data, the model performed much worse than in internal cross-validation. Further analysis of the model outcomes suggests that this may be partly due to difference in labelling procedures between different training data sets.

ORCID iDs

D Murphy https://orcid.org/0000-0001-9662-3840
H Gifford https://orcid.org/0000-0003-2797-852X
S D Relton https://orcid.org/0000-0003-0634-4587
D C Wong https://orcid.org/0000-0001-8117-9193

References

Alday E A P *et al* 2020 Classification of 12-lead ecgs: the physionet/computing in cardiology challenge 2020 *Physiol. Meas.* **41** 124003
Alvi M, Zisserman A and Nellåker C 2018 Turning a blind eye: explicit removal of biases and variation from deep neural network embeddings *Proc. European Conf. on Computer Vision (ECCV) Workshops* (https://doi.org/10.1007/978-3-030-11009-3_34)
Anh D, Krishnan S and Bogun F 2006 Accuracy of electrocardiogram interpretation by cardiologists in the setting of incorrect computer analysis *J. Electrocardiol.* **39** 343–5
Chen W B, Pan X L, Wan X H, Lu X F, Liu C, Hu S J, Kang X X and Yang J 2013 *Diagnosis* 8th edn (People's Republic of China: People's Medical Publishing House)
Cook D A, Oh S-Y and Pusic M V 2020 Accuracy of physicians' electrocardiogram interpretations a systematic review and meta-analysis *JAMA Intern. Med.* **180** 1461–71
Eschalier R, Ploux S, Ritter P, Haïssaguerre M, Ellenbogen K A and Bordachar P 2015 Nonspecific intraventricular conduction delay: definitions, prognosis, and implications for cardiac resynchronization therapy *Heart Rhythm* **12** 1071–9
Estes N M III 2013 Computerized interpretation of ecgs: supplement not a substitute *Circ-Arrhythmia. Elec.* **6** 2–4
Feinberg W M, Blackshear J L, Laupacis A, Kronmal R and Hart R G 1995 Prevalence, age distribution, and gender of patients with atrial fibrillation: analysis and implications *Arch. Intern. Med.* **155** 469–73
Hampton J R 1997 *The ECG Made Easy* 8th edn (London: Churchill Livingstone)
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, Zhang X, Ren S and Sun J 2016 Deep residual learning for image recognition *Proc. IEEE Conference on Computer Vision and Pattern Recognition* pp 770–8
Hu J, Shen L and Sun G 2018 Squeeze-and-excitation networks *Proc. IEEE Conference on Computer Vision and Pattern Recognition* pp 7132–41
Kang B, Xie S, Rohrbach M, Yan Z, Gordo A, Feng J and Kalantidis Y 2020 Decoupling representation and classifier for long-tailed recognition *Int. Conf. on Learning Representations*
Natarajan A, Chang Y, Mariani S, Rahman A, Boverman G, Vij S and Rubin J 2020 A wide and deep transformer neural network for 12-lead ecg classification *2020 Computing in Cardiology (CinC)* 47 (Piscataway, NJ: IEEE)
Reyna M A *et al* 2021 Will two do? varying dimensions in electrocardiography: the physionet/computing in cardiology challenge 2021 *2021 Computing in Cardiology (CinC)* 48 (Piscataway, NJ: IEEE)
Ribeiro A H *et al* 2020 Automatic diagnosis of the 12-lead ecg using a deep neural network *Nat. Commun.* **11** 1–9

Salerno S M, Alguire P C and Waxman H S 2003 Competency in interpretation of 12-lead electrocardiograms: a summary and appraisal of published evidence *Ann. Intern. Med.* **138** 751–60

Schläpfer J and Wellens H J 2017 Computer-interpreted electrocardiograms: benefits and limitations *J. Am. Coll. Cardiol.* **70** 1183–92

Shang Z, Zhao Z, Fang H, Relton S, Murphy D, Hancox Z, Yan R and Wong D 2021 Deep discriminative domain generalization with adversarial feature learning for classifying ecg signals *2021 Computing in Cardiology (CinC)* 48 (Piscataway, NJ: IEEE)

Willems J L *et al* 1991 The diagnostic performance of computer programs for the interpretation of electrocardiograms *New Engl. J. Med.* **325** 1767–73

Zhao Z, Fang H, Relton S D, Yan R, Liu Y, Li Z, Qin J and Wong D C 2020 Adaptive lead weighted resnet trained with different duration signals for classifying 12-lead ecgs *2020 Computing in Cardiology (CinC)* (Piscataway, NJ: IEEE)

Zhu Z *et al* 2021 Identification of 27 abnormalities from multi-lead ecg signals: an ensembled se_resnet framework with sign loss function *Physiol. Meas.* **42** 065008