Arrhythmia classification of 12-lead and reduced-lead electrocardiograms via recurrent networks, scattering, and phase harmonic correlation

Philip Warrick, Vincent Lostanlen, Michael Eickenberg, Masun Nabhan Homsi, Adrián Campoy Rodriguez, Joakim Andén

Arrhythmia Classification of 12-Lead and Reduced-Lead Electrocardiograms via Recurrent Networks, Scattering, and Phase Harmonic Correlation

May 3, 2022 **Philip A. Warrick, Vincent Lostanlen, Michael Eickenberg, Masun Nabhan Homsi, Adrián Campoy Rodríguez, Joakim Andén**

Abstract. We describe an automatic classifier of arrhythmias based on 12-lead and reduced-lead electrocardiograms. Our classifier comprises four modules: scattering transform (ST), phase harmonic correlation (PHC), depthwise separable convolutions (DSC), and a long short-term memory (LSTM) network. It is trained on PhysioNet/Computing in Cardiology Challenge 2021 data. The ST captures short-term temporal ECG modulations while the PHC characterizes the phase dependence of coherent ECG components. Both reduce the sampling rate to a few samples per typical heart beat. We pass the output of the ST and PHC to a depthwise-separable convolution layer (DSC) which combines lead responses separately for each ST or PHC coefficient and then combines resulting values across all coefficients. At a deeper level, two LSTM layers integrate local variations of the input over long time scales. We train in an end-to-end fashion as a multilabel classification problem with a normal and 25 arrhythmia classes. Lastly, we use canonical correlation analysis (CCA) for transfer learning from 12-lead ST and PHC representations to reduced-lead ones. After local cross-validation on the public data from the challenge, our team "BitScattered" achieved the following results: 0.682 ± 0.0095 for 12-lead; 0.666 ± 0.0257 for 6-lead; 0.674 ± 0.0185 for 4-lead; 0.661 ± 0.0098 for 3-lead; and 0.662 ± 0.0151 for 2-lead.

Keywords: electrocardiography, scattering transform, phase harmonic correlation, canonical correlation analysis, convolutional neural networks, long short-term memory networks
Submitted to: *Physiol. Meas.*

1. Introduction

The World Health Organization estimates that cardiovascular diseases (CVDs) caused 17.9 million deaths worldwide in 2016, and may reach 23.6 million in the year 2030. In this context, electrocardiography (ECG) plays a vital role in CVD prevention, diagnosis, and treatment. This is because each electrode in an ECG can reveal cardiac abnormalities, which are risk factors for CVDs. The main advantage of ECG is that its acquisition is inexpensive and non-invasive. However, the visual interpretation of ECG is tedious, time-consuming, and requires expert knowledge. To address this, the PhysioNet/Computing in Cardiology Challenge 2021 offers a benchmark for automatic classification of cardiac abnormalities from 12-lead and reduced-lead ECGs.

Prior literature on ECG classification exhibits a methodological divide: signal processing versus machine learning. On one hand, digital signal processing methods include low-pass filters, fast Fourier Transform, and wavelet transform. On the other hand, machine learning methods include approaches such as random forests, support vector machines, convolutional

neural networks (CNNs) and long short-term memory (LSTM) networks. While feature engineering lacks flexibility to represent fine-grain class boundaries, a purely learned pipeline may lead to uninterpretable overfitting.

The 40 ranked contributors to the PhysioNet/CinC Challenge 2021 emphasize these methodological aspects in varying degrees, and well represent the state of the art in ECG classification. There have been numerous other formulations of the automated ECG interpretation problem, such as single-lead detection of atrial fibrillation (Clifford et al. 2017). Yet the uniqueness of the Physionet/CinC Challenge 2021 is the range of 29 arrhythmias considered for multi-label classification, and the availability of a large, full 12-lead ECG database with approximately 88,000 public records. Therefore we mainly restrict the scope of our methodological comparison to approaches used by other participants of the Challenge. The scores we report below refer to the all-lead Challenge metric on the hidden test set.

Team ISIBrno-AIMT Nejedly et al. (2021) (with winning score 0.58) developed a residual CNN network with an attention mechanism and a mixture of loss functions, including the differentiable Challenge metric approximation developed by Vicar et al. (2020). Preprocessing included a third-order Butterworth bandpass filter ($1\,\mathrm{Hz}$–$47\,\mathrm{Hz}$), z-score normalization and $16.4\,\mathrm{s}$ batch time span. They created a common model for all lead configurations and randomized lead selection during training, zeroing unused leads. An evolutionary optimization estimated probability thresholds for each class. The final step was majority voting on an ensemble of three such models.

Team CeZIS Bugata et al. (2021) (score 0.52) proposed a two-phase method. In the first phase, a 1-D variant of the ResNet50 was trained using data from different sources to first extract quality latent features. To address presumed label inconsistency, they relaxed the label semantics to include "unknown" in addition to positive and negative. In the second phase, the trained model was tuned using a loss function that approximated the Challenge metric. Separate models were trained for specific databases (CPSC2018, Georgia or "other") and a trained predictor of database was used at inference time to choose the most appropriate model.

The approach proposed by team snu_adsl Jangwon et al. (2021) (score 0.55) used an EfficientNet-B3 neural network as a base classifier with threshold optimization and label masking using auxiliary data sources. Team ami_kagoshima Hiroshi et al. (2021) (score 0.49) presented a CNN based EfficientNet model that incorporated DivideMix and stochastic weight averaging (SWA) to address label inconsistency. The network architecture introduced by team SMS+1 Gallego Vázquez et al. (2021) (score 0.52) combined hand-crafted features (demographic, morphological, and heart-rate-variability metrics) with ECG features extracted via CNNs. The network was trained with Asymmetric Loss (ASL) for multi-label classification to address class imbalance, along with a self-learning label correction method to further mitigate noisy labels in the dataset.

Transfer learning exploits knowledge from past tasks to better learn how to perform a new task. Its goal is to reuse previous learning to learn novel (but related) tasks more efficiently or solve new problems Yang et al. (2013). In this paper, we adopted this paradigm with the use of canonical correlation analysis (CCA), which maximizes the correlation between two sets of multidimensional signals. CCA has been applied elsewhere in automatic signal

classification. Noh & De Sa (2013) used CCA to discriminate electroencephalogram (EEG) patterns, while Lin et al. (2018) used it to mitigate the impact of electromyography electrode shift on classification accuracy. Fan et al. (2016) used CCA for EEG feature extraction and realtime artifact detection. Kuzilek et al. (2014) showed that CCA could be used to estimate unstructured ECG noise although it was sensitive to (structured) $50\,\mathrm{Hz}$ power line interference.

Our contribution to this research area aims to overcome the methodological divide by combining insights from signal processing and machine learning. At a first stage, we extract time scattering transform (ST) and phase-harmonic correlation (PHC) coefficients for each ECG channel. Although this stage is not trainable, ST offers time-frequency analysis with numerical guarantees of stability to time warps. The PHC describes the signal with numerically stable estimates localized in time, frequency and phase that characterize the phase dependence of coherent ECG components Mallat et al. (2019). At a second stage, we train a depthwise separable convolution (DSC) network, followed by a bidirectional long short-term memory (BiLSTM) network. While DSC combines scattering and phase coefficients from multiple leads simultaneously, the BiLSTM can also capture longer-term trends in cardiac activity. We also investigated transfer learning to the reduced-lead models using canonical correlation analysis (CCA). This novel use of CCA transfers 12-lead information in the learning phase of the reduced-lead models. The work is inspired from our previous publications, which aimed at detecting sleep arousals from polysomnographic recordings Warrick et al. (2019). Our system extends previous Challenge work Warrick et al. (2020, 2021) with more robust preprocessing using all datasets; consideration of recordings with arbitrary duration and sampling rate; inclusion of PHC, to our knowledge the first such application to biomedical signals; and the use of an asymmetrical loss function (ASL) that is more appropriate to imbalanced multi-label data.

2. Methods

This section describes the datasets used to construct the ECG classifier and explains the role of each of the system components in isolation. Figure 1 summarizes our proposed system.

2.1. Data

The PhysioNet/CinC Challenge 2021 data, described in detail in (Reyna et al. 2021), includes approximately 88,000 public and 26,000 private ECG records. Each record is assigned one or more arrhythmia diagnoses (or normal sinus rhythm) by experts and a subset of thirty of these were considered for the Challenge scoring. Four pairs of these diagnoses were merged, giving a total of twenty-six diagnosis labels. A complete list of the scored diagnoses is shown in Table 1. The ECG records vary in their duration and sampling rates; their relative contributions to the public and hidden datasets; and their distribution of arrhythmia diagnoses. Table 2 shows some key characteristics of the seven data sources.

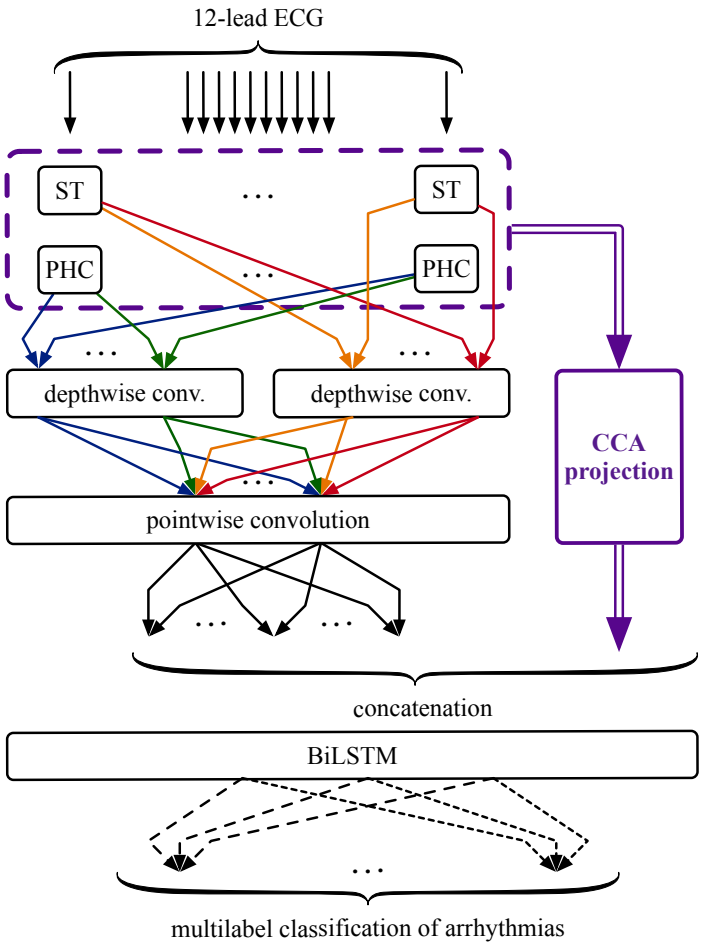

Figure 1. System overview for k-lead ECG. Top: channel-wise scattering transform (ST) and phase harmonic correlations (PHC). Arrow colors denote ST paths and PHC coefficients (shown for two channels only). Middle left: depthwise and pointwise convolution layers of the depthwise separable convolutional (DSC) neural network. For brevity, only two ST paths, two PHC coefficients and two feature maps are shown as pointwise convolution inputs. Middle (optionally, in violet): CCA projection of k-lead ST and PHC uses coefficients precalculated from 12-lead data before training. Bottom: bidirectional long short-term memory network (BiLSTM) followed by classification. Arrow styles denote output units. Only three hidden units and two arrhythmia classes are shown.

2.2. Preprocessing

In a preprocessing step in this study, we resampled the data to a common sampling rate of $f_s = 500\,\mathrm{Hz}$ and split the recordings into multiple segments when required to support arbitrary acquisition duration. Thus we were able to include subsets that we had omitted in our PhysioNet/CinC Challenge 2021 submission due to the presence of recordings with long durations or non-uniform acquisition rates f_{ac} (in the St. Petersburg and PTB data). This also enabled us to process the non-uniform f_{ac} of the University of Michigan and Undisclosed data, which likely failed in our final official-phase Challenge submission because our processing assumed a fixed f_{ac} of $500\,\mathrm{Hz}$.

We resampled the data using the Python function **scipy.signal.resample**. Although some ECG recording durations in the training set were as long as $30\,\mathrm{min}$, the vast majority (78,181 of

Diagnosis	Abbreviation	Merged
atrial fibrillation	AF	
atrial flutter	AFL	
bundle branch block	BBB	
bradycardia	Brady	
complete left bundle branch block	CLBBB	LBBB
complete right bundle branch block	CRBBB	RBBB
first-degree AV block	IAVB	
incomplete right bundle branch block	IRBBB	
left axis deviation	LAD	
left anterior fascicular block	LAnFB	
left bundle branch block	LBBB	CLBBB
low QRS voltages	LQRSV	
nonspecific intraventricular conduction disorder	NSIVCB	
normal sinus rhythm	NSR	
premature atrial contraction	PAC	SVPB
pacing rhythm	PR	
poor R-wave progression	PRWP	
premature ventricular contractions	PVC	VPB
prolonged PR interval	LPR	
prolonged QT interval	LQT	
Q wave abnormal	QAb	
right axis deviation	RAD	
right bundle branch block	RBBB	CRBBB
sinus arrhythmia	SA	
sinus bradycardia	SB	
sinus tachycardia	STach	
supraventricular premature beats	SVPB	PAC
T-wave abnormal	TAb	
T-wave inversion	TInv	
ventricular premature beats	VPB	PVC

Table 1. PhysioNet/CinC Challenge 2021 scored arrhythmias, with indicated abbreviations and label merging.

87,663≈89%) were $10\,\mathrm{s}$ or less. Therefore to reduce computational requirements, we reduced the time span of the learning batches to $10\,\mathrm{s}$. Longer recordings were truncated at $10\,\mathrm{s}$ or split into multiple sub-sequences of $10\,\mathrm{s}$. We also split the final two sub-sequences into two equal lengths when necessary to avoid very short sub-sequences less than one-third of the $10\,\mathrm{s}$ batch time span. For all sub-sequences $< 10\,\mathrm{s}$ we padded the input by reflection to reduce filtering artifacts; we excluded output samples corresponding to these padded inputs from contributing to the training loss function. We excluded 24 Georgia and 388 Ningbo records from training due to the presence of invalid (NaN) values in the ECG recordings.

	Source	Train	Valid	Test	Dur	f_{ac} (Hz)
1.	China Physiological Signal Challenge 2018	10,330	1,463	1,463	6-144 s	500
2.	Chapman University, Shaoxing and Ningbo hospitals (China)	45,152	-	-	10	500
3.	St. Petersburg Institute of Cardiological Technics (Russia)	74	-	-	30 min	257
4.	Physikalisch-Technische Bundesanstalt (PTB, Germany)	22,353	-	-	10-120 s	500 or 1000
5.	Georgia (United States)	10,344	5,167	5,161	5-10 s	500
6.	Undisclosed (United States)	-	-	10,000	?	?
7.	University of Michigan (United States)	-	-	19,642	10 s	250 or 500

Table 2. Characteristics of the PhysioNet/CinC Challenge 2021 data, showing number of records for training, hidden validation and hidden test sets, acquisition durations ("Dur") and sampling rates f_{ac}. Dashes indicate no records and question marks indicate that the values were not provided.

2.3. Scattering transform

The scattering transform is a deep convolutional network whose filters are defined a priori instead of being learned from data. The earliest application of the scattering transform to cardiology is due to Chudáček et al. (2014), in the context of fetal heart rate classification. We refer to Mallat (2016) for a mathematical introduction and to Warrick et al. (2019) for a recent review of the state of the art. Specifically, every layer in the scattering network contains filters of the form

$$\boldsymbol{\psi}_j : t \mapsto 2^{-j/Q}\boldsymbol{\psi}(2^{-j/Q}t), \tag{1}$$

where $\boldsymbol{\psi}$ is a wavelet, Q is a constant number of filters per octave, and the scale variable j is an integer ranging between 0 and J. Hereafter, we take the "mother wavelet" $\boldsymbol{\psi}$ to be a Morlet wavelet with a quality factor of $Q = 1$, center frequency of $\xi = 186.4\,\mathrm{Hz}$ and frequential width $\sigma = 19.4\,\mathrm{Hz}$. This choice of ξ and σ ensures a frequency support below the Nyquist frequency for $f_s = 500\,\mathrm{Hz}$. The Morlet wavelet is a complex-valued function with a Gaussian envelope while being approximately analytic, i.e., with negligible Fourier coefficients outside of the half-line of positive frequencies ($\omega > 0$) . Furthermore, we set the maximum wavelet scale to $J = 11$ after a process of trial and error.

Let $\boldsymbol{\phi}_T$ be a Gaussian filter of cutoff frequency equal to $1/T$. The first two orders of the scattering transform are

$$\begin{aligned} \mathbf{S}_1\boldsymbol{x}(t, j_1) &= |\boldsymbol{x} * \boldsymbol{\psi}_{j_1}| * \boldsymbol{\phi}_T(t) \quad \text{and} \\ \mathbf{S}_2\boldsymbol{x}(t, j_1, j_2) &= \Big||\boldsymbol{x} * \boldsymbol{\psi}_{j_1}| * \boldsymbol{\psi}_{j_2}\Big| * \boldsymbol{\phi}_T(t), \end{aligned} \tag{2}$$

where the vertical bars and the asterisk denote complex modulus and convolution product respectively.

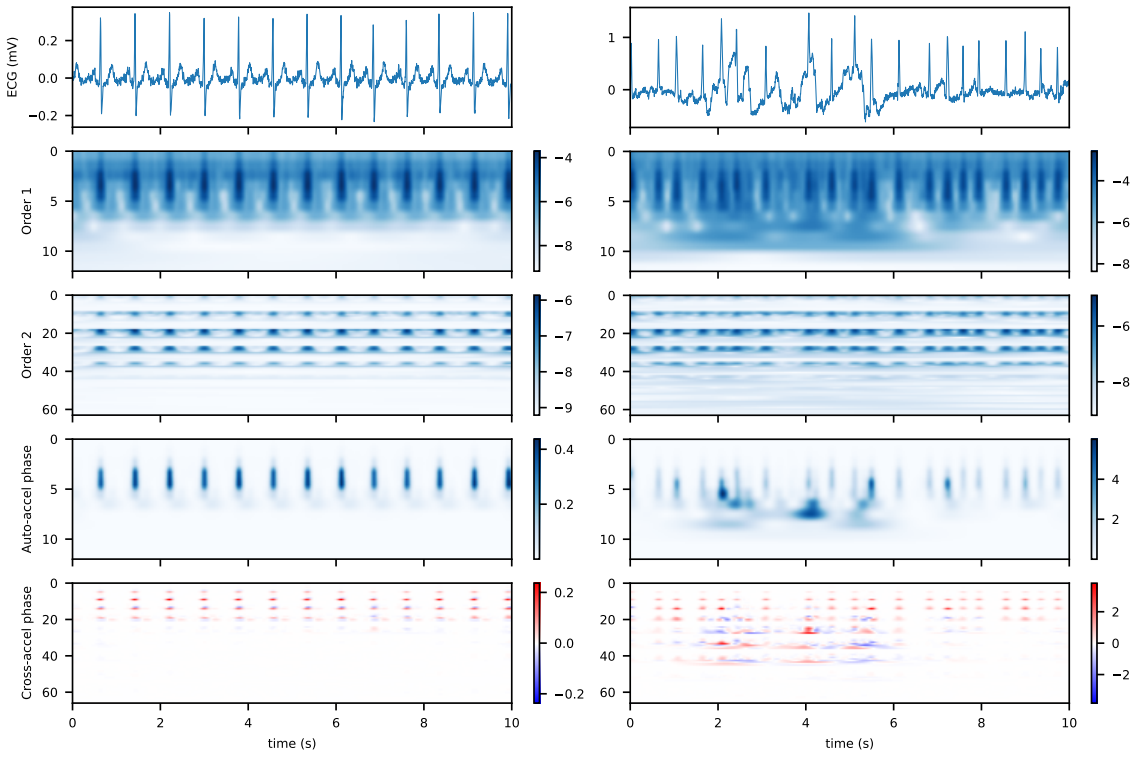

Figure 2. Scattering transform and phase harmonic correlation of ECG lead I 10 s recordings for (left) normal sinus rhythm (record A0002) and (right) atrial fibrillation (record A0023). Top to bottom: input ECG, 12 first-order and 63 second-order ST paths (log scale), 12 auto-correlation and 63 cross-correlation phase harmonic coefficients (linear scale).

For every discretized value of time t, we concatenate first-order coefficients $\mathbf{S}_1\boldsymbol{x}(t, j_1)$ and second-order coefficients $\mathbf{S}_1\boldsymbol{x}(t, j_1, j_2)$ to produce a multidimensional time series $\mathbf{S}\boldsymbol{x}(t, p)$; where the multiindex p, known as scattering *path*, either denotes an singleton (j_1) or a pair (j_1, j_2). With $J = 11$, this results in 12 first-order and 63 second-order paths for a total number of $P = 75$ paths.

To control the degree of time invariance, we modified the Python scattering package Kymatio‡ to set the time scale of Gaussian averaging to $T = 62.5\,\mathrm{ms}$. Note that this T is less than the customary $2^J/\xi$. Rather, the filterbank $\{\psi_j\}_j$ covers the frequency range $[2^{-J}\xi; \xi] = [0.091\,\mathrm{Hz}; 186\,\mathrm{Hz}]$ whereas the scattering transform is discretized at a Nyquist rate of $2/T = 32\,\mathrm{Hz}$. This rate is chosen to be higher than typical patient heart rates yet considerably lower than f_s.

We apply a pointwise compressive nonlinearity to the output of the ST, namely, an offset log function: $x \mapsto \log(x + \varepsilon)$ where $\varepsilon = 10^{-4}$. We "standardize" the result by subtracting its empirical mean and dividing it by its empirical standard deviation. Previous literature has shown that such transformations can bring the empirical histogram of scattering transform magnitudes closer to a standard normal distribution and thus improve classification accuracy Lostanlen et al. (2018).

‡ Official website of Kymatio: https://www.kymat.io

2.4. Phase harmonic correlation

Originally introduced in Mallat et al. (2019) to theoretically analyze the behavior of linear filtering followed by rectification in convolutional neural networks, wavelet phase harmonic descriptors have since led to impressive performance in applied fields, such as cosmological parameter estimation Allys et al. (2020). Here we use the explicit definition of wavelet phase harmonic descriptors from Allys et al. (2020), and adapt them to compute temporally local phase correlations (instead of fully time-invariant descriptors). The phase harmonics of a complex number $z \in \mathbb{C}$ is a sequence defined for all $k \in \mathbb{Z}$ by

$$[z]^k = |z|e^{jk\varphi(z)}, \tag{3}$$

where $\varphi(z)$ is the complex phase of z. Note that in general, $[z]^k \neq z^k$, because phase harmonics preserve the magnitude, whereas the simple power does not. We see that the phase derivative of the phase harmonics increases by a factor of k. The phase derivative is also referred to as the analytic instantaneous frequency; scaling frequency without affecting the time distribution is called frequency transposition, analogous to the same musical term.

The outputs of the Morlet filters across scales j_i for $i \in \{1, \ldots, J\}$ are

$$\boldsymbol{\Psi}_{j_1}(\boldsymbol{x}) = \boldsymbol{x} * \boldsymbol{\psi}_{j_1}. \tag{4}$$

In Mallat et al. (2019) it is shown that $\Psi_{j_1}(\boldsymbol{x})$ is not correlated to $\Psi_{j_2}(\boldsymbol{x})$ unless the frequency supports of Fourier transforms $\hat{\Psi}_{j_1}(\boldsymbol{\omega})$ and $\hat{\Psi}_{j_2}(\boldsymbol{\omega})$ overlap. Furthermore, since the correlation of $w \in \mathbb{C}$ and $z \in \mathbb{C}$ is

$$\begin{aligned}
\mathbf{C} &= \mathrm{Re}(wz^*) \\
&= \mathrm{Re}(|w|e^{j\varphi(w)}|z|e^{-j\varphi(z)}) \\
&= |w||z|\mathrm{Re}(e^{j[\varphi(w)-\varphi(z)]}) \\
&= |w||z|\cos[\varphi(w) - \varphi(z)].
\end{aligned} \tag{5}$$

$\mathbf{C}$ is maximized for given $|w|$ and $|z|$ when $\varphi(w) \approx \varphi(z)$. Therefore, for each filter output pair $(\Psi_{j_1}(\boldsymbol{x}), \Psi_{j_2}(\boldsymbol{x}))$ we need to ensure that their phases move at roughly the same velocity to stay in alignment. Since the wavelet instantaneous frequencies differ by a factor related to the difference in scales, we correlate phase harmonics of $\Psi_{j_1}(\boldsymbol{x})$ with $\Psi_{j_2}(\boldsymbol{x})$ using transposition factor $k' = x_{j_2}/x_{j_1} \in \mathbb{R}$, where x_{j_1} and x_{j_2} are the center frequencies of corresponding filters $\boldsymbol{\psi}_{j_1}$ and $\boldsymbol{\psi}_{j_2}$ as

$$\mathbf{Cx}(t, j_1, j_2) = \mathrm{Re}\left\{[\Psi_{j_1}(\boldsymbol{x})]^{x_{j_2}/x_{j_1}} \cdot [\Psi^*_{j_2}(\boldsymbol{x})]\right\}. \tag{6}$$

As a final step, we applied the Gaussian averaging filter $\boldsymbol{\phi}_T(t)$ and discretized at a Nyquist rate of $2/T$ = 32 Hz:

$$\mathbf{Cx}_T(t, j_1, j_2) = \mathbf{Cx}(t, j_1, j_2) * \boldsymbol{\phi}_T(t). \tag{7}$$

This local averaging provides robustness while preserving the temporal variations of the phase harmonics. For $j_1 = j_2$ we refer to the resulting phase harmonic correlation coefficients $\mathbf{C}\mathbf{x}_T$ as autocorrelations; for $j_1 \neq j_2$ they are cross correlations.

We modified the Kymatio scattering package to support the phase harmonic correlation calculations for one-dimensional time-series. We computed these correlations independently per ECG lead. Figure 2 illustrates the first two orders of the ST and corresponding PHC for a single lead of two ECG recordings, one labelled with normal sinus rhythm and the other labelled as atrial fibrillation.

2.5. Depthwise separable convolution

A depthwise separable convolution (DSC) splits the computation into two operations: depthwise convolution X linearly combines the leads for each ST path or PHC coefficient while the pointwise convolution Y linearly combines these transformed paths, as in (8a) and (8b)

$$X\left[p\right] = \sum_{l=1}^{L} S\left[l,\ p\right] F\left[p, l\right] \tag{8a}$$

$$Y\left[n\right] = \beta\left[B\left[n\right] + \sum_{p=1}^{P} X\left[p\right] G\left[p, n\right]\right] \tag{8b}$$

where $L \in \{12, 6, 4, 3, 2\}$ is the the number of leads and P represents the number of ST paths and/or PHC coefficients. F and G refer to the filter maps, N is the number of pointwise mixes, B is the bias and β represents the activation function. The total number of convolution coefficients including the bias weights is therefore $P \times L + (P + 1) \times N$. This is often a reduction in parameters compared to regular convolution. We used a DSC layer with ReLU activation and $N \in \{66, 150\}$, chosen to be on the order of P.

2.6. Long-short term memory (LSTM)

An LSTM is a type of recurrent neural network that can model temporal sequences Hochreiter & Schmidhuber (1997). It preserves information from inputs that have already passed through it using a hidden state. Bidirectional LSTMs (BiLSTM) process data in forward and reverse directions to capture both past and future contexts with two separate hidden layers.

2.7. Transfer learning for reduced-lead models

For reduced-lead models, we apply transfer learning from the 12-lead data using canonical correlation analysis (CCA). CCA finds a pair of linear transformations for two sets of multidimensional variables (views V_i), such that the linear projections of the two views, $(V_1 w_1, V_2 w_2)$ are maximally correlated Hardoon et al. (2004). In our case, view V_i is the ST and/or PHC coefficients of lead sets: V_1 corresponds to the lead set used for prediction (2,

3, 4 or 6 leads) and V_2 corresponds to the respective complement from 12-lead data (10, 9, 8 and 6 leads). This is done by maximizing the following equation:

$$\begin{aligned}\rho &= \max_{w_1,w_2} \operatorname{corr}(V_1 w_1, V_2 w_2) \\ &= \max_{w_1,w_2} \frac{w_1^T \Sigma_{12} w_2}{\sqrt{w_1^T \Sigma_{11} w_1 w_2^T \Sigma_{22} w_2}}\end{aligned} \tag{9}$$

where Σ_{11}, Σ_{22} and Σ_{12} are the covariances and cross-covariance of V_1 and V_2; and w_1 and w_2 are determined by singular-value decomposition.

We calculate w_1 and w_2 from fold training data prior to network training. CCA uses V_1 and V_2 to find the projection vectors corresponding to the *k* highest left- and right- singular values, and $k = P \times L$ was chosen to include all the singular values.

During training and prediction, V_1 is projected with fixed w_1. This projection is intended to transfer information from (possibly unavailable) V_2, correlated with the complementary lead set, such that classification of reduced-lead ECG records is improved.

2.8. Asymmetric Loss Function

In multilabel classification problems, there are typically a few positive and many negative labels per instance. This is also the case for this ECG classification study. In the Chapman-Shaoxing-Ningbo dataset, for example, there were 45,152 recordings and a total of 68,852 (scored and unscored) labels. Thus the average number of labels per recording was 1.52, and this rate was similar for the other datasets. Such label sparsity coupled with the large range of label incidence in the training data (see Figure 3) leads to under-emphasizing the gradients from the positive labels during training. Accordingly, numerous loss functions have been proposed to treat positive and negative labels differently in the the loss calculation.

One such approach, called the asymmetrical loss function (ASL) (Ridnik et al. 2021), reduces the loss contribution of the easiest-to-classify negative instances and can potentially discard instances that are mislabeled as negative. The mathematical description of the ASL begins with the general form of a binary loss function L:

$$L = -yL_+ - (1-y)L_- \tag{10}$$

where y is the ground-truth of a label and L_+ and L_- are the loss contributions for positive and negative instances, respectively. In the typical one-hot network output architecture, the total loss is the sum of such losses over the K possible labels. The ASL extends focal loss Lin et al. (2017) which applies the focusing parameter γ to the loss contributions of negative and positive instances having network output probability p:

$$\begin{cases} L_+ = (1-p)^{\gamma_+} \log(p) \\ L_- = p^{\gamma_-} \log(1-p) \end{cases} \tag{11}$$

where $\gamma = \gamma_+ = \gamma_-$ for focal loss and $\gamma = 0$ gives the binary cross-entropy (BCE) loss. This reduces the loss contribution of negative instances presumably easiest to classify, that is, those

with low probability $p \ll 0.5$. However, it may also reduce the contribution of the rare positive instances. To address this, ASL uses asymmetric focusing that sets larger values to γ_+ than γ_-.

Finally, the label imbalance may be such that the effect of γ_- is not sufficient to reduce the influence of negative instances, including those that are mislabelled as such. For this reason, ASL introduces hard threshold $p_m = \max(p - m, 0)$ to further remove the easiest negative samples, modifying L_- in (11) to

$$L_- = (p_m)^{\gamma_-} \log(1 - p_m). \tag{12}$$

2.9. Decision Rule

Although our Challenge approach used all sub-sequences during training, inference was limited to the first sub-sequence. In this study, we removed this limitation to consider predictions over the entire recording by averaging probabilities over all sub-sequence time outputs. Our decision rule chose any predicted class that exceeded probability threshold $p = 0.5$; otherwise the maximum probability class was chosen.

2.10. Implementation

Keras with TensorFlow as backend was used for building the neural network, with a Kymatio Keras layer for the scattering and phase harmonic correlations. We used a machine with 200 GB of available system memory and a GPU with 32 GB of memory.

We used two BiLSTM layers of 100 hidden units. Performance degraded for fewer layers and did not improve with more layers, so we retained this BiLSTM architecture. The number of hidden units was chosen arbitrarily and was not tuned. For the final dense layer we compared the use of BCE and ASL losses to support multilabel classification. The batch size was 35 $10\,\mathrm{s}$ blocks, chosen to be within memory limits using a 12 GB GPU.

The 10-fold cross-validation data partitions were 90% training and 10% testing for each fold. All subsequences for a recording belonged to only one of these partitions for each fold. The validation set, 10% of training, was used for as the loss criteria for early stopping (20 epochs) after completing a minimum of 50 epochs.

2.11. Evaluation

The Challenge metric described in (Perez Alday et al. 2020, Reyna et al. 2021) was used to assess performance. Differences between experimental results were often subtle, so we chose a principled comparison of experiment pairs: we performed a 2-sided t-test of the test Challenge metrics over all 10 folds to reject the null hypothesis that the metrics were equal. We considered rejection of the null hypothesis at level $p < 0.05$ to be statistically significant.

We compared the effectiveness of the ST/PHC layers to a set of reference CNN architectures and to no filtering (see Table 3). The architectures included CNN, a single convolutional layer; FCN, a fully connected three-layer CNN network (Wang et al. 2016);

MCDCNN, a multi-channel deep CNN with convolutions applied to individual channels (Cui et al. 2016); and 1-D Resnet18, one of the first very deep yet trainable networks (He et al. 2015). All comparisons included the final BiLSTM and Dense layers. We used the default parameter settings of the public implementations listed in Table 3. We adapted the three Resnet18 block outputs to the first BiLSTM layer by maxpool operations to obtain a common rate of $f_s/32$.

3. Results

We found that normalizing the ST coefficients improved training convergence and marginally improved classification performance, but had a negative impact on results that used CCA. For this reason, we report CCA results without ST normalization. For PHC coefficients, we found that the distribution was highly skewed towards very low absolute values and normalization mappings such as offset log or the ± 1 range did not lead to training convergence. However, simply scaling the PHC coefficients by a factor of $1,000$ led to convergence and classification performance with only PHC coefficients that was on par with ST coefficients alone. We did not attempt to further tune this scaling factor.

We used the default ASL parameters $p_m = 0.05$ and $\gamma_+ = 1$ and tuned γ_-. We found that for all front-end configurations of ST and PHC coefficients, $\gamma_- = 4$ gave the best Challenge scores for the validation partition on cross-validation validation for $\gamma_- \in \{2, 3, 4, 5\}$. We used these ASL settings for our subsequent comparisons.

Table 4 shows cross-validation Challenge metric results for several 12-lead architectures. Model (2) compared to Model (1) shows the beneficial effect of ASL (0.671 ± 0.0167) compared to BCE (0.586 ± 0.0125) for a classifier with ST coefficients at the front end (statistically significant at level p=2.42E-06). Model (3) indicates that a model using only PHC coefficients performed almost as well (0.657 ± 0.0094) as Model (2) that used only ST coefficients. Models (4) and (5) combined both ST and PHC coefficients, improving the score over Model (2). However only Model (5), which increased the capacity of the DSC pointwise filter from 66 to 150, had a score increase (to 0.682 ± 0.0095) that was statistically significant (p=3.61E-02).

Table 4 also shows cross-validation results for the CNN architectures. All architectures performed better than no filtering (RawECG). Note that the RawECG training at full sampling rate f_s had long training times and high memory usage, so we truncated cross-validation after 4 folds. The best ST-PHC Model (5) performed better than all these networks except for MCDCNN (0.698 ± 0.0093 vs. 0.682 ± 0.0095).

Table 5 compares cross-validation Challenge metric results for the reduced-lead models with and without CCA using ST coefficients at the front end. These results were only slightly inferior to the 12-lead results. The use of CCA increased scores in all models, and the increases from 0.645 ± 0.0156 to 0.662 ± 0.0151, p=2.45E-02 and from 0.661 ± 0.0098 to 0.651 ± 0.0119, p=3.63E-02 for the lead 2- and 3- lead models, respectively, were statistically significant.

Figure 3 compares the per-class cross-validation test F-measure performance for Model (2) for 12-lead and 2-lead and the 2-lead model with CCA, ST-DSC$_{66}$-CCA-ASL$_{\gamma_-=4}$. Generally, the 12-lead model performed best (for 14 of 26 labels), followed by the 2-lead

	Model	Description	Implementation	Filters	#Filters	↓
(6)	RawECG	No filtering		-	-	-
(7)	CNN	Single-layer		1×7	150	MP16
(8)	FCN	Fully Connected Network	(Ismail Fawaz et al. 2019)			
				1×8	128	
				1×5	256	
				1×3	256	MP16
(9)	MCDCNN	Multi-channel Deep CNN	(Ismail Fawaz et al. 2019)			
				1×5	8/lead	MP2
				1×5	8/lead	MP2
(10)	Resnet18	Residual NN	(Goodman 2019)			
				1×7	64	Str2,MP2
				1×3	64 ×2	
				1×3	64 ×2	
				1×3	128 ×2	Str2
				1×3	128 ×2	
				1×3	256 ×2	Str2
				1×3	256 ×2	
				1×3	512 ×2	Str2
				1×3	512 ×2	

Table 3. Reference architectures. "RawECG" indicates no filtering. Layers are described by row. ↓ indicates downsampling by maxpool (MP) or stride (STR) operations.

model with CCA, followed by the 2-lead model without CCA, which had the lowest score for 19 of 26 labels.

Figure 4 compares the per-class cross-validation test F-measure performance for the 12-lead ST model ST-DSC$_{66}$-ASL$_{\gamma_-=4}$ and the 12-lead PHC model, PHC-DSC$_{66}$-ASL$_{\gamma_-=4}$. The two model performed similarly although the ST model performed better for 18 of the 26 labels.

Accordingly, for our revised post-Challenge entry, we used the architecture of Model (5) and included CCA in the reduced-lead models. Unfortunately, this submission failed due to memory depletion on the Challenge server after just over 10 hours of processing. Therefore we were unable to report a score on the hidden test set.

4. Discussion

We observe slight performance degradation for models with decreasing numbers of leads, as observed by many other Challenge participants, suggesting that the correlation between leads is considerable. With such similar test Challenge metrics, Figure 3 indicates that the 2-lead model without CCA follows the general trend of the 12-lead model in terms of per-

	Model_1	#Leads	Cross-validation	Model_2	p-value
(1)	ST-DSC_{66}-BCE	12	0.586 ± 0.0125		
(2)	ST-DSC_{66}-$\text{ASL}_{\gamma_-=4}$	12	0.671 ± 0.0167	(1)	2.42E-06*
(3)	PHC-DSC_{66}-$\text{ASL}_{\gamma_-=4}$	12	0.657 ± 0.0094		
(4)	ST-PHC-DSC_{66}-$\text{ASL}_{\gamma_-=4}$	12	0.679 ± 0.0091		
(5)	**ST-PHC-DSC_{150}-$\text{ASL}_{\gamma_-=4}$**	12	$\mathbf{0.682 \pm 0.0095}$	(2)	3.61E-02*
(6)	RawECG-$\text{ASL}_{\gamma_-=4}$	12	0.561 ± 0.0439	(5)	4 folds only
(7)	CNN-$\text{ASL}_{\gamma_-=4}$	12	0.648 ± 0.0425	(5)	4.77E-02*
(8)	FCN-$\text{ASL}_{\gamma_-=4}$	12	0.642 ± 0.0464	(5)	2.50E-02*
(9)	MCDCNN-$\text{ASL}_{\gamma_-=4}$	12	0.698 ± 0.0093	(5)	1.42E-03*
(10)	Resnet18-$\text{ASL}_{\gamma_-=4}$	12	0.636 ± 0.0609	(5)	3.57E-02*

Table 4. Cross-validation test Challenge metric for 12-lead models. ST and PHC indicate the use of scattering and phase harmonic correlations, respectively. The DSC subscript indicates the number of pointwise filters in the depth-separable convolution filter. BCE and ASL indicates the use of the binary cross entropy and asymmetric loss functions, respectively. An asterisk (*) indicates two-sided t-test $p < 0.05$ for Model_1 compared to Model_2.

Model	#Leads	Cross-validation	p-value
ST-DSC_{66}-$\text{ASL}_{\gamma_-=4}$	6	0.653 ± 0.0145	
ST-DSC_{66}-CCA-$\text{ASL}_{\gamma_-=4}$	6	0.666 ± 0.0257	2.28E-01
ST-DSC_{66}-$\text{ASL}_{\gamma_-=4}$	4	0.662 ± 0.0125	
ST-DSC_{66}-CCA-$\text{ASL}_{\gamma_-=4}$	4	0.674 ± 0.0185	1.18E-01
ST-DSC_{66}-$\text{ASL}_{\gamma_-=4}$	3	0.651 ± 0.0119	
ST-DSC_{66}-CCA-$\text{ASL}_{\gamma_-=4}$	3	0.661 ± 0.0098	3.63E-02*
ST-DSC_{66}-$\text{ASL}_{\gamma_-=4}$	2	0.645 ± 0.0156	
ST-DSC_{66}-CCA-$\text{ASL}_{\gamma_-=4}$	2	0.662 ± 0.0151	2.45E-02*

Table 5. Effect of CCA on Challenge metric for reduced-lead models in test partition of cross-validation. An asterisk (*) indicates two-sided t-test $p < 0.05$ compared to previous row.

class performance. However for several classes in particular, the 2-lead performance was significantly lower (e.g., CRBBB|RBBB), suggesting that the increased information provided by the 12-lead configuration may be especially important for these arrhythmias.

Thus the potential for transfer learning from the 12-lead data to the reduced-lead models was limited given this similar performance across lead models. Nevertheless we observed modest and, in the case of the 2- and 3-lead models, statistically significant improvement applying CCA to the reduced-lead models. Indeed, these two models with the fewest leads have the most to gain from transferring 12-lead information. We see this in the trend in Table 5 towards statistical significance of the improved performance using CCA with decreasing numbers of leads. Figure 3 shows that for the majority of labels, the 2-lead CCA model performance was slightly inferior to the 12-lead model, but better than the 2-lead model without CCA. This also suggests that the 12-lead information transferred via CCA helped to improve the 2-lead model.

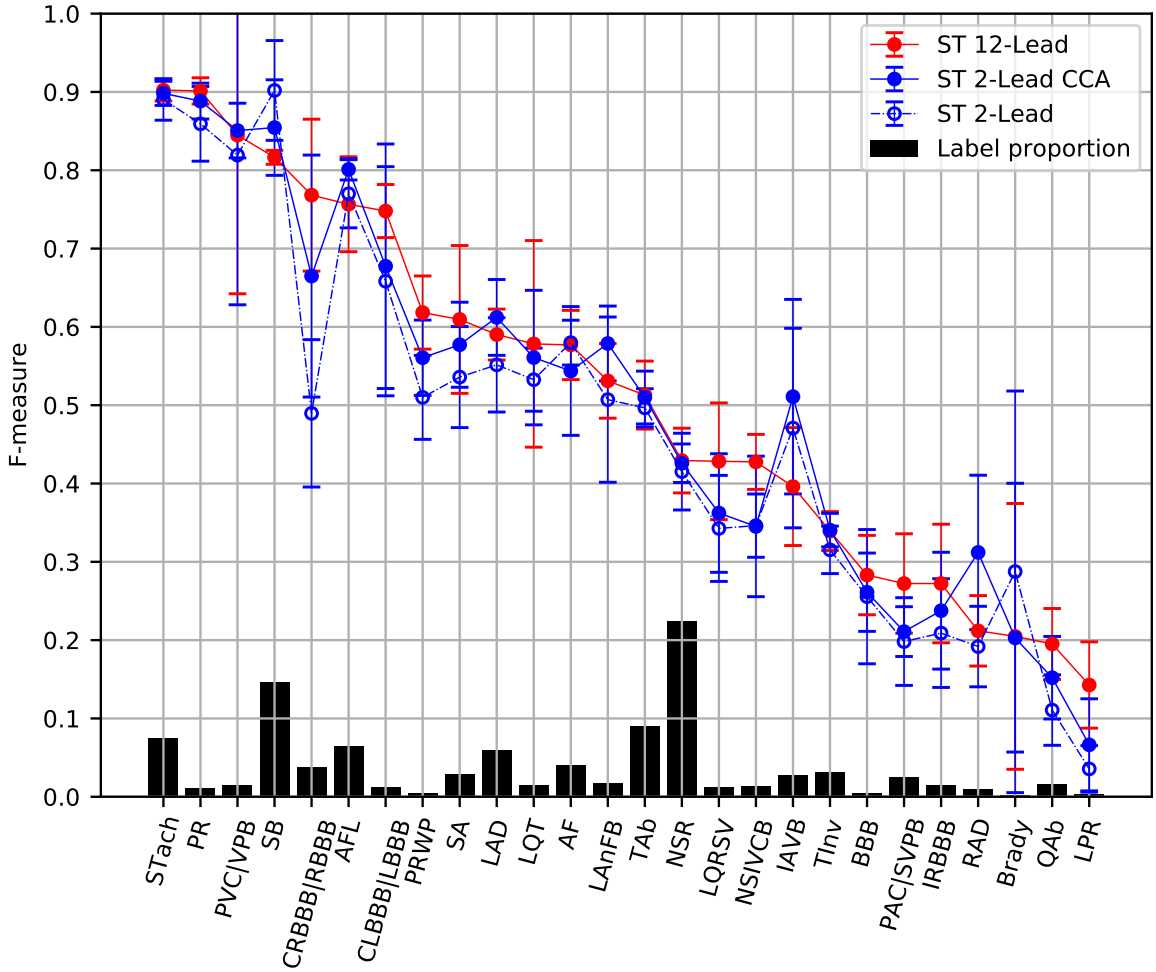

Figure 3. Per-class cross-validation test F-measure (mean$\pm$ 1 standard deviation error bars) for 12- (red solid) and 2-lead (blue dashed) models ST-DSC$_{66}$-ASL$_{\gamma_-=4}$ as well as 2-lead with CCA (blue solid) model ST-DSC$_{66}$-CCA-ASL$_{\gamma_-=4}$. The bar graph indicates arrhythmia label proportions. Label pairs separated by a bar ("|") indicate that the two labels were merged (an "OR" operation) in the official scoring and in the classifier.

ST and PHC considerably reduce the network sampling rate by a factor of $f_s/(2/T) = 16$. Because this front-end filtering uses fixed coefficients, it can be directly computed, cached and serves to ease the data training load. This is especially important for LSTMs where the dependency of computational complexity on sequence length is worse than linear, evidenced by the excessive resources required for our RawECG experiment conducted at rate f_s.

The results show that ST and PHC gave comparable results in isolation, suggesting that they have correlated information. Combining the two gave slightly improved results with statistical significance, showing some complementarity in the two measures. This is suggested in Figure 4 where the PHC classifier outperformed the ST classifier for certain labels.

(Mallat et al. 2019) points out that the first layers of convolutional neural networks (CNNs) often learn filters similar in frequency but with different phases. Furthermore, the rectified linear unit (ReLU) acts as a phase filter on the coefficients of this linear filtering when used

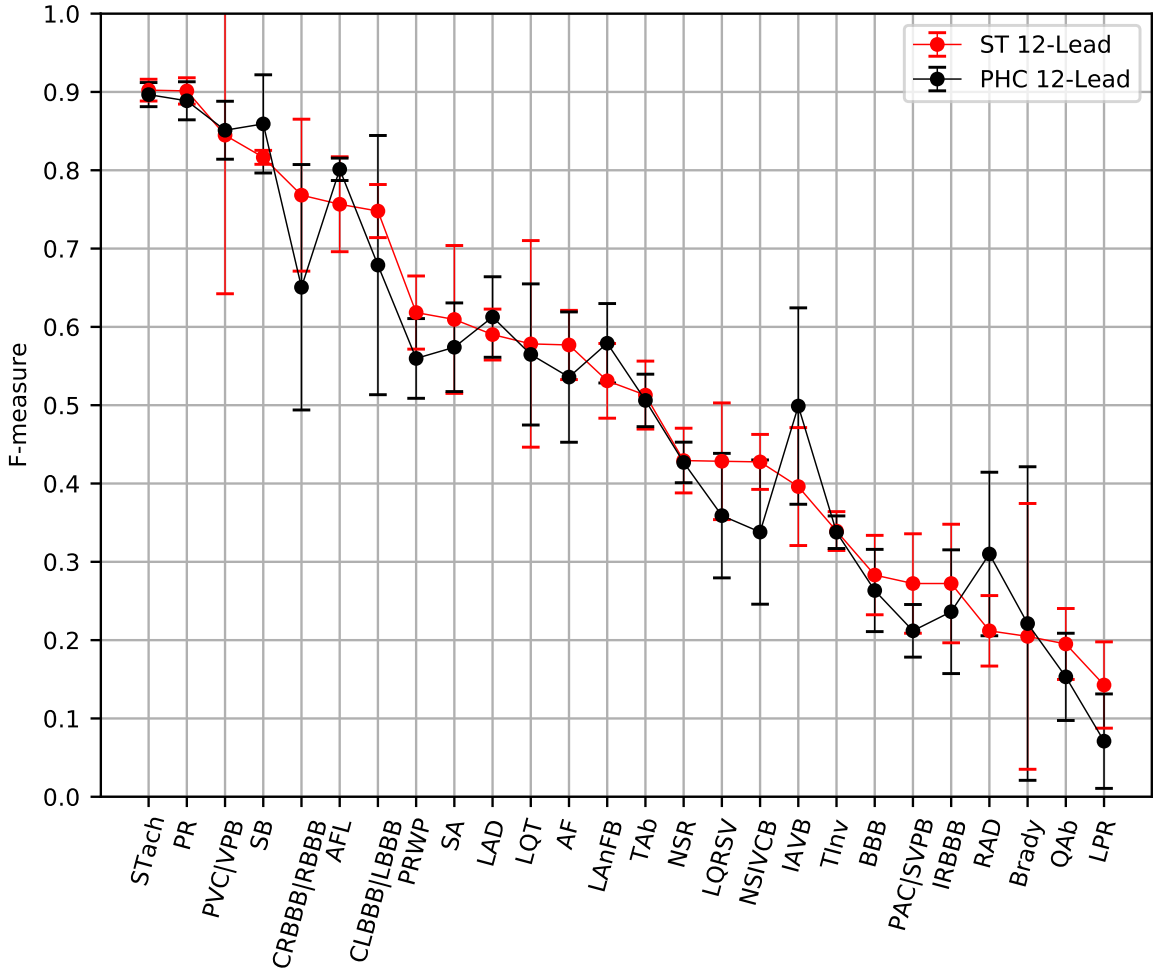

Figure 4. 12-lead per-class cross-validation test F-measure (mean± 1 standard deviation error bars) for Model (2) ST-DSC$_{66}$-ASL$_{\gamma_-=4}$ (red) and Model (3) PHC-DSC$_{66}$-ASL$_{\gamma_-=4}$ (black).

as the CNN pointwise non-linearity. This is analogous to our first network layers since PHC applies phase filtering to wavelet coefficients.

In our official-phase Challenge model (Warrick et al. 2021) which used only BCE loss, we observed that label performance was strongly dependent on label incidence, presumably because per-label loss function contribution is directly proportional to incidence. Replacing BCE with ASL significantly improved the Challenge score and Figure 3 shows a weaker association of incidence and F-measure performance. This is especially true for the highest incidence label NSR, whose performance was near the top with BCE but shifted significantly lower with the ASL, ranking 15^{th} of 26 labels. This indicates that the use of ASL helped to adapt the learning to the wide range of incidence across labels, and as noted in Section 2.8, doing so despite high sparsity in the 26 labels.

To summarize, the best model in this study improved on the official-phase Challenge entry in the following ways: 1) the use of PTB and St. Petersburg datasets during training; 2) the addition of the ASL; 3) the use of CCA for reduced lead models; and 4) the addition of PHC

coefficients. Our successful official-phase Challenge entry (Warrick et al. 2021) completed training of the baseline models in just over $18\,\mathrm{h}$ and prediction of the hidden validation set in $18\,\mathrm{min}$, within the maximum allowable times of $48\,\mathrm{h}$ and $24\,\mathrm{h}$, respectively. We would expect that the architecturally similar model of this paper would modestly increase processing time. The Challenge entry obtained an all-lead Challenge score of 0.43 on the hidden validation set, but a very low score of 0.10 on the hidden test set. As mentioned in Section 2.2, this is likely due to the fact that our official-phase Challenge submission assumed a fixed f_{ac} of $500\,\mathrm{Hz}$, but the University of Michigan and Undisclosed cohorts of the hidden test data had non-uniform f_{ac}. Including the two final datasets in training reduced cross-validation score for the 12-lead model from 0.601 ± 0.015 to 0.586 ± 0.0125. Thus our official-phase Challenge entry was roughly equivalent to BCE-based Model (1) apart from its inability to process non-uniform f_{ac}. Therefore we would expect that the model additions of this paper would generalize with improved hidden test set Challenge scores and to other cohorts of carefully curated ECG data with similar labelling. However, we were not able to verify this with our post-Challenge entry.

A limitation of the current system is that it cannot make a prediction from an ECG recording having missing samples or with entirely completely missing channels for a particular lead model. This could be addressed, for example, by within-lead interpolation, training augmentation with artificially removed data ("cropping") as in Nejedly et al. (2021), and the addition of dropout layers. Other extensions to our approach to explore include: applying PHC across channels to better capture the trajectory of the cardiac vector; using the demographic data of sex and age, recognized risk factors for cardiac pathology; improving the decision rule to better calibrate the class probabilities, and further searching of hyperparameters.

5. Conclusions

Our approach achieved experimental success without need for feature engineering, with few parameters to select. The fixed-coefficient ST/PHC layers required no learning yet generated results competitive with a range of trained CNN layers. The use of ASL was a very important contribution to address the label sparsity and imbalance of this multi-label problem. The use of CCA successfully transferred information to the reduced-lead models, especially those with the most to gain, the 2- and 3- lead models. The phase information of the PHC coefficients had similar performance to ST, and using both provided complementary information and improved the model.

References

Allys, E., Marchand, T., Cardoso, J.-F., Villaescusa-Navarro, F., Ho, S. & Mallat, S. (2020), 'New interpretable statistics for large-scale structure analysis and generation', *Phys. Rev. D* **102**, 103506.

Bugata, P., Bugata Jr., P., Kmecova, V., Stankova, M., Gajdos, D., Hudak, D., Stana, R., Horvat, S., Antoni, L., Vozarikova, G., Bruoth, E. & Szabari, A. (2021), 'A two-phase

multilabel ECG classification using one-dimensional convolutional neural network and modified labels', *Computing in Cardiology* **48**, 1–4.

Chudáček, V., Andén, J., Mallat, S., Abry, P. & Doret, M. (2014), 'Scattering transform for intrapartum fetal heart rate variability fractal analysis: A case-control study', *IEEE Transactions on Biomedical Engineering* **61**(4), 1100–1108.

Clifford, G., Liu, C., Moody, B., Lehman, L., Silva, I., Li, Q., Johnson, A. & Mark, R. (2017), AF Classification from a Short Single Lead ECG Recording: The Physionet Computing in Cardiology Challenge, *in* 'Computing in Cardiology', Vol. 44.

Cui, Z., Chen, W. & Chen, Y. (2016), 'Multi-scale convolutional neural networks for time series classification', *arXiv* **1603.06995**.

Fan, Z., Wang, Z., Li, G. & Wang, R. (2016), A canonical correlation analysis based EMG classification algorithm for eliminating electrode shift effect, *in* '2016 38th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)', IEEE, pp. 867–870.

Gallego Vázquez, C., Breuss, A., Gnarra, O., Portmann, J. & Da Poian, G. (2021), 'Two will do: Convolutional neural network with asymmetric loss, self-learning label correction, and hand-crafted features for imbalanced multi-label ECG data classification', *Computing in Cardiology* **48**, 1–4.

Goodman, A. (2019), 'keras-resnet'. (accessed May 1, 2022).
URL: *https://github.com/broadinstitute/keras-resnet*

Hardoon, D. R., Szedmak, S. & Shawe-Taylor, J. (2004), 'Canonical correlation analysis: an overview with application to learning methods', *Neural Computation* **16**(12), 2639–2664.

He, K., Zhang, X., Ren, S. & Sun, J. (2015), 'Deep residual learning for image recognition', *arXiv* **1512.03385**.

Hiroshi, S., Takashi, N., Koshiro, I., Shinji, H., Takaaki, K., Mitsutomo, Y., Shumpei, S., Toshitaka, Y. & Shimpei, O. (2021), 'Reduced-lead ECG classifier model trained with DivideMix and model ensemble', *Computing in Cardiology* **48**, 1–4.

Hochreiter, S. & Schmidhuber, J. (1997), 'Long short-term memory', *Neural Computation* **9**(8), 1735–1780.

Ismail Fawaz, H., Forestier, G., Weber, J., Idoumghar, L. & Muller, P.-A. (2019), 'Deep learning for time series classification: a review', *Data Mining and Knowledge Discovery* (4), 917–963.

Jangwon, S., Jimyeong, K., Eunjung, L., Jaeill, K., Duhun, H., Jungwon, P., Junghoon, L., Jaeseung, P., Seo-Yoon, M., Yeonsu, K., Min, K., Soonil, K., Eue-Keun, C. & Wonjong, R. (2021), 'Learning ECG representations for multi-label classification of cardiac abnormalities', *Computing in Cardiology* **48**, 1–4.

Kuzilek, J., Kremen, V. & Lhotska, L. (2014), Comparison of JADE and canonical correlation analysis for ECG de-noising, *in* '2014 36th Annual International Conference of the IEEE Engineering in Medicine and Biology Society', IEEE, pp. 3857–3860.

Lin, C.-T., Huang, C.-S., Yang, W.-Y., Singh, A. K., Chuang, C.-H. & Wang, Y.-K. (2018), 'Real-time EEG signal enhancement using canonical correlation analysis and Gaussian mixture clustering', *Journal of healthcare engineering* **2018**.

Lin, T.-Y., Goyal, P., Girshick, R. B., He, K. & Dollár, P. (2017), 'Focal loss for dense object detection', *2017 IEEE International Conference on Computer Vision (ICCV)* pp. 2999–3007.

Lostanlen, V., Lafay, G., Andén, J. & Lagrange, M. (2018), 'Relevance-based quantization of scattering features for unsupervised mining of environmental audio', *EURASIP Journal on Audio, Speech, and Music Processing* **2018**(1), 15.

Mallat, S. (2016), 'Understanding deep convolutional networks', *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* **374**(2065), 20150203.

Mallat, S., Zhang, S. & Rochette, G. (2019), 'Phase harmonic correlations and convolutional neural networks', *Information and Inference: A Journal of the IMA* **9**(3), 721–747.

Nejedly, P., Ivora, A., Smisek, R., Viscor, I., Koscova, Z., Jurak, P. & Filip, P. (2021), 'Classification of ECG using ensemble of residual CNNs with attention mechanism', *Computing in Cardiology* **48**, 1–4.

Noh, E. & De Sa, V. R. (2013), Canonical correlation approach to common spatial patterns, *in* '2013 6th International IEEE/EMBS Conference on Neural Engineering (NER)', IEEE, pp. 669–672.

Perez Alday, E. A., Gu, A., Shah, A., Robichaux, C., Wong, A.-K. I., Liu, C., Liu, F., Rad, B. A., Elola, A., Seyedi, S., Li, Q., Sharma, A., Clifford, G. D. & Reyna, M. A. (2020), 'Classification of 12-lead ECGs: the PhysioNet/Computing in Cardiology Challenge 2020', *Physiological Measurement* **47**, 1–4.

Reyna, M. A., Sadr, N., Perez Alday, E. A., Gu, A., Shah, A., Robichaux, C., Rad, B. A., Elola, A., Seyedi, S., Ansari, S., Li, Q., Sharma, A. & Clifford, G. D. (2021), 'Will two do? Varying dimensions in electrocardiography: the PhysioNet/Computing in Cardiology Challenge 2021', *Computing in Cardiology* **48**, 1–4.

Ridnik, T., Ben-Baruch, E., Zamir, N., Noy, A., Friedman, I., Protter, M. & Zelnik-Manor, L. (2021), Asymmetric loss for multi-label classification, *in* 'Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)', pp. 82–91.

Vicar, T., Hejc, J., Novotna, P., Ronzhina, M. & Janousek, O. (2020), 'Abnormalities recognition using convolutional network with global skip connections and custom loss function', *Computing in Cardiology* **47**, 1–4.

Wang, Z., Yan, W. & Oates, T. (2016), 'Time series classification from scratch with deep neural networks: A strong baseline', *arXiv* **1611.06455**.

Warrick, P. A., Lostanlen, V., Eickenberg, M., Andén, J. & Homsi, M. N. (2020), 'Arrhythmia classification of 12-lead electrocardiograms by hybrid scattering-LSTM networks', *Computing in Cardiology* **47**, 1–4.

Warrick, P. A., Lostanlen, V., Eickenberg, M., Homsi, M. N., Rodríguez, A. C. & Andén, J. (2021), 'Arrhythmia classification of reduced-lead electrocardiograms by scattering-recurrent networks', *Computing in Cardiology* **48**, 1–4.

Warrick, P. A., Lostanlen, V. & Homsi, M. N. (2019), 'Hybrid scattering-LSTM networks for automated detection of sleep arousals', *Physiological Measurement* **40**(7), 074001.

Yang, L., Hanneke, S. & Carbonell, J. (2013), 'A theory of transfer learning with applications to active learning', *Machine Learning* **90**(2), 161–189.