Transfer learning for ECG classification

Msc. Pasit Supholkhan ID 6536115

Introduction

- The paper addresses the need for reliable methods for **automatic ECG interpretation** to assist physicians in analyzing the considerable amounts of ECG data recorded by **remote monitoring devices.**
- The research explores the use of **deep** convolutional neural networks (CNN) for **classifying raw ECG** recordings, specifically focusing on the classification of **Atrial Fibrillation (AFib)**, the most common heart arrhythmia.

Introduction

- **Training CNNs** for ECG classification often **requires a large number of annotated samples**, which can be expensive to acquire. To overcome this challenge, the authors employ **transfer** learning by **pretraining** CNNs on a **large public dataset** of continuous raw ECG signals.
- The paper investigates both **supervised** and **unsupervised** **pre training** approaches, exploring their relevance and effectiveness in **reducing** the need for **expensive ECG annotations.**

Methodology

Methodology

- The paper employs deep convolutional neural networks (**CNN**) for classifying raw ECG recordings.
- Transfer learning is utilized, where CNNs are pretrained on the largest public dataset of continuous raw ECG signals, the Icentia11K dataset.
- The pre trained CNNs are then **fine tuned** on a **smaller dataset** for the classification of **Atrial Fibrillation (AFib)**, the most common heart arrhythmia.
- Both **supervised** and **unsupervised** pre training approaches are investigated, with the aim of reducing the need for expensive ECG annotations.
- The performance of the pretraining methods is **evaluated using metrics** such as **macro avg F1 score** on validation and test sets.

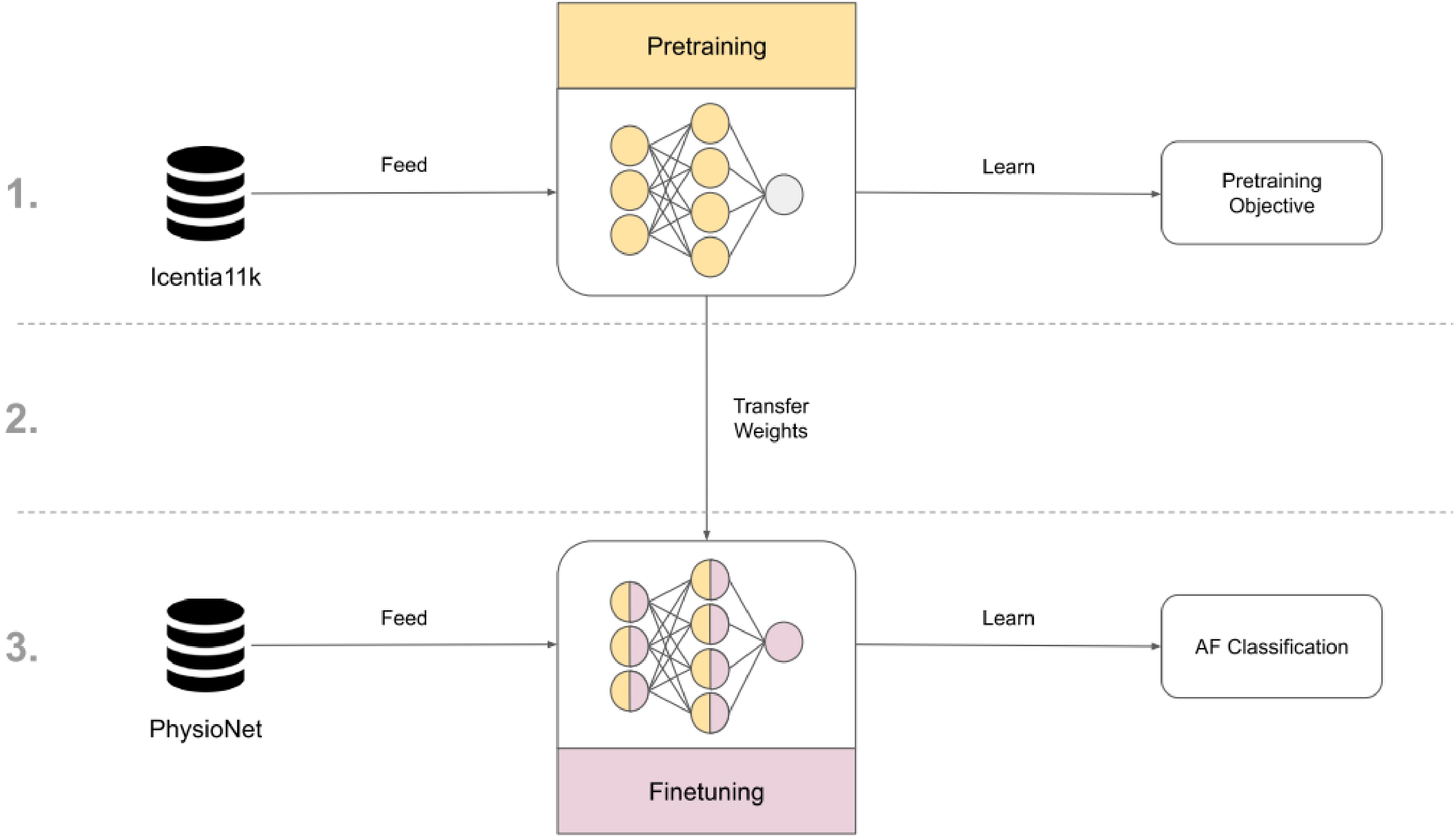
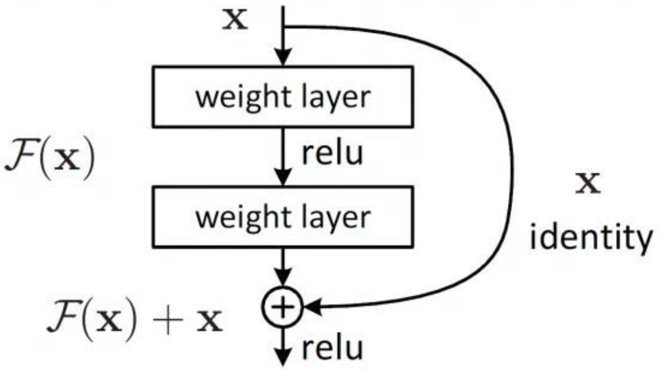
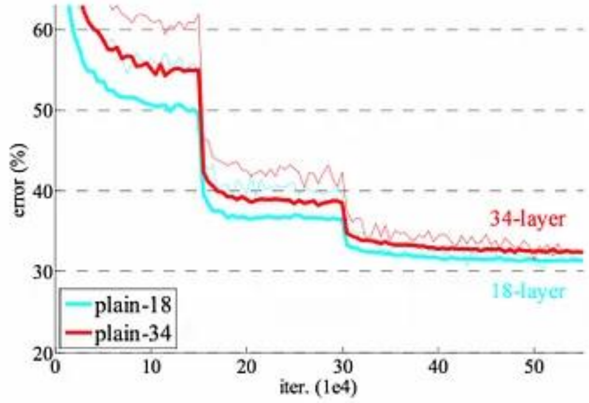


Figure 1. Visualization of transfer learning in this work. The process is divided into 3 steps: (1) deep convolutional neural network (CNN) is pretrained on the Icentia11K[5] data set for a selected pretraining objective, e.g. classification of heart rate; (2) the pretrained weights are used as initial weights of a new CNN; (3) this CNN is finetuned on the PhysioNet/CinC Challenge 2017[7,8] data set to classify Atrial Fibrillation (AF).

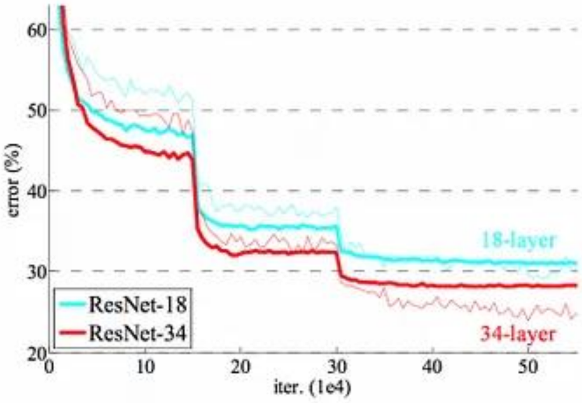
RestNet



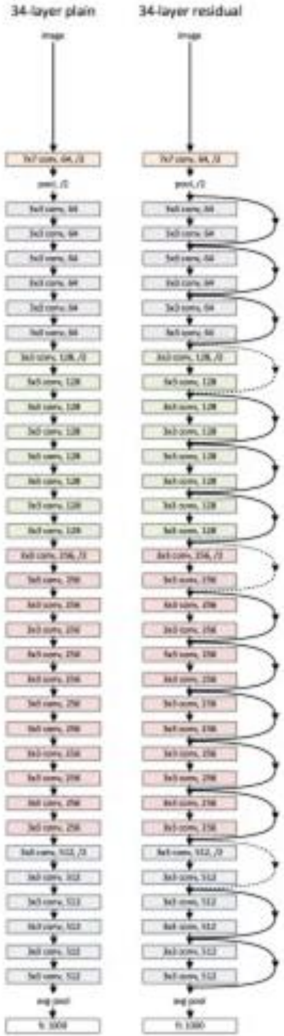
$\mathbf{x}$
weight layer
$\mathcal{F}(\mathbf{x})$
relu
weight layer
$\mathbf{x}$
identity
$\mathcal{F}(\mathbf{x}) + \mathbf{x}$
relu



error (%)
iter. (1e4)
34-layer
18-layer
plain-18
plain-34



error (%)
iter. (1e4)
18-layer
34-layer
ResNet-18
ResNet-34

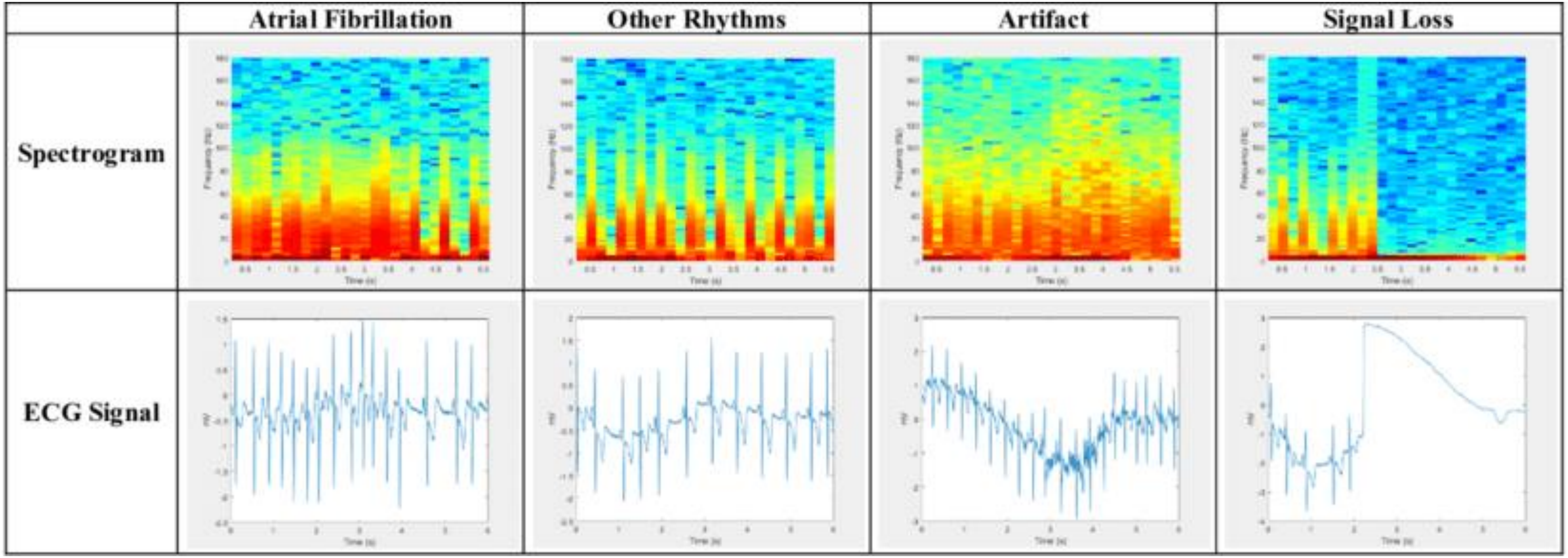


34-layer plain
34-layer residual

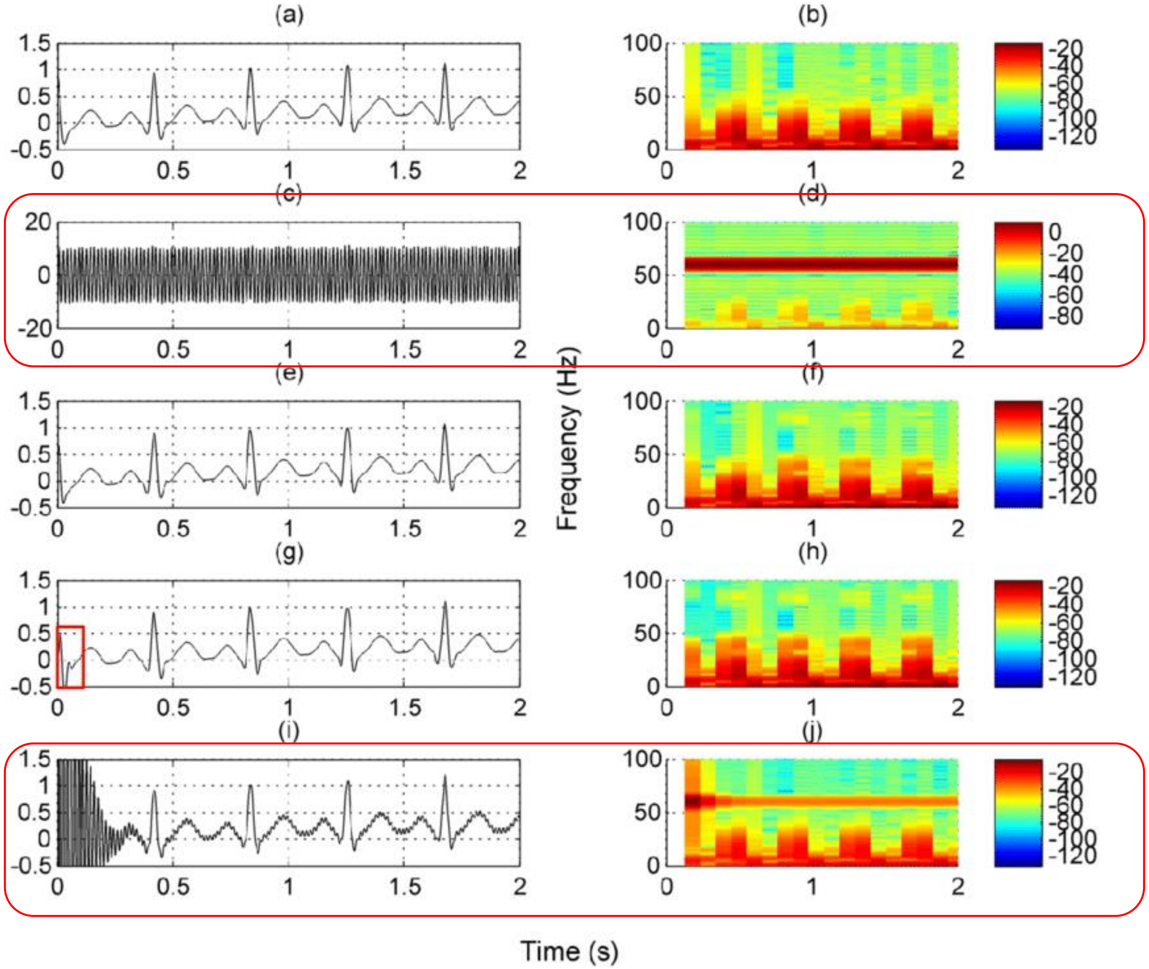
ImageNet

ImageNet Large Scale Visual Recognition Challenge (ILSVRC) from 2010 to 2017.

14,197,122 annotated images, **20k** categories



$$X(k) = \frac{1}{N} \sum_{n=0}^{N-1} x(n) \cdot e^{-j \frac{2\pi}{N} kn}$$



(a)
(b)
(c)
(d)
(e)
(f)
(g)
(h)
(i)
(j)
Frequency (Hz)
Time (s)

DATA SOURCE

Data used in this paper

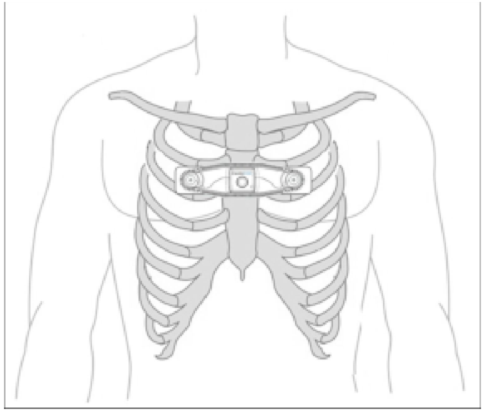
- The paper utilizes the largest public dataset of continuous raw ECG signals, the **Icentia11K** dataset, for pretraining the CNNs.
- The authors also mention the use of a smaller dataset for fine tuning the networks specifically for the classification of **Atrial Fibrillation (AFib)**, the most common heart arrhythmia.
- The **sampling frequency** of the ECG data is varied to investigate the performance of the pretrained networks on data with different frequencies.
- The paper mentions the use of **single lead ECG** data for **pretraining** the CNNs, which are then **fine tuned on 12 lead ECG data**.
- The authors highlight the exploration of both supervised and unsupervised pre training approaches, indicating the use of **labeled** and **unlabeled** data for training the CNNs.

Icentia11K Dataset

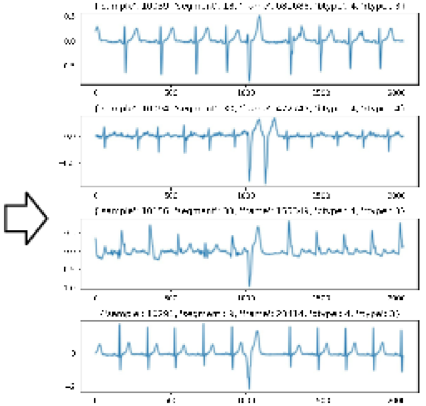
From 11,000 patients

CardioSTAT device **2 week**

EKG lead : **lead I** position, **250 Hz**



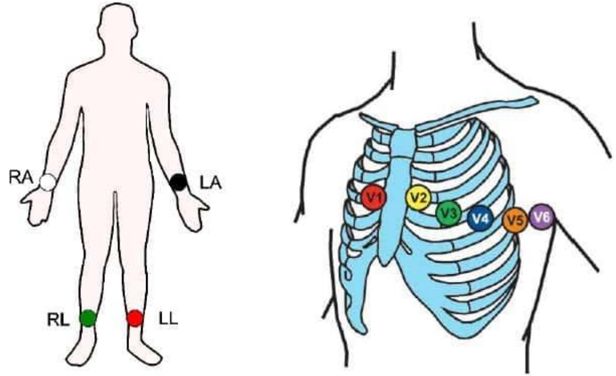
Device worn by over 11k patients



Dataset containing over 2 billion labelled beats
(Released free to the public)

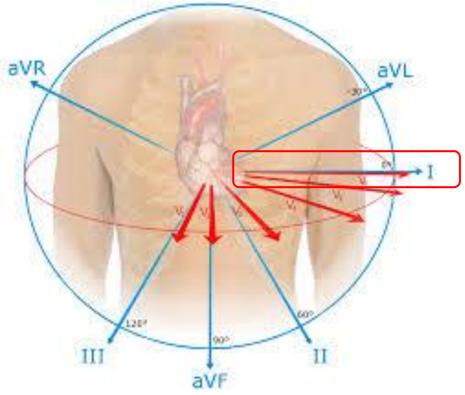
Beat labels
Normal Premature Atrial Contractions Premature Ventricular contractions

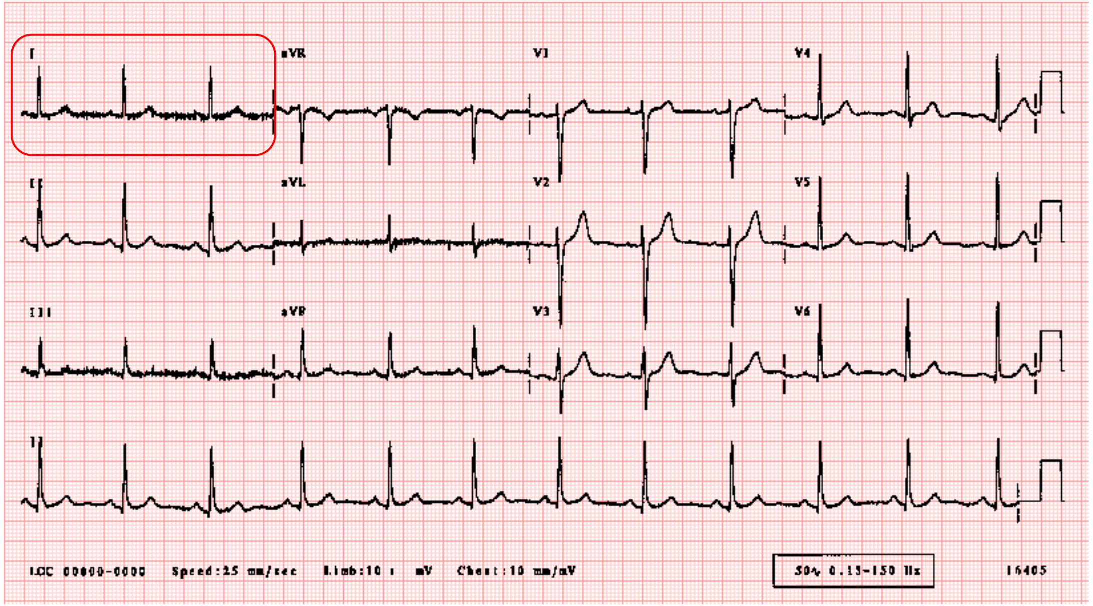
Rhythm Labels
NSR (Normal Sinusal Rhythm) AFib (Atrial Fibrillation) AFlutter (Atrial Flutter)



RA – right forearm or wrist
LA – left forearm or wrist
LL – left lower leg, proximal to ankle
RL – right lower leg, proximal to ankle
V1 – 4-th intercostal space, right sternal edge
V2 – 4-th intercostal space, left sternal edge
V3 – midway between V2 and V4
V4 – 5-th intercostal space, mid-clavicular line
V5 – anterior axillary line in straight line with V4
V6 – mid-axillary line in straight line with V4 and V5

Figure 23: 12 leads resting ECG electrode placement





PhysioNet/CinC Challenge 2017 Datasets

A total of 12,186 ECGs were used

8,528 in the public training set and

3,658 in the private hidden test set.

AliceCor device

(9-60s, 300Hz)

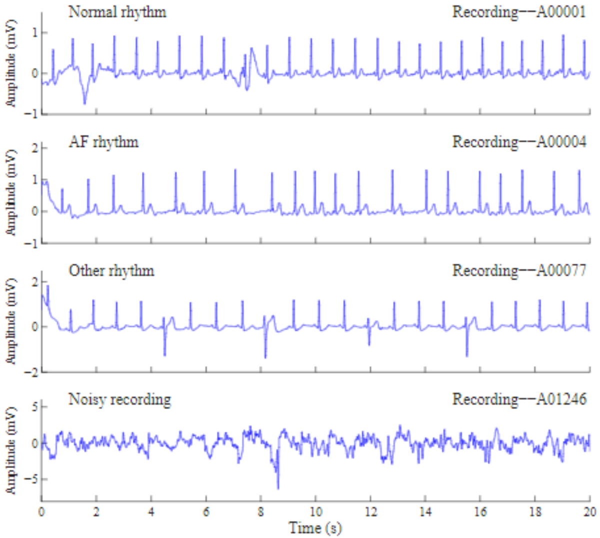
Adjust 300 Hz to 250Hz

Zero padding to 60s

Table 2: Data profile for the training set.

Type	# recording	Time length (s)				
		Mean	SD	Max	Median	Min
Normal	5154	31.9	10.0	61.0	30	9.0
AF	771	31.6	12.5	60	30	10.0
Other rhythm	2557	34.1	11.8	60.9	30	9.1
Noisy	46	27.1	9.0	60	30	10.2
Total	**8528**	**32.5**	**10.9**	**61.0**	**30**	**9.0**

Figure 1. Examples of the ECG waveforms.



Results of the paper

- The paper demonstrates the effectiveness of transfer learning for ECG classification, specifically for the classification of Atrial Fibrillation (AFib), the most common heart arrhythmia. Pretraining the CNNs on the **Icentia11K dataset** and fine tuning them on a smaller dataset for AFib classification **improves** the **performance** of the **CNNs** by up **to 6.57%** in terms of **macro F1 score**.
- The pretrained networks outperform **random weight initialization** in predicting every class, indicating the **effectiveness** of the **transfer learning** approach.

		Disease		Predictive Value	
		⊕	⊖		
Test	⊕	**A** True Positive (TP)	**B** False Positive (FP)	Positive Predictive Value (PPV) $\frac{TP}{TP+FP} = \frac{A}{A+B}$	Total Positive Results (A + B)
	⊖	**C** False Negative (FN)	**D** True Negative (TN)	Negative Predictive Value (NPV) $\frac{TN}{FN+TN} = \frac{D}{C+D}$	Total Negative Results (C + D)
Sensitivity & Specificity		Sensitivity $\frac{TP}{TP+FN} = \frac{A}{A+C}$	Specificity $\frac{TN}{FP+TN} = \frac{B}{B+D}$		
		All diseased patients (A + C)	All non-diseased patients (B + D)		

$$\text{Precision} = \frac{TP}{TP + FP}$$

$$\text{Recall} = \frac{TP}{TP + FN}$$

$$\text{F1 Score} = \frac{2}{\frac{1}{\text{Precision}} + \frac{1}{\text{Recall}}} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

$$\text{F1 Score} = \frac{TP}{TP + \frac{1}{2}(FP + FN)}$$

Pretraining method				Frame	F_1	F_{1n}	F_{1a}	F_{1o}	F_{1p}
None (random weight initialization)					.731 (± .019)	.898 (± .005)	.711 (± .027)	.701 (± .017)	.613 (± .062)
Beat classification				512	.769 (± .011)	.911 (± .010)	.760 (± .018)	.758 (± .016)	.647 (± .022)
				2048	**.779 (± .014)**	**.915 (± .007)**	**.777 (± .014)**	**.763 (± .014)**	**.661 (± .040)**
				4096	.768 (± .010)	.908 (± .009)	.764 (± .021)	.754 (± .015)	.646 (± .025)
Rhythm classification				512	.742 (± .017)	.896 (± .007)	.721 (± .026)	.716 (± .032)	.636 (± .045)
				2048	.767 (± .012)	.908 (± .004)	.753 (± .020)	.745 (± .018)	.660 (± .026)
				4096	.755 (± .005)	.903 (± .008)	.745 (± .022)	.735 (± .012)	.635 (± .017)
Heart rate classification				512	.766 (± .011)	**.915 (± .004)**	.759 (± .019)	.756 (± .015)	.635 (± .029)
				2048	.753 (± .013)	.910 (± .005)	.743 (± .037)	.738 (± .011)	.619 (± .039)
				4096	.751 (± .010)	.909 (± .006)	.744 (± .019)	.739 (± .016)	.611 (± .025)
	Context	**ns**	**Offset**	**Frame**					
Future prediction	8	4	2	512	.756 (± .008)	.903 (± .007)	.742 (± .011)	.730 (± .017)	.649 (± .021)
	16	8	2	512	.744 (± .016)	.905 (± .005)	.730 (± .027)	.730 (± .009)	.612 (± .041)
	16	8	8	512	.758 (± .013)	.908 (± .005)	.753 (± .021)	.745 (± .012)	.627 (± .026)
	16	16	8	512	.745 (± .013)	.897 (± .006)	.724 (± .024)	.722 (± .009)	.639 (± .034)

Table 1. Comparison of different configurations of the pretraining methods. For each method, we report the average macro F_1 score (and the standard deviation) on our test set for the PhysioNet/CinC Challenge 2017[7,8]. Additionally, we report the average F_1 score for each class: normal (F_{1n}), AF (F_{1a}), other (F_{1o}) and noisy (F_{1p}). *Frame* refers to the length of an ECG frame, *context* to the number of frames in the context, *ns* to the number of negative samples and *offset* to the distance between the context and the future frame measured in frames. All pretraining methods outperform random weight initialization in predicting every class.

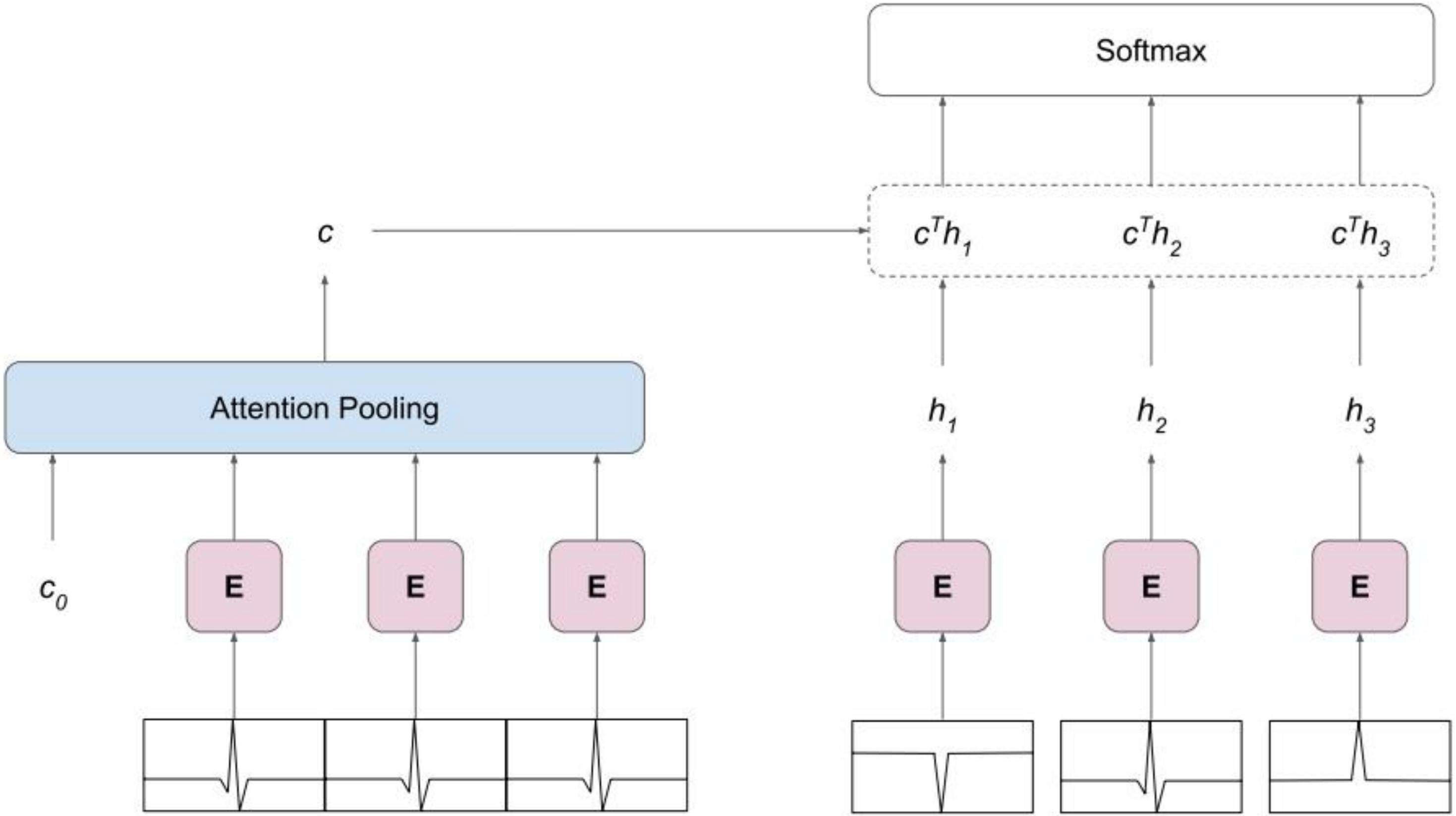
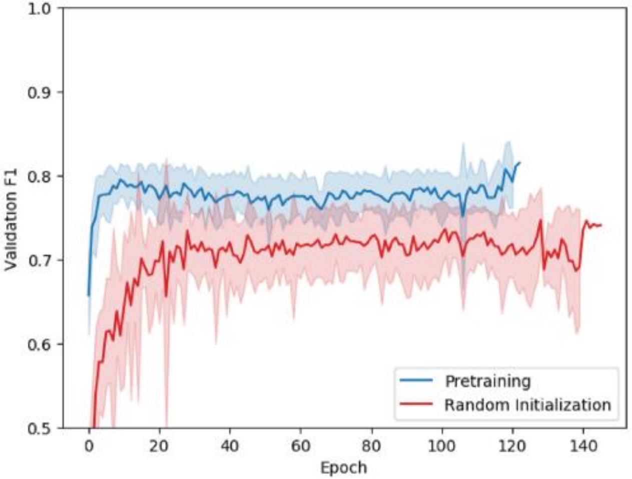
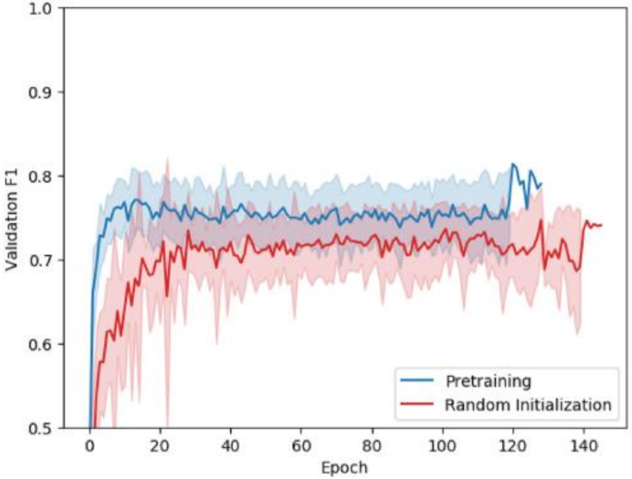


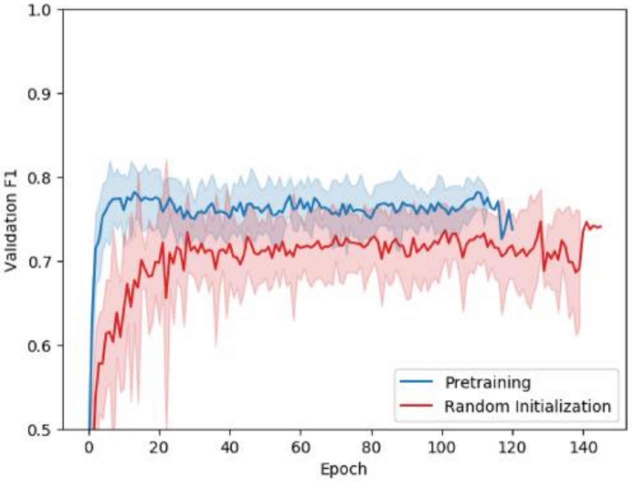
Figure 2. Architecture of the future prediction framework. The model predicts the correct future frame among negative samples (right) based on the present frames (left). Encoder **E** extracts features from ECG frames, producing a feature vector for each frame. Attention pooling summarizes feature vectors into a single context vector c describing the present. A dot product between c and frame encodings h_i gives the similarity between the context and future frames. The entire model is trained end-to-end with gradients backpropagated from the cross-entropy loss of classifying the future frame correctly.



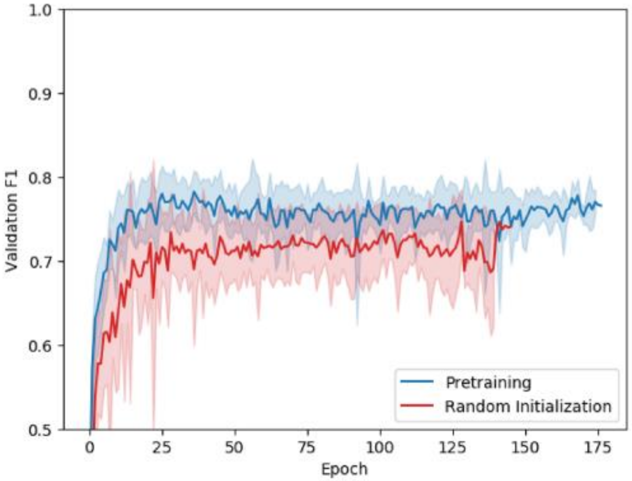
(a) Beat Classification



(b) Rhythm Classification



(c) Heart Rate Classification



(d) Future Prediction

Pretraining method	25% train	50% train	75% train
None (random weight initialization)	.670 (± .013)	.712 (± .010)	.731 (± .019)
Beat classification	**.739 (± .014)**	**.763 (± .011)**	**.779 (± .014)**
Rhythm classification	.707 (± .018)	.727 (± .028)	.767 (± .012)
Heart rate classification	.722 (± .010)	.749 (± .018)	.766 (± .011)
Future prediction	.694 (± .014)	.734 (± .011)	.758 (± .013)

Table 2. Comparison of the pretraining methods depending on the size of the downstream train set. For each method, we report the average macro F_1 score (and the standard deviation) on our test set for the PhysioNet/CinC Challenge 2017[7,8]. We examine 3 sizes of the train set as a proportion of the entire data set: 25%, 50% and 75% (original split). Pretraining allows models to be trained on less data and still achieve the same degree of performance as the same models that are not pretrained.

Pretraining method	128 Hz	250 Hz (Icentia11K)	300 Hz (PhysioNet)
None (random weight initialization)	.701 (± .017)	.731 (± .019)	.715 (± .023)
Beat classification	**.779 (± .012)**	**.779 (± .014)**	**.770 (± .011)**
Rhythm classification	.748 (± .012)	.767 (± .012)	.747 (± .017)
Heart rate classification	.761 (± .011)	.766 (± .011)	.767 (± .010)
Future prediction	.747 (± .008)	.758 (± .013)	.734 (± .016)

Table 3. Comparison of the pretraining methods depending on the sampling frequency (Hz) of the downstream data set. For each method, we report the average macro F_1 score (and the standard deviation) on our test set for the PhysioNet/CinC Challenge 2017[7,8]. Note that all networks are pretrained on ECG data sampled at 250 Hz, regardless of the sampling frequency during finetuning. Pretraining is beneficial even if networks are not specifically trained to deal with ECG data sampled at different frequencies.

Pretraining method	ResNet-18v2	ResNet-34v2	ResNet-50v2
None (random weight initialization)	.731 (± .019)	.764 (± .012)	.708 (± .023)
Beat classification	**.779 (± .014)**	**.794 (± .018)**	**.775 (± .015)**
Rhythm classification	.767 (± .012)	.775 (± .020)	.760 (± .008)
Heart rate classification	.766 (± .011)	.771 (± .008)	.761 (± .019)
Future prediction	.758 (± .013)	.761 (± .014)	.743* (± .010)

Table 4. Comparison of the pretraining methods depending on the architecture of the model (i.e. residual network). For each method, we report the average macro F_1 score (and the standard deviation) on our test set for the PhysioNet/CinC Challenge 2017[7,8]. Employing the ResNet-34v2 improves the performance of every pretraining method. We suspect that ResNet-34v2 lies in a sweet spot between model complexity and performance, whereas ResNet-18v2 underfits and ResNet-50v2 overfits to the training data. *Due to a spike in the model complexity, we only pretrain the first 3 stages of the ResNet-50v2.

Conclusions from the paper

- **Transfer learning** using deep convolutional neural networks (CNNs) **improves** the performance of ECG classification for **Atrial Fibrillation (AFib)** by up to **6.57%** compared to CNN's that are not pretrained. This reduces the number of annotations required for training CNNs for ECG classification.
- Both **supervised** and **unsupervised** pre training approaches are explored, with **supervised** pre training showing **greater improvement** in performance compared to **unsupervised** pre training. However, unsupervised pre training is considered relevant as it does not rely on expensive ECG annotations.
- The paper highlights the use of the largest public dataset of continuous raw ECG signals, the Icentia11K dataset, for pretraining the CNNs. Additionally, a smaller dataset is used for fine tuning the networks specifically for AFib classification.
- The pretrained CNNs outperform **random weight initialization** in predicting different heart rhythms, indicating the effectiveness of transfer learning for ECG classification.

Limitations of this paper

- The paper does not investigate the **impact** of **different label choices** in the classification tasks, which could potentially affect the performance of the pretrained **feature extractors**.
- The study **focuses** on the classification of Atrial Fibrillation (**AFib**) and does not explore the transfer learning approach for **other types** of heart **arrhythmias** or ECG abnormalities.


THANK YOU