Real-world Challenges in Leveraging Electrocardiograms for Coronary Artery Disease Classification

Jessica K. De Freitas
Hasso Plattner Institute for Digital Health
Icahn School of Medicine at Mount Sinai
New York, NY 10065
jessica.defreitas@icahn.mssm.edu

Alexander W. Charney
Department of Genetics and Genomic Sciences
Icahn School of Medicine at Mount Sinai
New York, NY 10065
alexander.charney@icahn.mssm.edu

Isotta Landi
Department of Medicine
Icahn School of Medicine at Mount Sinai
New York, NY 10065
isotta.landi2@mssm.edu

Abstract

This work investigates coronary artery disease (CAD) prediction from electrocardiogram (ECG) data taking into account different windows with respect to the time of diagnosis. We report that ECG waveform measurements automatically collected during ECG recordings contain sufficient features for good classification of CAD using machine learning models up to five years before diagnosis. On the other hand, convolutional neural networks trained on the ECG signals themselves appear to best extract CAD related features when processing data collected within one year before or after a diagnosis is made. Through this work we demonstrate that the type of ECG data and the time window with respect to diagnosis should guide model selection.

1 Introduction

Coronary artery disease (CAD) is a chronic and heterogeneous condition affecting millions of people and is a leading cause of death in the US and worldwide [Tsao et al., 2022]. A diagnosis is usually confirmed by invasive, time-consuming, and costly methods (e.g., angiography) [Fihn et al., 2012]. The electrocardiogram (ECG), on the other hand, is a quick and affordable test that measures the electrical activity of the heart, providing information on cardiac function. However, ECG changes indicative of CAD may only become apparent to a trained physician when the condition has progressed in severity [Mahmoodzadeh et al., 2011]. Machine learning (ML) and, in particular, deep learning (DL) methods have successfully detected a variety of cardiac conditions from ECG signals with high performance [Hughes et al., 2021, Galloway et al., 2019]. This suggests that there may be subtle changes in an ECG that are beyond human detection but that a trained system could identify and use to classify complex conditions such as CAD.

Such an approach would allow for routine CAD screening leading to earlier detection, efficient resource allocation, and ultimately prevention of disease progression. However, in most studies, models are trained on data within a relatively small window from diagnosis and do not take into account the possible challenges arising when working with electronic health records (EHRs). Within an EHR system, multiple measurements are collected at varying frequency for patients in the same

36th Conference on Neural Information Processing Systems (NeurIPS 2022).

cohort. The diagnosis itself can be given to a patient at various points in the clinical course of the disease. Therefore, data associated with the same diagnosis can be heterogeneous and reflect different disease states, which can both facilitate or hinder classification. Leveraging EHRs from a large healthcare system, we aim to classify patients with CAD from their ECGs while investigating how signals taken at different times from diagnosis modulate CAD prediction.

1.1 Related Works

A number of studies using ML/DL show promising performance in identifying a variety of cardiac abnormalities, however these have been leveraging data collected within a small window. Raghunath et al. [2021] used ECGs within 1 year from diagnosis to predict atrial fibrillation in patients that had no history of the condition. Another study focused on CAD prediction using ECGs collected up to 30 days prior to an angiogram to predict the presence and location of CAD from the procedure reports [Huang et al., 2022]. Chen et al. [2022] leveraged ECGs taken within 7 days from diagnosis to predict structural abnormalities found in echocardiograms. These papers demonstrate the ability of ML/DL to identify disease in a more scalable and cost-effective manner using ECG around the same time point of a gold-standard diagnosis.

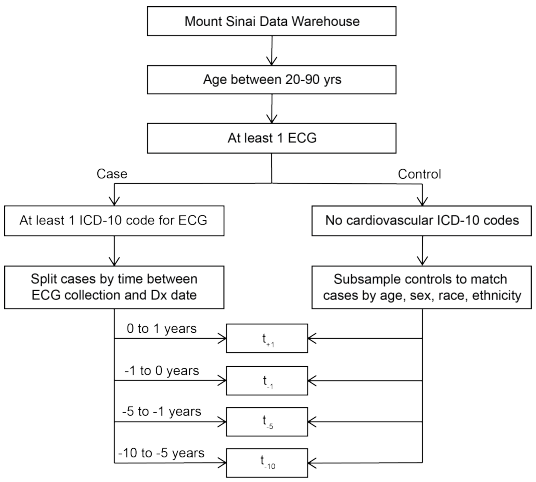
However, it remains an open question whether a model could also diagnose patients with appreciable disease but with a diagnosis that was recorded later in their EHRs, and thus excluded by design from previous studies. The first contribution of this work is the evaluation of how model performance in predicting CAD is affected by the time from ECG to diagnosis. Second, we investigate whether, to efficiently classify CAD in the real-world, simpler ML models, that require less extensive experimentation and lower computational resources, could be preferable than DL approaches. Towards this goal we evaluated: (1) baseline models applied to automatically extracted waveform measurements from ECG; (2) convolutional neural networks (CNNs) trained on ECG signals.

2 Data Description

2.1 Cohort Selection

Using EHRs from the Mount Sinai Health System, a large, ethnically and racially diverse hospital network in New York City, cases were selected as those patients between the age of 20 and 90 year with at least one ICD-10-CM code for "Atherosclerotic heart disease of native coronary artery" (i.e., "I25.1"). We define diagnosis date as the first appearance of the ICD code in the patient's EHR and use it to divide the cases into four datasets: a 't-after' (t_{+1}), comprising ECGs collected up to one year after diagnosis, and three 't-before' datasets comprising ECGs collected up to one year (t_{-1}), from one to 5 years (t_{-5}), and from 5 to 10 years (t_{-10}) before the diagnosis date.

Figure 1: Cohort Selection Strategy



Controls were defined as those patients without any cardiac ICD-10-CM codes, i.e. falling under the "I" category of "Diseases of the circulatory system." We selected controls to match the case population within each of the four datasets according to age, sex, race and ethnicity. See Figure 1 for summary of cohort selection strategy. The demographic characteristics of the final cohort can be found in appendix A, Tables 4 and 5.

	Datasets			
	t_{+1}	t_{-1}	t_{-5}	t_{-10}
Case: train+dev (test)	$49,684\ (12,422)$	$40,611\ (10,152)$	$21,820\ (5,455)$	$14,434\ (3,608)$
Control: train+dev (test)	$49,685\ (12,421)$	$40,611\ (10,152)$	$21,820\ (5,455)$	$14,434\ (3,608)$

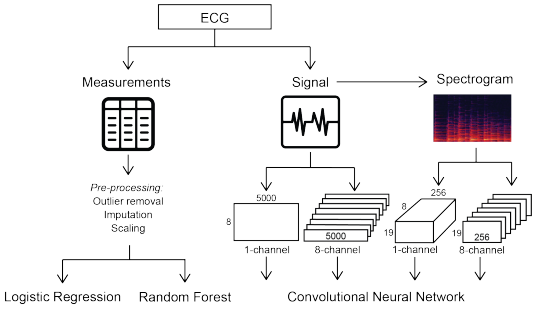
Table 1: Counts of ECGs in each time-split dataset.

2.2 ECG Data

For each patient in the cohort we considered ECGs repeatedly collected throughout their medical history, however we included at most one ECG per dataset as to avoid bias from patients with high numbers of ECGs collected within the given time period. We obtained a total of $317,768$ ECGs, split evenly between cases and controls ($158,186$ in each group respectively). See appendix B Figure 3 for a visualization of time from diagnosis date to ECG recording. The entire cohort is comprised of $196,159$ unique patients with $87,512$ cases and $108,661$ controls. The case group has an average of 1.8 ECGs per patient while control group has an average of 1.5 ECGs. It is expected that CAD cases have a higher number of ECGs conducted as they would likely have more encounters with the health system in order to monitor their condition.

ECG data presents as (1) automatically calculated measurements describing characteristics of the signal; (2) signals of voltage over time. Measurement data are numeric and represent counts, rates, lengths, and onset/offset of sub-waves. Signal data are comprised of 8 waveforms, known as leads, which show voltage over time and are sampled at a frequency of $500\ s^{-1}$ for 10 seconds, resulting in vectors of length $5,000$. The leads measure the same electrical event but from different angles.

Figure 2: Description of ECG data



3 Methods

3.1 Predicting CAD from ECG Measurements

Preprocessing Measurement features were filtered by dropping those with $> 10\%$ missingness across the entire dataset, resulting in the 15 variables listed in appendix B Table 6. Individual observations were dropped if missing $> 30\%$ of features or if containing outliers as determined by a 0.005% quantile filter. Missing values were imputed using a k-nearest neighbors strategy with $k = 5$, euclidean distance, and uniform weights. Finally, data was scaled using a min-max strategy.

Models For CAD classification from ECG measurements, we chose logistic regression and random forest. Details about models and hyperparameter selection can be found in the Appendix C.

3.2 Predicting CAD from ECG Signal and Spectrogram

Preprocessing We applied band-pass and median signal filtering to the signal data to reduce the baseline drift of the biological signal and the effect of motion artifacts and high frequency noise [Fedotov, 2016]. Along with the processed signal, we also considered its spectrogram that, converting the signal from the time domain into the frequency domain, represents the signal's amplitude at different frequencies over time [Rioul and Vetterli, 1991]. We decided to include the spectrogram data format because a frequency-focused representation may add information that can be leveraged by the models.

Model Architectures Separate CNNs were trained on signal and spectrogram data formats. Input configurations included a 1-channel, as well as an 8-channel configuration for both signal and spectrogram input formats, resulting in four models with different input configurations. For the ECG signal: (A) tensors of dimension $1 \times (8 \times 5,000)$ and (B) matrices of dimension $8 \times (5,000)$; for the spectrogram: (C) tensors of dimension $1 \times (8 \times 19 \times 256)$ and (D) tensors of dimension $8 \times (19 \times 256)$.

A learning rate scheduler was used to reduce the learning rate by a factor of 10 after the cross-entropy loss reached a plateau over a number of epochs, i.e., no decrease in loss after 8 epochs. The models used an Adam optimizer starting at a learning rate of 10^{-3}. Because the datasets vary in number of samples, models were trained for the same number of steps (3.7 million) for fair comparison. The specific convolutional kernel size, max pool kernel size, and number of convolutional/ReLU blocks for each model and other hyperparameters are described in the appendix D, Table 9.

3.3 Study Design

For both ML and CNN classification, each of the four datasets was split into 60% training, 20% development, and 20% testing. When creating the splits, data was grouped at the patient level to ensure that an individual patient's ECGs only appeared in a single split and to avoid biasing results.

All models were trained on each of the four time-windowed datasets and hyperparameters were selected based on development set performance. We then evaluated the best models on the held-out test set. We also evaluated the models trained on t_{+1} on each of the three test sets t_{-1}, t_{-5}, and t_{-10}. This was done to ultimately investigate whether CAD features can be automatically leveraged in real-world scenarios for early predictions. Performance was assessed with the F_1 score. Computational time for cohort selection, ML model training, and CNN training totals $\sim 1,200$ hours on CPU, A100 GPU, or V100 GPU.

4 Results

As observed in Table 2, when identifying CAD from ECG measurements, a CNN model with 8-channel input format performed best in the t_{+1} and t_{-1} datasets but random forest performed best in the t_{-5} and t_{-10} datasets.

	Datasets			
Models	t_{+1}	t_{-1}	t_{-5}	t_{-10}
Log. Regression	0.760 (0.763)	0.733 (0.722)	0.747 (0.741)	0.594 (0.600)
Random Forest	0.769 (0.771)	0.763 (0.754)	**0.753 (0.747)**	**0.635 (0.633)**
Signal 1-channel	0.788 (0.866)	0.765 (0.758)	0.707 (0.703)	0.580 (0.608)
Signal 8-channel	**0.791 (0.856)**	**0.781(0.778)**	0.709 (0.710)	0.617 (0.623)
Spect. 1-channel	0.773 (0.838)	0.764 (0.758)	0.740 (0.734)	0.667 (0.665)
Spect. 8-channel	0.775 (0.830)	0.777 (0.765)	0.749 (0.734)	0.665 (0.656)

Table 2: F_1 score of models evaluated on the same time frame used for training. Test (dev).

Among the random forest models, the best performance was in t_{+1}, with performance decreasing the further away from diagnosis date (0.78% reduction in the t_{-1} dataset, 2.1% reduction in the t_{-5} dataset, and 17.4% reduction in the t_{-10} dataset).

The best performing CNN models used signal with 8-channel configuration. However across the datasets, there was no clear preference for signal over spectrogram, as shown in the higher performance of spectrogram with 1-channel inputs performing best in the t_{-5} and t_{-10} datasets. With the 8-channel signal CNN model, the best performance is in t_{+1}, which then decreases the further away from diagnosis date (1.3% reduction in the t_{-1} dataset, 10.4% reduction in the t_{-5} dataset, and 22.0% reduction in the t_{-10} dataset). In conclusion, we found that signal data achieves higher performance at CAD identification closer to diagnosis date in t_{+1} and t_{-1} but that measurement data performs better in the t_{-5} and t_{-10} datasets. Additionally we find that models maintain good performance in t_{-1} and t_{-5} datasets relative to data taken after diagnosis in t_{+1}.

Table 3: F_1 score: trained 't-after' model evaluated on 't-before' datasets; test(dev)

	Datasets		
Models	t_{-1}	t_{-5}	t_{-10}
Logistic Regression	$0.762\ (0.754)$	$0.744\ (0.741)$	$0.649\ (0.641)$
Random Forest	$0.770\ (0.777)$	**0.766 (0.752)**	**0.671 (0.656)**
Signal 1-channel	$0.771\ (0.763)$	$0.703\ (0.696)$	$0.545\ (0.545)$
Signal 8-channel	**0.771 (0.766)**	$0.696\ (0.694)$	$0.531\ (0.544)$
Spectrogram 1-channel	$0.768\ (0.751)$	$0.732\ (0.728)$	$0.605\ (0.594)$
Spectrogram 8-channel	$0.762\ (0.749)$	$0.705\ (0.699)$	$0.555\ (0.547)$

Table 3 reports the results of the models trained on the 't-after' and evaluated on the 't-before' datasets. Again we observe that the CNN performs better in the t_{-1} dataset while the ML models were able to achieve better results over the CNN models in t_{-5} and t_{-10} datasets although less striking than in the results reported above. Again we see performance decreasing the further away from diagnosis.

5 Discussion

In this study we acknowledge a number of limitations. We use the CAD ICD-10 code as the label and its first occurrence as the patient's diagnosis date, however it is possible that patients received a diagnosis earlier at another health system or that it was recorded in a clinical note or in the report of another diagnostic test. Although this is the most common way of cohort selection, in future work the CAD label could also be extracted from clinical notes, or for the sub-group of patients with angiograms in their procedure reports. Another limitation is that by grouping the 't-before' datasets into one to five year bins, we lose some temporal granularity in assessing model performance. In future work we could create smaller time bins to perform a more fine-grain temporal classification analysis.

Our results suggest that ECG measurements, up to 5 years before diagnosis, can contain sufficient features for CAD prediction. With respect to study design, a larger time window could allow for increased sample size as more ECGs could be considered for cohort inclusion. Additionally, using a machine learning model would offer benefits in terms of reduced carbon footprint, training time, and more manageable model size over deep learning models trained on ECG signal. Although the CNN models do perform slightly better overall, they do show a preference for datasets closer to the diagnosis date, +/- 1 year. Moreover, when models were trained on the 't-after' dataset and tested on the 't-before' datasets, CNN showed results comparable to baselines, in particular on the t_{-1} dataset. This warrants further investigation into what features are extracted from the ECG signal after diagnosis by the CNN models, that might not be caught by ECG measurements. In conclusion, we argue that, when leveraging EHRs, the type of ECG data and the time window with respect to diagnosis should guide model decisions for optimized CAD predictions towards precision medicine.

References

Connie W. Tsao, Aaron W. Aday, Zaid I. Almarzooq, Alvaro Alonso, Andrea Z. Beaton, Marcio S. Bittencourt, Amelia K. Boehme, Alfred E. Buxton, April P. Carson, Yvonne Commodore-Mensah, Mitchell S.V. Elkind, Kelly R. Evenson, Chete Eze-Nliam, Jane F. Ferguson, Giuliano Generoso, Jennifer E. Ho, Rizwan Kalani, Sadiya S. Khan, Brett M. Kissela, Kristen L. Knutson, Deborah A. Levine, Tené T. Lewis, Junxiu Liu, Matthew Shane Loop, Jun Ma, Michael E. Mussolino, Sankar D. Navaneethan, Amanda Marma Perak, Remy Poudel, Mary Rezk-Hanna, Gregory A. Roth, Emily B. Schroeder, Svati H. Shah, Evan L. Thacker, Lisa B. VanWagner, Salim S. Virani, Jenifer H. Voecks, Nae-Yuh Wang, Kristine Yaffe, Seth S. Martin, and null null. Heart disease and stroke statistics—2022 update: A report from the american heart association. *Circulation*, 145 (8):e153–e639, 2022. doi: 10.1161/CIR.0000000000001052. URL `https://www.ahajournals.org/doi/abs/10.1161/CIR.0000000000001052`.

Stephan D Fihn, Julius M Gardin, Jonathan Abrams, Kathleen Berra, James C Blankenship, Apostolos P Dallas, Pamela S Douglas, JoAnne M Foody, Thomas C Gerber, Alan L Hinderliter, et al. 2012 accf/aha/acp/aats/pcna/scai/sts guideline for the diagnosis and management of patients with stable ischemic heart disease: a report of the american college of cardiology foundation/american heart association task force on practice guidelines, and the american college of physicians, american association for thoracic surgery, preventive cardiovascular nurses association, society for cardiovascular angiography and interventions, and society of thoracic surgeons. *Circulation*, 126 (25):e354–e471, 2012.

Solmaz Mahmoodzadeh, Mansour Moazenzadeh, Hamidreza Rashidinejad, and Mehrdad Sheikhvatan. Diagnostic performance of electrocardiography in the assessment of significant coronary artery disease and its anatomical size in comparison with coronary angiography. *Journal of research in medical sciences: the official journal of Isfahan University of Medical Sciences*, 16(6):750, 2011.

J. Weston Hughes, Jeffrey E. Olgin, Robert Avram, Sean A. Abreau, Taylor Sittler, Kaahan Radia, Henry Hsia, Tomos Walters, Byron Lee, Joseph E. Gonzalez, and Geoffrey H. Tison. Performance of a Convolutional Neural Network and Explainability Technique for 12-Lead Electrocardiogram Interpretation. *JAMA Cardiology*, 6(11):1285–1295, 11 2021. ISSN 2380-6583. doi: 10.1001/jamacardio.2021.2746. URL `https://doi.org/10.1001/jamacardio.2021.2746`.

Conner D Galloway, Alexander V Valys, Jacqueline B Shreibati, Daniel L Treiman, Frank L Petterson, Vivek P Gundotra, David E Albert, Zachi I Attia, Rickey E Carter, Samuel J Asirvatham, et al. Development and validation of a deep-learning model to screen for hyperkalemia from the electrocardiogram. *JAMA cardiology*, 4(5):428–436, 2019.

Sushravya Raghunath, John M Pfeifer, Alvaro E Ulloa-Cerna, Arun Nemani, Tanner Carbonati, Linyuan Jing, David P vanMaanen, Dustin N Hartzel, Jeffery A Ruhl, Braxton F Lagerman, et al. Deep neural networks can predict new-onset atrial fibrillation from the 12-lead ecg and help identify those at risk of atrial fibrillation–related stroke. *Circulation*, 143(13):1287–1298, 2021.

Pang-Shuo Huang, Yu-Heng Tseng, Chin-Feng Tsai, Jien-Jiun Chen, Shao-Chi Yang, Fu-Chun Chiu, Zheng-Wei Chen, Juey-Jen Hwang, Eric Y Chuang, Yi-Chih Wang, et al. An artificial intelligence-enabled ecg algorithm for the prediction and localization of angiography-proven coronary artery disease. *Biomedicines*, 10(2):394, 2022.

Hung-Yi Chen, Chin-Sheng Lin, Wen-Hui Fang, Chia-Cheng Lee, Ching-Liang Ho, Chih-Hung Wang, and Chin Lin. Artificial intelligence-enabled electrocardiogram predicted left ventricle diameter as an independent risk factor of long-term cardiovascular outcome in patients with normal ejection fraction. *Frontiers in medicine*, 9, 2022.

AA Fedotov. Selection of parameters of bandpass filtering of the ecg signal for heart rhythm monitoring systems. *Biomedical Engineering*, 50(2):114–118, 2016.

O. Rioul and M. Vetterli. Wavelets and signal processing. *IEEE Signal Processing Magazine*, 8(4): 14–38, 1991. doi: 10.1109/79.91217.

A Appendix

The following tables describe demographic information about the selected cohort.

	Age			
Cohort	t_{+1}	t_{-1}	t_{-5}	t_{-10}
Train+dev: mean (sd)	$67.3(11.6)$	$66.9(12.1)$	$65.6(12.0)$	$62.4(12.1)$
Test: mean (sd)	$67.3(11.7)$	$67.1(12.0)$	$65.3(12.0)$	$62.1(11.6)$
Train+dev: median (range)	$68(20-90)$	$67(20-90)$	$66(20-90)$	$63(20-90)$
Test: median (range)	$68(20-90)$	$67(20-90)$	$66(20-90)$	$63(20-89)$

Table 4: Age of patients in cohort

Race/Ethnicity	t_{+1}	t_{-1}	t_{-5}	t_{-10}
Caucasian or White	42.7%	39.4%	38.9%	38.9%
Black or African-American	13.1%	14.5%	16.9%	16.7%
Hispanic or Latino	11.8%	13.2%	16.6%	17.1%
Other	11.5%	15.3%	13.9%	14.7%
Unknown	18.2%	13.6%	10.9%	9.7%
Asian	2.6%	3.6%	2.5%	2.7%
American Indian or Alaska Native	0.13%	0.18%	0.16%	0.15%
Native-Hawaiian or Pacific Islander	0.06%	0.12%	0.14%	0.10%

Table 5: Race/Ethnicity of patients in cohort

B Appendix

The following figure and table describe ECG data characteristics.

Figure 3: Time from Diagnosis to ECG

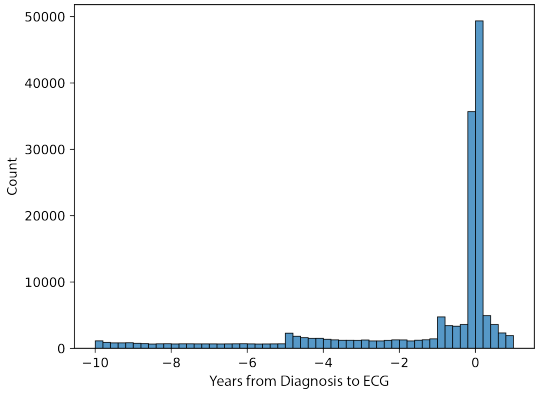


Table 6: Description of measurement data.

Variable name	Description
VentricularRate	Ventricular rate in BPM
AtrialRate	Atrial rate (in BPM)
PRInterval	P-R interval (in msec)
QRSDuration	QRS duration (in msec)
QTInterval	QT interval (in msec)
QTCorrected	Bazett's Algorithm
PAxis	P axis
RAxis	R axis
TAxis	T axis
QRSCount	QRS count
QOnset	Q onset (median complex sample point)
QOffset	P onset (median complex sample point)
POnset	P onset (median complex sample point)
POffset	P offset (median complex sample point)
TOffset	T offset (median complex sample point)

C Appendix

Hyperparameters were chosen as those performing best on the development set in terms of $F1$ score.

Logistic Regression (LR) Hyperparameters were obtained via grid search over the following parameter space:

```
lr_grid =
    {"C": [100, 10, 1.0, 0.1, 0.01, 0.001, 0.0001, 0.00001, 0.000001, 0.0000001],
    "penalty": ["l2", "l1"]}
```

Table 7: Best performing hyperparameters.

	t_{+1}	t_{-1}	t_{-5}	t_{-10}
C	0.1	0.001	100	0.1
Penalty type	L2	L2	L2	L2

Random Forest (RF) Hyperparameters were obtained via randomized search over the following parameter space:

```
rf_grid =
    {"n_estimators": [int(x) for x in np.linspace(start=10, stop=300, num=4)],
    "max_depth": [int(x) for x in np.linspace(1, 80, num=4)],
    "max_features": ["auto", "sqrt"],
    "min_samples_leaf": [1, 2, 4, 8],
    "min_samples_split": [2, 5, 10]}
```

Table 8: Best performing random forest hyperparameters.

	t_{+1}	t_{-1}	t_{-5}	t_{-10}
Number of estimators	300	300	300	203
Maximum depth of the tree	80	80	27	27
Number of features when considering split	auto	auto	auto	auto
Minimum number of samples at a Leaf node	1	1	2	8
Minimum number of samples to Split node	2	2	2	2

D Appendix

In general the CNNs, have a batch norm layer, followed by a variable number of blocks consisting of a convolutional layer, ReLU layer, and a max pooling layer. Following these blocks, there is single fully connected linear layer, and softmax layer for binary classification.

Table 9: CNN architectures.

Data type	Input configuration	Kernel	Max pool	Conv. blocks
Signal	1-channel	(8, 5)	(3,3)	5
Signal	8-channel	5	3	5
Spectrogram	1-channel	(5, 5, 5)	(2, 2, 2)	3
Spectrogram	8-channel	(5, 5)	(2, 2)	4