Detecting Atrial Fibrillation in ICU Telemetry data with Weak Labels

Brian Chen BRIAN.CHEN@QUEENSU.CA
School of Computing
Queen's University
Kingston, ON, Canada

Golara Javadi GOLARA@ECE.UBC.CA
Department of Electrical and Computer Engineering
University of British Columbia
Vancouver, BC, Canada

Amoon Jamzad A.JAMZAD@QUEENSU.CA
School of Computing
Queen's University
Kingston, ON, Canada

Alexander Hamilton ALEXANDER.HAMILTON@QUEENSU.CA
Human Mobility Research Centre
Queen's University
Kingston, ON, Canada

Stephanie Sibley STEPHANIE.SIBLEY@KINGSTONHSC.CA
Department of Critical Care Medicine
Queen's University
Kingston, ON, Canada

Purang Abolmaesumi PURANG@ECE.UBC.CA
Department of Electrical and Computer Engineering
University of British Columbia
Vancouver, BC, Canada

David Maslove DAVID.MASLOVE@QUEENSU.CA
Department of Critical Care Medicine
Queen's University
Kingston, ON, Canada

Parvin Mousavi MOUSAVI@QUEENSU.CA
School of Computing
Queen's University
Kingston, ON, Canada

Abstract

State of the art techniques for creating ML models in healthcare often require large quantities of clean, labelled data. However, many healthcare organizations lack the capacity to generate the large-scale annotations required to create and validate reliable labels. In this paper, we demonstrate how raw data from an information-rich area of care can be exploited without the need for mass manual annotation via the use of weak labels. We evaluate the proposed framework on telemetry data from the intensive care unit for application of atrial fibrillation (AF) detection. We generate an in-house dataset of over 60,000 ECG segments with weak labels, derived from a model trained on publicly available data. We then show that building a deep learning model based on these weakly generated labels can significantly improve (more than 30%) the performance of AF detection in comparison with only using limited expert-annotated ground truth labels. We further demonstrate how weakly supervised learning techniques can be used to augment and control the level of noise in these weak labels. Lastly, we explore how supervised fine-tuning effects the performance of these models and discuss the viability of leveraging weak labels for large-scale atrial fibrillation detection and identification.

1. Introduction

As machine learning (ML) techniques gain more of a foothold in the healthcare space, many practical challenges have arisen around the data required for developing these models. Chief among these challenges is how to gather and handle a sufficient quantity of data for ML algorithms, especially where ground truth labels are needed. This has been exacerbated by the recent rise of techniques like deep learning that necessitate orders of magnitude more samples in order to achieve state-of-the-art levels of performance.

The challenge of gathering an adequately-sized ground truth for ML-based analysis crosses medical domains and modalities. Medical imaging, digital pathology and computational genomics all encounter similar issues with a paucity of annotated data Maslove et al. (2017). Physiological data is no exception despite its prevalence compared to other sources. Datasets with clean, comprehensive and sufficiently granular ground truth are rare, while the heterogeneity of different patient populations changes how well external datasets can generalize to a particular institution's patient population. Most notably, there is a severe lack of cost effective approaches for manually annotating these data. Unlike non-medical ML domains such as general computer vision, where one can crowd-source data annotation using systems such as CAPTCHA, medical data often require expert interpretation from clinicians and healthcare staff, and thus would be prohibitively expensive to employ for large scale annotation.

Intensive care unit (ICU) records of inpatient data offer a rich and diverse source of physiological signals for potential analysis. These data include many vital signs (usually taken at periodic intervals) and continuous waveform signals collected at high frequency such as electrocardiography (ECG) for heart rhythm analysis, photoplethysmography (PPG) for pulse oximetry, and arterial blood pressure (ABP) measurement via indwelling catheters. These data are less sparse but can be very noisy and suffer from issues such as sensor drop-off because they are sensitive to patient movement and other environmental factors. ICU monitoring of ECG waveforms alone can generate GBs of data daily. The majority of these data points come in the form of streams that are continuously displayed at the bedside. However, the quantity of data in these streams generally limits manual inspection to deviations that trigger alarms based on static thresholds and signal quality, or at semi-regular intervals when a clinician deems it necessary. Moreover, clinician-created annotations at the bedside are susceptible to collection time versus entry time drift (e.g. when a bedside nurse prints out a 10 second diagnostic strip and when they create the corresponding medical record entry). From a labelling standpoint, particularly for ML, these continuous data become sparse in much the same way that plagues the handling of vital signs. Working with raw signals in near-real time could be especially valuable for automated monitoring and alerts as they contain a wealth of information about the patient's state, possible conditions, and risk factors, all of which could be used to inform swift clinical interventions, if necessary. As such, physiological signals are a promising proving ground for investigating methods of generating less confident but more numerous labels.

Atrial fibrillation (AF) is one of the most prevalent arrhythmias in the ICU (Bosch et al., 2018). Estimates vary, but between 4% and 30% of ICU patients will develop AF during their stay (Seguin and Launey, 2010). From a diagnostic perspective, this condition can be identified through changes in the morphology of an ECG signal. This includes loss of the P-wave, and irregularity in the R-R interval (Hindricks et al., 2021). AF carries a great deal of downstream risk in the ICU. It has been shown to be strongly associated with ischemic stroke, congestive heart failure and other high-risk cardiopulmonary conditions, as well as generally increased risk of mortality and increased length of hospital stay (Bosch et al., 2018; Klein Klouwenberg et al., 2016). In addition to the direct risk of patient harm, delayed recognition and intervention has risks for the quality of healthcare delivery and can potentially incur much higher costs from more invasive and later interventions that could be prevented through more proactive effort. Post-hoc identification of AF in the ICU typically relies on one of three data sources: structured diagnosis codes on admission or discharge, unstructured clinical notes, and diagnostic ECG reports taken at the bedside. However, each of these may lack both sensitivity and specificity for the diagnosis of AF, and only capture a brief snapshot of the overall burden of AF carried by a patient during their ICU stay. To that end, manual expert annotation is generally required to capture details such as paroxysmal AF and other shifts in rhythm over time. Because expert clinician and technician time is a limited, inelastic resource, this necessitates making a trade-off between the number of patients considered and number of data points to consider for each patient. Even then, many studies are restricted to large academic medical institutions that can scale up enough time and personnel.

In this paper, we propose a method for training automated AF classifiers on ICU ECG data without large-scale manual data annotation. We detail a method for generating weak labels on institutional data without additional metadata by pre-training models on existing external ECG databases. We then show that weakly supervised learning can be used in tandem with these weak labels when only a limited ground truth is available. Lastly, we explore combining labelled and weakly labelled data for training and find that fine-tuning does not provide any benefit over only using weak labels and weak supervision.

Generalizable Insights about Machine Learning in the Context of Healthcare

1. Many healthcare domains still lack a wealth of clean, labelled in-house data for ML applications.

2. Training with weakly labelled data can dramatically improve model performance when ground truth labels are limited.

3. Weakly supervised learning techniques can be used to further boost model performance and reduce the impact of label noise.

4. Only applying weak supervision may be superior to conducting additional fine-tuning in the absence of a sufficiently large or representative tuning dataset.

2. Related Work

2.1. Machine Learning on ECG Data

A large body of work has explored the application of machine learning techniques on ECG signals. Hong et al. (2020) identified over 100 studies from 2019-2020 alone with a clinically relevant objective. Also notable are the Computing in Cardiology competitions (Moody et al., 2001; Clifford et al., 2015, 2017) which aim to provide standard benchmarks for evaluating ML model performance for detecting or localizing various rhythms in ECGs, as well as other classification tasks. Recent work has increasingly focused on using deep learning for diagnostic and rhythm detection purposes (Goodfellow et al., 2018; Hannun et al., 2019; Strodthoff et al., 2020).

2.2. Machine Learning with Limited Labelled Data

Many studies have applied techniques to circumvent the need for large quantities of ground truth data in ML training (Cheplygina et al., 2019). Representation learning approaches seek to accomplish this through training a model to first understand the structure of the underlying data. This general feature extractor can then be fine-tuned on a significantly smaller ground truth to learn about the underlying classes. Unsupervised techniques such as direct input reconstruction are common for segmentation-based tasks on images (Hesamian et al., 2019), have also been applied to text (Huang et al., 2020) and physiological signals (Perslev et al., 2019). Self-supervised learning takes a similar approach, but favours learning representations through proxy tasks such as re-arranging permuted inputs (Taleb et al., 2020; Sarkar and Etemad, 2020), or via contrastive methods that try to group related samples together in an embedding space (Banville et al., 2020; Chaitanya et al., 2020).

Instead of assuming an overall lack of training labels, weakly supervised learning attempts to extract signal out of inherently noisy or coarse labels. These labels may originate from unreliable external metadata (Hu et al., 2020) or non-expert labels (Saab et al., 2020). Other techniques allow for hierarchical or finer-grained narrowing of labels for different spatial or temporal segments (Sudharshan et al., 2019; Shen et al., 2021; Isaev et al., 2020). In the absence of any external label sources, weak supervision can also be leveraged on predicted labels from lower-powered classifiers or collections of heuristics (Fries et al., 2019; Ratner et al., 2020).

3. Methods

A summary of the data and training pipeline is illustrated in Figure 1. We create or make use of three main datasets: a publicly available 12-lead ECG dataset, a small set of ground truth labelled in-house data and a much larger set of weakly labelled in-house data.

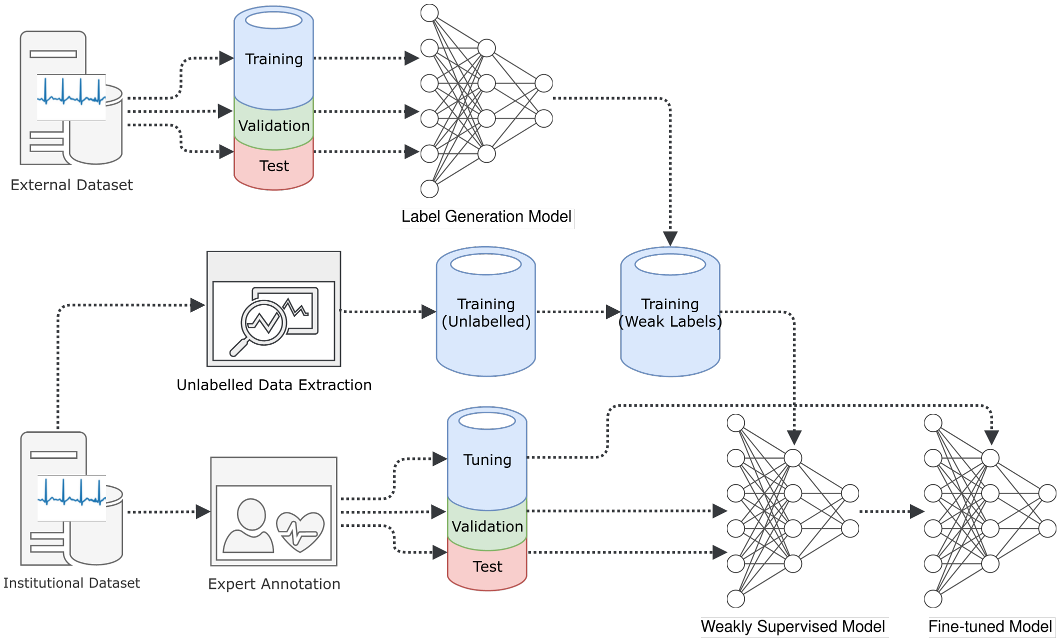


Figure 1: Summary of the data extraction, annotation and training process. We sample two disjoint sets of data from the patient cohort, only one of which receives human annotations (less than 1% of cohort). External data are used purely for training the weak label generation model. The weak label data is then used to train the the model that is validated, tested, and later fine tuned by expert annotated cohort.

3.1. Cohort

Out primary dataset was derived from a single-centre cohort at a tertiary North American academic health centre. This facility has a 33-bed mixed-use medical, surgical, neurological, and trauma ICU. Data were collected between 2015 to 2020 from GE Solar bedside monitors. Data were captured using Bedmaster software (Hill-Rom Holdings Inc, Chicago, Ill), which created a single file for each patient stay. Prior to analysis, each Bedmaster file was converted from its proprietary format to XML, and then to HDF5. The resulting dataset consists over 11TB of vital sign and waveform data, including mean arterial pressure, oxygen saturation (SpO2), and ECG leads I, II, III and V1 (each collected at 240Hz). To create our patient cohort, we matched clinical metadata from our EMR system to find 1,043 patients with unambiguous admission and discharge times (i.e. with no transfers between ICU beds during their stay).

3.2. Labelled Segment Extraction

From the primary cohort, we selected 655 patients and extracted one random contiguous 10 second segment from their record for clinical annotation. We defined 8 disjoint classes to use for annotation (Table 1). Each segment was annotated by two critical care physicians with extensive experience in ECG interpretation. Inter-annotator agreement (Cohen's Kappa) on

Table 1: Distribution of data segments in each ECG class label, extracted from the annotated subset of the internal cohort. Each segment is associated with one patient. No segments in the cohort were labelled as ventricular tachycardia/fibrillation or other tachycardia

Class	Count	Percentage
Sinus Rhythm	502	76.6%
Atrial Fibrillation/Atrial Flutter	89	13.6%
Pacemaker	12	1.8%
Bigemeny and Trigemeny	2	0.3%
Ventricular Tachycardia/Fibrillation	0	0%
Other Tachycardia	0	0%
Other Bradycardia	1	0.2%
Noise (non-diagnostic)	12	1.8%

the entire labelled set was 0.81. Ties were broken by a 3rd critical care physician annotator and then by consensus among all three physicians.

We set aside a random split of 70% of the data for baseline training and fine-tuning. The remaining data were split into 15% for validation and 15% for testing. All splits were done in a stratified fashion using label counts. Due to the low prevalence of the other classes, we only consider AF and Sinus Rhythm for our experiments.

3.3. Unlabelled Segment Extraction

For the unlabelled dataset, we extracted a further set of 100,722 10-second segments from the 1,043 patient cohort. For each patient, we performed a linear scan of the ECG signal until we found a continuous 10-second interval with no sensor drop-off. This interval was extracted as an ECG segment and incorporated into the unlabelled dataset. To reduce the risk of redundancy among segments from the same patient, we ensured adjacent ECG segments found during the linear scan were situated at least 1 hour apart in the patient's stay. Likewise, no identifiers (anonymized or otherwise) or demographic information was retained to ensure models could not directly associate segments from the same patient or between datasets. The mean and median segment counts for each patient were 101 and 57 respectively, with a minimum of 1 and artificially enforced maximum of 1000 segments. To create a more balanced dataset, a random set of 267 (1.75 $\times$ the IQR + the median) segments was sampled for each patient, for a grand total of 84,614 unlabelled segments.

We also employed the Chapman public 12-lead ECG dataset (Zheng et al., 2020) for our experiments. This consists of 12-lead ECGs from 10,645 patients (10 seconds each) extracted from Shaoxing People's Hospital in China. Per the authors' recommendations, we collapsed the 11 rhythm classes into 4 superclasses: Atrial Fibrillation (AFIB), Sinus Rhythm (SR), Sinus Bradycardia (SB) and Generalized Supraventricular Tachycardia (GSVT). We resampled the 500Hz signals to 240Hz to match our institutional dataset, and used splits of 70% for training, 10% for validation and 20% for testing.

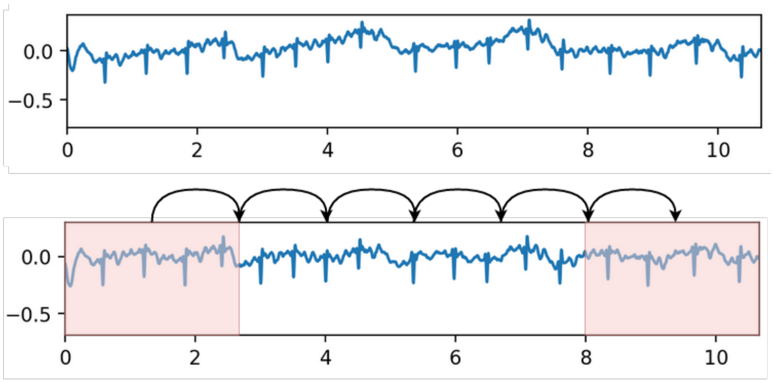


Figure 2: Preprocessing and windowing an ECG segment. The raw signal (top) is band-passed filtered and normalized. If the signal is in the training set, a single 2.5s random crop of it is taken for training the model. For validation and test, a sliding window of 2.5s with overlap of 1.25s is used (bottom) and the predictions are aggregated.

3.4. Preprocessing and Windowing

Only minimal processing was performed on the raw signals. Each 10 second sample was filtered using an order 3 Butterworth filter (Butterworth, 1930) with band 0.5-40Hz following Clifford et al. (2006). To ensure a consistent range of data for training, each lead of each segment was also independently min-max scaled. Per Strodthoff et al. (2020), we took a random crop of 2.5s from the signal while training and aggregate predictions from a sliding 2.5s window with 1.25s of overlap during validation and testing (Figure 2). This provided some limited train-time augmentation and smooths out per-label predictions during evaluation.

3.5. Model Architecture

We employ a fully-convolutional network derived from the model in Goodfellow et al. (2018). This model consists of 13 convolutional blocks and a final linear head. Each block is composed of a 1D convolution, batch normalization, ReLU and 30% dropout. The number of channels is gradually decreased from 256 to 64 while the per-block dilation is increased from 1 to 8. Blocks 6 and 11 also incorporate max-pooling with a kernel size and dilation of 2. To allow for input size invariance, global average pooling is applied before the final linear layer. A full architecture diagram may be found in Appendix A. All models were created using PyTorch (Paszke et al., 2019) and PyTorch Lightning (Falcon et al., 2020).

3.6. Baseline

We trained the model described in Section 3.5 from scratch on our set of labelled in-house data. Training was conducted over a maximum of 200 epochs with a batch size of 64, learning rate of 0.05, Adam optimizer (Kingma and Ba, 2017) and early stopping with a patient threshold of 10 epochs.

3.7. Training with Weak Labels

To create the weak labelling model, we trained a copy of the CNN classifier from Section 3.5 on the Chapman dataset. All hyperparameters were identical to Section 3.6, with the exception of a batch size of 128 to accommodate the larger training set. The resulting model achieved a AUC of 0.91 on our Chapman test set, roughly in line with Zheng et al.'s results.

Once trained, the model was then used to generate labels via running inference on our unlabelled data. We map the Chapman AFIB class to our AF class and the SR and SB classes to Sinus Rhythm. Samples with the Chapman GSVT class are discarded because there is no direct analogue in our classes. Under this scheme, we extracted 51,218 weak AF and 16,105 weak Sinus Rhythm labels for a total training set size of 67,323. We then trained a model directly on the weak generated labels, while using a subset of the annotated cohort for validation and testing.

3.8. Weakly Supervised Learning

We explored two complementary approaches for weakly supervised learning on the our large dataset of inferred model labels.

3.8.1. Confident Learning

Confident learning seeks to estimate the level of uncertainty in weak labels (Northcutt et al., 2021). This is accomplished through estimating the confident joint $C_{\widetilde{y},y^*}$ of noisy labels $\widetilde{y}$ and uncorrupted or clean labels y^* in the dataset. Much like a confusion matrix, this is derived from counting (dis)agreements between predicted class labels and given target labels. To create a set of noisy samples $\hat{X}_{\widetilde{y}=i,y^*=j}$, the conditional probability $\hat{p}(\widetilde{y}=j)$ of each sample being misclassified is calculated. To minimize the impact of class imbalance, samples are only included if $\hat{p}(\widetilde{y}=j)$ exceeds a threshold t_j based on the expected per-class confidence score.

Once noisy label candidates are found, the corresponding samples can be down-weighted or pruned from the dataset. In our experiments, we use iterative cleaning and re-training procedures provided via the cleanlab library (Northcutt et al., 2021). This is conducted every 3 epochs (see Figure 5 and 6 in Appendix B for a more in-depth exploration of different intervals), calibrated based on how often the validation loss hits an inflection point and begins increasing. To effectively prune samples from the loss function, we keep track of the indices of potentially noisy samples. Any predictions from those indices are then masked with zero when they appear in a batch.

3.8.2. Co-teaching

In contrast to confident learning, co-teaching (Han et al., 2018) relies on a pair of networks to share information about noisy samples and high confidence samples during the training process. Instead of backpropagating the loss signal to its own weights, each network will rank and sample low loss instances to pass to the other. This can be expressed in the following form:

$$L_f = L(\theta_f, X, y), L_g = L(\theta_g, X, y) \tag{1}$$

$$i_f = \text{argsort}(L_g)[0, ..., R(T) * |X|] \tag{2}$$

$$i_g = \text{argsort}(L_f)[0, ..., R(T) * |X|] \tag{3}$$

$${L_f}' = L(\theta_f, X[i_f], y[i_f]), {L_g}' = L(\theta_g, X[i_g], y[i_g]) \tag{4}$$

Where f and g are networks with parameters θ_f and θ_g, X is the input batch, y is the target labels and L is the loss function and $R(T)$ is the proportion of samples to propagate at epoch T. Han et al. demonstrate that this improves robustness and reduces the risk of large deep learning models overfitting on noisy labels. The dual network architecture allows for more heterogeneity and reduces the impact of outliers on training performance. The proportion of low loss instances is slowly annealed over the course of training, effectively pruning out noisy samples and providing a fine-tuning like approach for later epochs. For our experiments, we followed the equation in section 4.2 of Han et al. (2018) and used a sampling proportion $R(T)$ of 1. This was annealed to 0.75 using a τ of 0.25 over the course of 15 epochs.

3.8.3. Combined Learning Approaches

As confident learning and co-teaching address different parts of the training pipeline, we can combine them and evaluate if the performance of both approaches is superior in aggregate. The hyperparameters and training procedure are unchanged from Section 3.8.1 and 3.8.2.

3.9. Fine-tuning

In addition to the above, we also fine-tune each model trained with weak labels on our labelled training set (see Figure 1). This fine tuning follows the same training procedure as Section 3.8 with additional steps for gradually unfreezing the model. Firstly, the backbone of the network up to the final linear layer was frozen for the first 10 epochs. This reduces variance in the vital first few steps of fine tuning and prevents severe performance degradation. After the 10 epochs, the remainder of the network was unfrozen and training continued as before.

4. Results

We calculated the binary AUC and average precision for each training configuration on our labelled holdout test set. We also derive F1 and specificity scores, both calibrated using Youden's J statistic (Youden, 1950) on the validation dataset. Figure 3 and Table 2 summarize these results. Training directly on weak labels offered a significant increase in all metrics over the baseline. However, applying the label generation model directly onto our labelled data was also comparable to training on the weak labels that the model generated. Co-teaching exhibited the best overall performance with a marginal lead, but not a significant difference compared to working with weak labels directly. Surprisingly, confident learning did not provide any measurable advantage over training on weak labels directly and even adversely affected performance when applied in tandem with co-teaching.

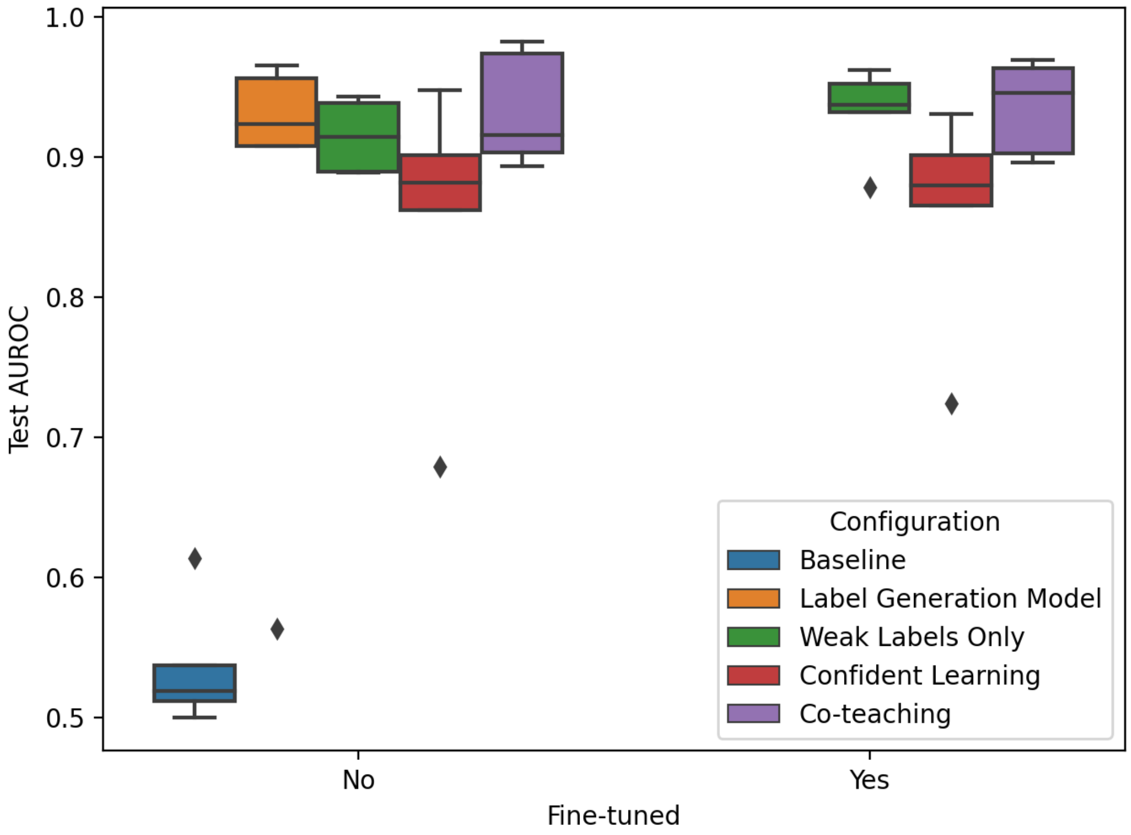


Figure 3: Boxplot comparison of AUROC for all configurations against the fully-supervised baseline. The "Weak Labels Only" legend refers a model that directly trained on the weak generated labels, while validated/tested using a subset of the annotated cohort.

Similarly, fine-tuning on our labelled training data actually incurred a performance deficit for all models. Here too, co-teaching suffers the least impact from fine-tuning and maintains a relatively low variance compared to the other techniques.

About 10-12% of training labels were identified as potentially noisy by the confident learning process at the first label-cleaning interval. This decreases to 2-3% after 15 epochs (5 rounds of cleaning). Although all configurations follow a similar monotonically decreasing trend, training with co-teaching appears to exhibit a higher variance in starting noise and a slower initial decrease until the 15 epoch mark (Figure 4).

Table 2: Comparison of average (and standard deviation) of curve-based (top) and thresholded (bottom) performance metrics over 10 trials for each training configuration.

Performance metric	AUROC		Average Precision	
Fine-tuning performed	No	Yes	No	Yes
Baseline	0.59 (0.09)		0.25 (0.06)	
Label Generation Model	0.90 (.01)		0.66 (.06)	
Weak Labels Only	0.91 (.03)	0.86 (.07)	0.63 (.10)	0.57 (.14)
Confident Learning	0.89 (.02)	0.82 (.07)	0.62 (.03)	0.52 (.13)
Co-teaching	0.91 (.01)	0.88 (.05)	0.61 (.06)	0.56 (.09)
Co-teaching + Confident Learning	0.88 (.01)	0.82 (.09)	0.38 (.08)	0.37 (.09)

Performance metric	F1 Score		Specificity	
Fine-tuning performed	No	Yes	No	Yes
Baseline	0.45 (.16)		0.56 (0.21)	
Label Generation Model	0.73 (.04)		0.81 (0.06)	
Weak Labels Only	0.75 (.11)	0.73 (.10)	0.84 (.03)	0.82 (.09)
Confident Learning	0.73 (.05)	0.70 (.09)	0.81 (.04)	0.84 (.07)
Co-teaching	0.78 (.03)	0.75 (.06)	0.82 (.03)	0.84 (.03)
Co-teaching + Confident Learning	0.74 (.04)	0.66 (.12)	0.67 (.13)	0.69 (.16)

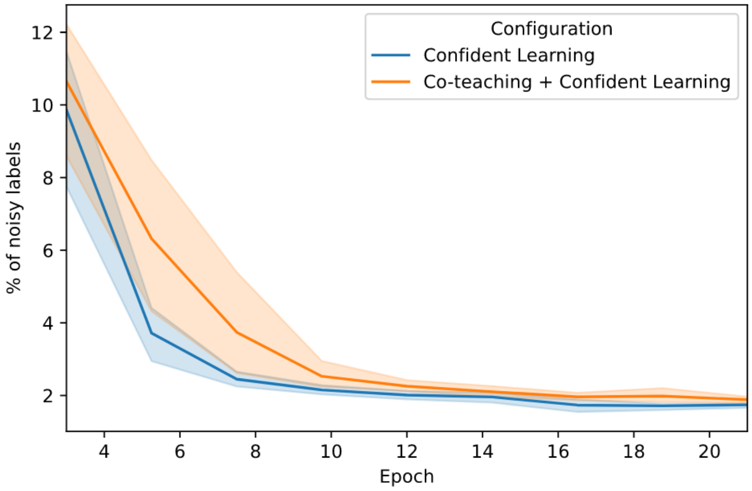


Figure 4: Trend in the estimated proportion of noisy labels from confident learning. The interval for iterative cleaning and re-training is 3 epochs. Note that no label cleaning occurs between epoch 0 and the first cleaning interval.

5. Discussion

In this study, we investigated and demonstrated that weak labels can help to bridge a labelling gap for institutions without the resources to create extensive manual annotation campaigns. Most notably, using weak labels is superior to a smaller set of clean ground truth labels even when weakly supervised learning techniques are not applied. Jiang et al.

(2020) found similar results when comparing inherent and synthetic label noise on natural images, but to our knowledge this has not been explored for physiological signals.

We also demonstrate that weakly supervised learning can be used to further build on the performance of using noisy labels directly. However, we do not find that weak supervision via label cleaning (confident learning) is beneficial. We hypothesize that the iterative pruning reduces the diversity of training samples available for a given class and allows the model to skip less confident examples. This effect can be seen in Appendix B, where both semi-frequent and extremely infrequent (i.e. equivalent to no confident learning) re-cleaning allow the model to re-gain or retain previously pruned samples respectively. In a similar vein, we counter-intuitively see a performance drop from fine-tuning on our labelled dataset. We believe this adds more evidence in support of prior work that has looked into the limitations of consensus-based hard labelling (Guan et al., 2018) and the advantages of having more labelled samples over more labels for each sample (Khetan et al., 2018).

The absence of copious annotated data is an endemic and well-described problem for machine learning in healthcare. From a technical perspective, there remains a wide solution space to explore in terms of approaches for generating weak labels, classification models and data sampling strategies. We propose two particular categories of weak labels sources to draw from. Intrinsic features are anything that can be extracted directly from the signal without out-of-band information. Existing heuristics based on gross features such as heart rate, as well as derived morphological features such as R-R intervals and heart rate variability are a promising source, as are learned feature representations from deep models. Conversely, extrinsic features include any additional clinical data that might be associated with a patient's ICU stay. For example, current clinical practice advocates for the use of antiarrhythmic drugs as a part of AF treatment; pharmacy records identifying patients receiving these medications could be used to derive labelling functions. Other sources include structured notes and labels that are both timestamped and commonly associated with AF (Ding et al., 2019), or other clinical conditions such as AF risk factors (age, history of hypertension). With weakly supervised learning, one can use this information to find rough ranges where a patient may have experienced AF without needing to pinpoint exact intervals on the signal to ensure accurate labels for model training.

5.1. Limitations

As a preliminary proof of concept, our study has a few limitations to consider. We focused on one condition (AF) in a single clinical setting (the ICU) and using one data modality (ECG). Our samples were derived from a smaller cohort and are likely less diverse than the entire spectrum of patients that present to our ICU. This diversity was only further reduced by filtering criteria such as excluding uncommon arrythmias or segments with more than 50% sensor drop-off. Therefore, our approach has only been thoroughly tested on data from one institution and may be influenced by specific attributes of our regional patient population as well as local practice patterns. Future work may explore cross-institution applications and whether the approach generalizes to disparate care systems as well.

Moreover, the value of weak labels from our label generation model can not be completely decoupled from the external dataset it was pre-trained on. Specifically, it is difficult to decouple uncertainty inherent in labels and label collection from the uncertainty that results from the stochastic nature of deep learning methods when evaluating results. In this paper, we aimed to address the former challenge. However, our preliminary experiments with varying levels of artificial noise on the Chapman dataset find that weak supervision can result in an over 0.1 average improvement in F1 compared to using weak labels with the label generation model when the proportion of noisy labels exceeds 25%.

5.2. Conclusion

In this paper, we evaluated the feasibility of leveraging external data and weakly supervised learning to improve AF detection for local ICU data. Many questions remain around the prevalence, burden and treatment of AF in the ICU. Our hope with this work is to provide techniques for accelerating research into this condition in the critical care context, with a focus on how it develops, and its downstream consequences, particularly as the overall burden of AF increases. This will be used to inform decisions about treatment best practices and answer long-standing clinical questions about the condition itself. Eventually, we hope to extend this beyond the domain of detection to predictive modelling, such that machine learning models may be used for early warning and monitoring directly. Lastly, we hope that our approach provides an avenue for more democratization of ML for healthcare in the ICU by allowing organizations to create competitive and useful models with tractable human and computational investment.

References

Hubert Banville, Omar Chehab, Aapo Hyvärinen, Denis-Alexander Engemann, and Alexandre Gramfort. Uncovering the structure of clinical EEG signals with self-supervised learning. *arXiv:2007.16104 [cs, eess, q-bio, stat]*, July 2020.

Nicholas A. Bosch, Jonathan Cimini, and Allan J. Walkey. Atrial Fibrillation in the ICU. *Chest*, 154(6):1424–1434, December 2018. ISSN 00123692. doi: 10.1016/j.chest.2018.03.040.

Stephen Butterworth. On the theory of filter amplifiers. *Wireless Engineer*, 7(6):536–541, 1930.

Krishna Chaitanya, Ertunc Erdil, Neerav Karani, and Ender Konukoglu. Contrastive learning of global and local features for medical image segmentation with limited annotations. *arXiv:2006.10511 [cs, eess, stat]*, October 2020.

Veronika Cheplygina, Marleen de Bruijne, and Josien P. W. Pluim. Not-so-supervised: A survey of semi-supervised, multi-instance, and transfer learning in medical image analysis. *Medical Image Analysis*, 54:280–296, May 2019. ISSN 1361-8415. doi: 10.1016/j.media.2019.03.009.

Gari D. Clifford, Francisco Azuaje, and Patrick Mcsharry. ECG statistics, noise, artifacts, and missing data. *Advanced methods and tools for ECG data analysis*, 6(1):18, 2006.

Gari D Clifford, Ikaro Silva, Benjamin Moody, Qiao Li, Danesh Kella, Abdullah Shahin, Tristan Kooistra, Diane Perry, and Roger G. Mark. The PhysioNet/Computing in Cardiology Challenge 2015: Reducing False Arrhythmia Alarms in the ICU. *Computing in cardiology*, 2015:273–276, September 2015. ISSN 2325-8861. doi: 10.1109/CIC.2015.7408639.

Gari D Clifford, Chengyu Liu, Benjamin Moody, Li-wei H. Lehman, Ikaro Silva, Qiao Li, A E Johnson, and Roger G. Mark. AF classification from a short single lead ECG recording: The PhysioNet/computing in cardiology challenge 2017. In *2017 Computing in Cardiology (CinC)*, pages 1–4, September 2017. doi: 10.22489/CinC.2017.065-469.

Eric Y. Ding, Daniella Albuquerque, Michael Winter, Sophia Binici, Jaclyn Piche, Syed Khairul Bashar, Ki Chon, Allan J. Walkey, and David D. McManus. Novel Method of Atrial Fibrillation Case Identification and Burden Estimation Using the MIMIC-III Electronic Health Data Set. *Journal of Intensive Care Medicine*, 34(10):851–857, October 2019. ISSN 1525-1489. doi: 10.1177/0885066619866172.

William Falcon, Jirka Borovec, Adrian Wälchli, Nic Eggert, Justus Schock, Jeremy Jordan, Nicki Skafte, Ir1dXD, Vadim Bereznyuk, Ethan Harris, Tullie Murrell, Peter Yu, Sebastian Præsius, Travis Addair, Jacob Zhong, Dmitry Lipin, So Uchida, Shreyas Bapat, Hendrik Schröter, Boris Dayma, Alexey Karnachev, Akshay Kulkarni, Shunta Komatsu, Martin.B, Jean-Baptiste SCHIRATTI, Hadrien Mary, Donal Byrne, Cristobal Eyzaguirre, cinjon, and Anton Bakhtin. PyTorchLightning/pytorch-lightning: 0.7.6 release. Zenodo, May 2020.

Jason A. Fries, Paroma Varma, Vincent S. Chen, Ke Xiao, Heliodoro Tejeda, Priyanka Saha, Jared Dunnmon, Henry Chubb, Shiraz Maskatia, Madalina Fiterau, Scott Delp, Euan Ashley, Christopher Ré, and James R. Priest. Weakly supervised classification of aortic valve malformations using unlabeled cardiac MRI sequences. *Nature Communications*, 10(1):3111, July 2019. ISSN 2041-1723. doi: 10.1038/s41467-019-11012-3.

Sebastian D. Goodfellow, Andrew Goodwin, Robert Greer, Peter C. Laussen, Mjaye Mazwi, and Danny Eytan. Towards Understanding ECG Rhythm Classification Using Convolutional Neural Networks and Attention Mappings. In *Machine Learning for Healthcare Conference*, pages 83–101, November 2018.

Melody Guan, Varun Gulshan, Andrew Dai, and Geoffrey Hinton. Who Said What: Modeling Individual Labelers Improves Classification. *Proceedings of the AAAI Conference on Artificial Intelligence*, 32(1), April 2018. ISSN 2374-3468.

Bo Han, Quanming Yao, Xingrui Yu, Gang Niu, Miao Xu, Weihua Hu, Ivor Tsang, and Masashi Sugiyama. Co-teaching: Robust training of deep neural networks with extremely noisy labels. *Advances in Neural Information Processing Systems*, 31, 2018.

Awni Y. Hannun, Pranav Rajpurkar, Masoumeh Haghpanahi, Geoffrey H. Tison, Codie Bourn, Mintu P. Turakhia, and Andrew Y. Ng. Cardiologist-level arrhythmia detection and classification in ambulatory electrocardiograms using a deep neural network. *Nature Medicine*, 25(1):65–69, January 2019. ISSN 1546-170X. doi: 10.1038/s41591-018-0268-3.

Mohammad Hesam Hesamian, Wenjing Jia, Xiangjian He, and Paul Kennedy. Deep Learning Techniques for Medical Image Segmentation: Achievements and Challenges. *Journal of Digital Imaging*, 32(4):582–596, August 2019. ISSN 1618-727X. doi: 10.1007/s10278-019-00227-x.

Gerhard Hindricks, Tatjana Potpara, Nikolaos Dagres, Elena Arbelo, Jeroen J Bax, Carina Blomström-Lundqvist, Giuseppe Boriani, Manuel Castella, Gheorghe-Andrei Dan, Polychronis E Dilaveris, Laurent Fauchier, Gerasimos Filippatos, Jonathan M Kalman, Mark La Meir, Deirdre A Lane, Jean-Pierre Lebeau, Maddalena Lettino, Gregory Y H Lip, Fausto J Pinto, G Neil Thomas, Marco Valgimigli, Isabelle C Van Gelder, Bart P Van Putte, Caroline L Watkins, and ESC Scientific Document Group. 2020 ESC Guidelines for the diagnosis and management of atrial fibrillation developed in collaboration with the European Association for Cardio-Thoracic Surgery (EACTS): The Task Force for the diagnosis and management of atrial fibrillation of the European Society of Cardiology (ESC) Developed with the special contribution of the European Heart Rhythm Association (EHRA) of the ESC. *European Heart Journal*, 42(5):373–498, February 2021. ISSN 0195-668X. doi: 10.1093/eurheartj/ehaa612.

Shenda Hong, Yuxi Zhou, Junyuan Shang, Cao Xiao, and Jimeng Sun. Opportunities and challenges of deep learning methods for electrocardiogram data: A systematic review. *Computers in Biology and Medicine*, 122:103801, July 2020. ISSN 0010-4825. doi: 10.1016/j.compbiomed.2020.103801.

Szu-Yen Hu, Shuhang Wang, Wei-Hung Weng, JingChao Wang, XiaoHong Wang, Arinc Ozturk, Quan Li, Viksit Kumar, and Anthony E. Samir. Self-Supervised Pretraining with DICOM metadata in Ultrasound Imaging. In *Machine Learning for Healthcare Conference*, pages 732–749. PMLR, September 2020.

Kexin Huang, Jaan Altosaar, and Rajesh Ranganath. ClinicalBERT: Modeling Clinical Notes and Predicting Hospital Readmission. *arXiv:1904.05342 [cs]*, November 2020.

Dmitry Yu Isaev, Dmitry Tchapyjnikov, C. Michael Cotten, David Tanaka, Natalia Martinez, Martin Bertran, Guillermo Sapiro, and David Carlson. Attention-Based Network for Weak Labels in Neonatal Seizure Detection. In *Machine Learning for Healthcare Conference*, pages 479–507. PMLR, September 2020.

Lu Jiang, Di Huang, Mason Liu, and Weilong Yang. Beyond Synthetic Noise: Deep Learning on Controlled Noisy Labels. In *International Conference on Machine Learning*, pages 4804–4815. PMLR, November 2020.

Ashish Khetan, Zachary C. Lipton, and Anima Anandkumar. Learning From Noisy Singly-labeled Data. *arXiv:1712.04577 [cs]*, May 2018.

Diederik P. Kingma and Jimmy Ba. Adam: A Method for Stochastic Optimization. *arXiv:1412.6980 [cs]*, January 2017.

Peter M. C. Klein Klouwenberg, Jos F. Frencken, Sanne Kuipers, David S. Y. Ong, Linda M. Peelen, Lonneke A. van Vught, Marcus J. Schultz, Tom van der Poll, Marc J. Bonten, and

Olaf L. Cremer. Incidence, Predictors, and Outcomes of New-Onset Atrial Fibrillation in Critically Ill Patients with Sepsis. A Cohort Study. *American Journal of Respiratory and Critical Care Medicine*, 195(2):205–211, July 2016. ISSN 1073-449X. doi: 10.1164/rccm.201603-0618OC.

David M. Maslove, Francois Lamontagne, John C. Marshall, and Daren K. Heyland. A path to precision in the ICU. *Critical Care*, 21(1):79, April 2017. ISSN 1364-8535. doi: 10.1186/s13054-017-1653-x.

G. Moody, A. Goldberger, S. McClennen, and S. Swiryn. Predicting the onset of paroxysmal atrial fibrillation: The Computers in Cardiology Challenge 2001. In *Computers in Cardiology 2001. Vol.28 (Cat. No.01CH37287)*, pages 113–116, September 2001.

Curtis G. Northcutt, Lu Jiang, and Isaac L. Chuang. Confident Learning: Estimating Uncertainty in Dataset Labels. *arXiv:1911.00068 [cs, stat]*, February 2021.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Köpf, Edward Yang, Zach DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. PyTorch: An Imperative Style, High-Performance Deep Learning Library. *arXiv:1912.01703 [cs, stat]*, December 2019.

Mathias Perslev, Michael Jensen, Sune Darkner, Poul Jø rgen Jennum, and Christian Igel. U-Time: A Fully Convolutional Network for Time Series Segmentation Applied to Sleep Staging. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d\textquotesingle Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32*, pages 4417–4428. Curran Associates, Inc., 2019.

Alexander Ratner, Stephen H. Bach, Henry Ehrenberg, Jason Fries, Sen Wu, and Christopher Ré. Snorkel: Rapid training data creation with weak supervision. *The Vldb Journal*, 29(2):709–730, 2020. ISSN 1066-8888. doi: 10.1007/s00778-019-00552-1.

Khaled Saab, Jared Dunnmon, Christopher Ré, Daniel Rubin, and Christopher Lee-Messer. Weak supervision as an efficient approach for automated seizure detection in electroencephalography. *npj Digital Medicine*, 3(1):1–12, April 2020. ISSN 2398-6352. doi: 10.1038/s41746-020-0264-0.

Pritam Sarkar and Ali Etemad. Self-supervised ECG Representation Learning for Emotion Recognition. *arXiv:2002.03898 [cs, eess, stat]*, February 2020.

Philippe Seguin and Yoann Launey. Atrial fibrillation is not just an artefact in the ICU. *Critical Care*, 14(4):182, July 2010. ISSN 1364-8535. doi: 10.1186/cc9093.

Yiqiu Shen, Nan Wu, Jason Phang, Jungkyu Park, Kangning Liu, Sudarshini Tyagi, Laura Heacock, S. Gene Kim, Linda Moy, Kyunghyun Cho, and Krzysztof J. Geras. An interpretable classifier for high-resolution breast cancer screening images utilizing weakly supervised localization. *Medical Image Analysis*, 68:101908, February 2021. ISSN 1361-8415. doi: 10.1016/j.media.2020.101908.

Nils Strodthoff, Patrick Wagner, Tobias Schaeffter, and Wojciech Samek. Deep Learning for ECG Analysis: Benchmarks and Insights from PTB-XL. *arXiv:2004.13701 [cs, stat]*, April 2020.

P. J. Sudharshan, Caroline Petitjean, Fabio Spanhol, Luiz Eduardo Oliveira, Laurent Heutte, and Paul Honeine. Multiple instance learning for histopathological breast cancer image classification. *Expert Systems with Applications*, 117:103–111, March 2019. ISSN 0957-4174. doi: 10.1016/j.eswa.2018.09.049.

Aiham Taleb, Christoph Lippert, Tassilo Klein, and Moin Nabi. Multimodal Self-Supervised Learning for Medical Image Analysis. *arXiv:1912.05396 [cs]*, October 2020.

W. J. Youden. Index for rating diagnostic tests. *Cancer*, 3(1):32–35, 1950. ISSN 1097-0142. doi: 10.1002/1097-0142(1950)3:1⟨32::AID-CNCR2820030106⟩3.0.CO;2-3.

Jianwei Zheng, Jianming Zhang, Sidy Danioko, Hai Yao, Hangyuan Guo, and Cyril Rakovski. A 12-lead electrocardiogram database for arrhythmia research covering more than 10,000 patients. *Scientific Data*, 7, February 2020. ISSN 2052-4463. doi: 10.1038/s41597-020-0386-x.

Appendix A.

A.1. Full Model Architecture

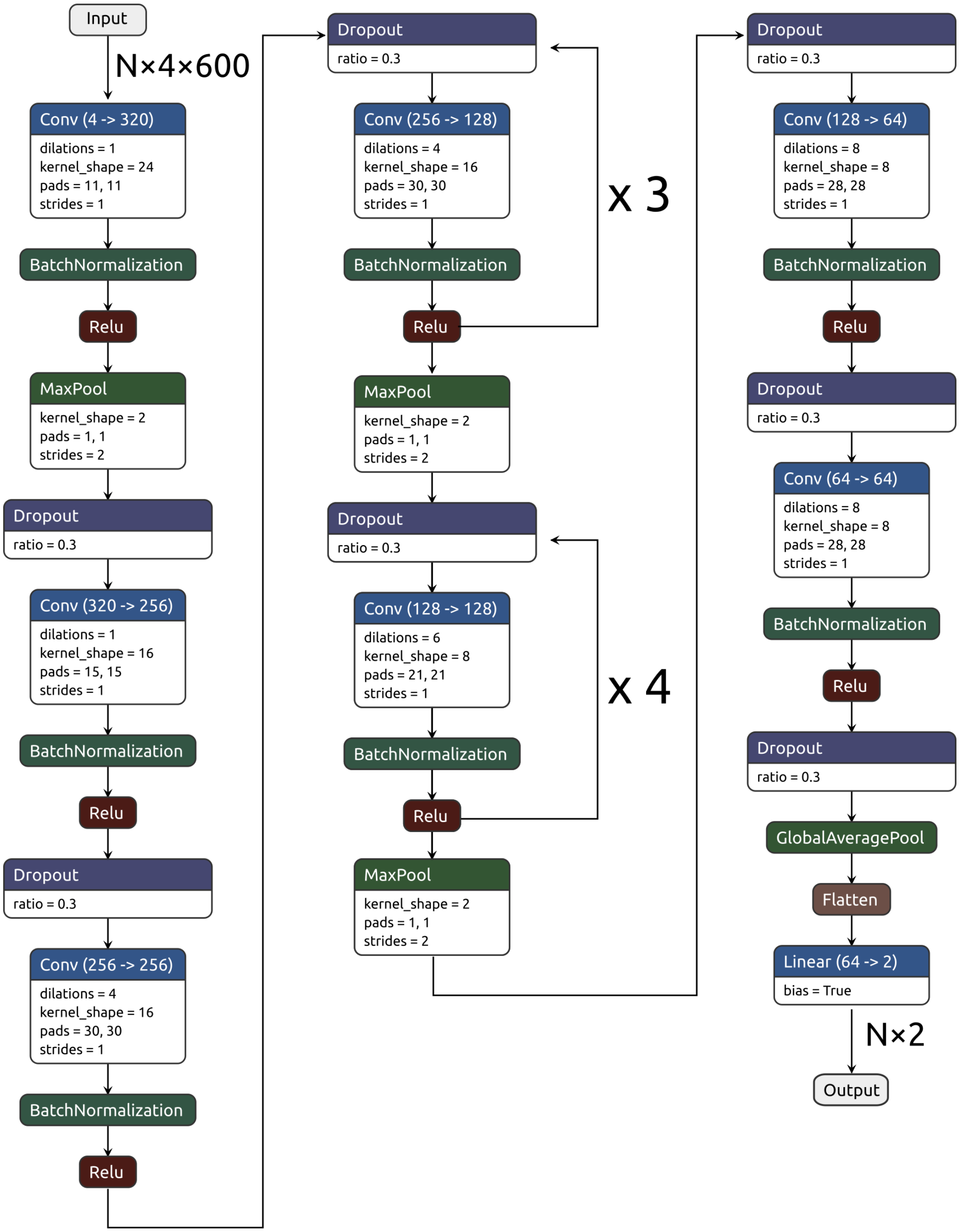
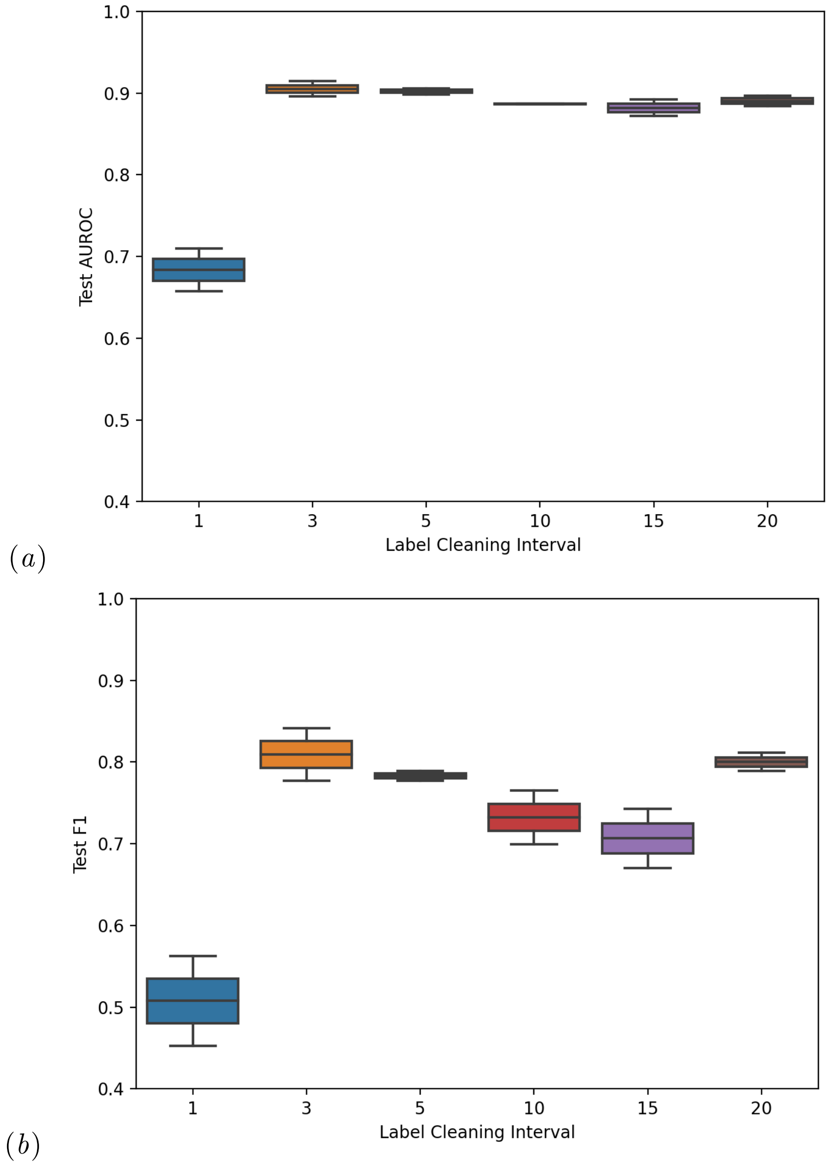


Figure 5: Deep learning model architecture used for training and label generation. The model takes in input of batch size N, 4 channels (leads) and arbitrary length (shown here as 2.5s $\times$ 240Hz $=$ 600 samples).

Appendix B.

B.1. Effect of Varying Cleaning Interval



(*a*)

(*b*)

Figure 6: AUROC (a) and F1 Score (b) performance metric comparison in confident learning with differing label cleaning intervals (number of epochs). A cleaning interval of 3 epochs shows the highest performance.