Automated Identification of Label Errors in Large Electrocardiogram Datasets

Peter Doggart[1,2], Alan Kennedy[1], Emily Foreman[1], Dewar Finlay[2], Raymond Bond[2]

[1] PulseAI Ltd, Belfast, United Kingdom
[2] Ulster Univeristy, Belfast, United Kingdom

Abstract

Background: *Training and testing Deep Neural Networks (DNNs) for automated electrocardiogram (ECG) interpretation requires large datasets. These datasets are commonly extracted at scale from Electronic Health Records (EHRs). Typically, a single physician over-reads the machine generated interpretation as part of standard care. Incorrect interpretation of the ECG occurs frequently, reducing the quality of the labels.*

Method: *We trained a DNN to identify seven ECG rhythms based on morphology; Sinus Rhythm, Junctional Rhythm, Ectopic Atrial Rhythm, Atrial Flutter, Atrial Fibrillation, Ventricular Rhythm and Pacemaker. The DNN was trained on a dataset of 368,202 ECGs taken from a proprietary database. We then applied confident learning techniques using the DNN to identify label errors in the Physionet PTB-XL database, which is publicly available.*

Results: *The confident learning algorithm identified 515 potential rhythm label errors in the 21,837 ECGs in PTB-XL database (2.36%). The labels were sorted by the likelihood of label error based on the self-confidence score, and the top 200 ECGs were manually reviewed. Of these 200 ECGs, 158 were found to be incorrectly labelled (79%). Confident learning successfully corrected the label in 156 cases (78%). The estimated labelling error rate for ECG rhythm in the PTB-XL database is 1.86%.*

1. Introduction

Large labelled datasets are critical to the success of all supervised machine learning techniques, regardless of the field of study. However, the process used to extract and construct these datasets often includes some form of automated labelling, which are inherently error-prone [1]. It has been shown that across the 10 most commonly-used computer vision, natural language and audio datasets, that label errors are numerous and widespread, with an average label error rate of at least 3.3% [2].

Large electrocardiogram (ECG) databases are commonly extracted from Electronic Health Records (EHRs). The primary source of ECG interpretation is often provided by automated algorithms, which are then over-read by a physician before being stored. However, these automated algorithms perform poorly, with overall classification accuracy reported as low as 58.9% [3].

We hypothesized that this poor automated interpretation performance is not always corrected by the over-reading physician, which leads to errors in ECG diagnoses recorded in the EHRs. These misdiagnoses subsequently become label errors in datasets which are used to train and test supervised classification models.

In this paper, we analyse the PhysioNet PTB-XL dataset [4, 5] to identify labelling errors in 12-lead ECG rhythm.

2. Materials and Methods

2.1. Datasets

To train our Deep Convolutional Neural Network (DCNN) we extracted 368,202 electrocardiograms (ECGs) from the proprietary PulseAI worldwide ECG database. This database contains labelled ECGs from over 1 million patients from 7 countries. The ECGs were labelled as part of standard clinical care, with a cardiologist or emergency medicine physician over-reading the automated ECG machine interpretation. This dataset was split into training (75%) and validation (25%) sets with stratification. The distribution of class labels is shown in Table 1.

During training, the majority class (Sinus Rhythm) was blind undersampled without replacement to contain

Label	Training	Validation
Sinus Rhythm	253,590	84,577
Junctional Rhythm	955	298
Ectopic Atrial Rhythm	1,200	392
Atrial Flutter	2,507	844
Atrial Fibrillation	12,633	4,154
Ventricular Rhythm	149	52
Pacemaker	5,117	1,734

Table 1. Class distribution in training and validation sets.

149,000 ECGs, which is 1000 times the number of examples in the lowest prevalence class. All the minor classes were then blind oversampled with replacement to create a balanced training dataset containing 149,000 ECGs in each class.

The PhysioNet PTB-XL dataset [4, 5] was downloaded from the PhysioNet/Computing in Cardiology Challenge 2021 [6] and the provided SNOMED-CT diagnostic codes were mapped to the classes shown in Table 1. All non-rhythm classes were discarded. No other modifications were made to the labelling of this dataset.

2.2. Deep Convolutional Neural Network

Our DCNN architecture is similar to the 13 layer architecture described by Goodfellow et al. [7] but adapted for 12-lead ECG input, rather than single lead as described. Each block is composed of 1D convolution, batch normalization, ReLU activation, 1D max pooling (excluding blocks 2, 4, 5, 7, 8, 10), and dropout (30% rate). The final layers of the network were composed of two dense layers with ReLU activation, followed by a dense layer with softmax activation for classification. The final network contained 5,476,103 parameters.

The network takes 12-lead ECG recordings of 10 seconds in length as input, sampled at 250Hz. All ECGs were resampled to 250Hz prior to training. The network was trained with binary cross entropy loss using an Adam optimizer with an initial learning rate of 0.0001. The learning rate was reduced by a factor of 0.1 on validation loss plateau until the network was fully converged.

2.3. Confident Learning

To identify label errors in the Physionet PTB-XL dataset, we utilize confident learning tools provided by the cleanlab python package [8]. Confident learning is based on the principles of pruning noisy data, counting to estimate noise, and ranking examples to train with confidence.

We computed the class probabilities for every ECG in the PTB-XL dataset using our DCNN and then provided these probabilities, along with the dataset labels, to cleanlab. These probabilities are all out-of-sample, as this data was held-out during training. Confident learning estimates the joint distribution of given, noisy labels and latent (unknown) uncorrupted labels to fully characterize class-conditional label noise. We then used this to find and extract the noisy examples with label issues.

Identified label issues were reviewed internally by two experienced ECG interpreters, reviewing the reference and cleanlab identified labels. Any ECGs where consensus could not be reached were subsequently reviewed by an external cardiologist.

3. Results

Confident learning identified 515 potential rhythm label errors from the 21,837 ECGs in the PTB-XL dataset (2.36%). The top 200 ECGs sorted by label self-confidence score were selected for manual review. Of these 200, an internal consensus was reached on 190 cases (95%), with 10 sent to an external cardiologist for review (5%).

There are two areas of interest in the results of this study. The main focus is quantifying the degree of label error in the PTB-XL dataset. However, we also assess the performance of the confident learning tools to automatically reclassify incorrectly labelled examples in the dataset. The results are shown in Table 2 and Table 3.

	Confident Learning		
Reference	**Correct**	**Incorrect**	**Total**
Correct	8	34	42
Incorrect	118	4	122
No Rhythm	30	6	36
Total	156	44	**200**

Table 2. Confusion matrix showing the breakdown of reference and confident learning labels after human review. Correct-correct occurs when PTB-XL contains more than one rhythm label, the second of which is incorrect.

Label	**PTB-XL**	**Human Review**	**Change**
Sinus Rhythm	56	59	+3
Junctional Rhythm	0	2	+2
Ectopic Atrial Rhythm	0	0	0
Other SVT *	0	4	+4
Atrial Flutter	2	90	+88
Atrial Fibrillation	124	40	-84
Ventricular Rhythm	0	0	0
Pacemaker	18	5	-13

Table 3. Distribution of class label changes in the 200 reviewed ECGs. *Other SVT refers to other supraventricular tachycardias that were not present in the DNN classification.

4. Discussion

The principal findings of this study are that: 1) Label errors are as prevalent in ECG datasets as those reported in other fields; 2) Confident learning tools can be successfully used to find label errors in large ECG datasets; 3) We can use those same tools to automatically correct the majority of label errors.

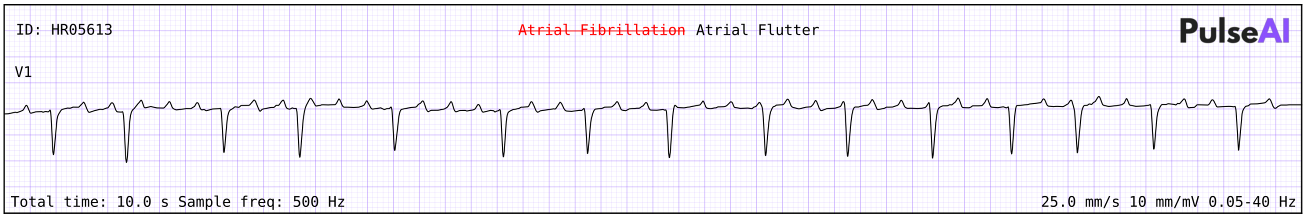
ID: HR05613
~~Atrial Fibrillation~~ Atrial Flutter
PulseAI
V1
Total time: 10.0 s Sample freq: 500 Hz
25.0 mm/s 10 mm/mV 0.05-40 Hz

Figure 1. An ECG rhythm strip (V1) from PTB-XL HR05613 showing clearly defined atrial flutter waves and variable ventricular conduction rather than atrial fibrillation as labelled.

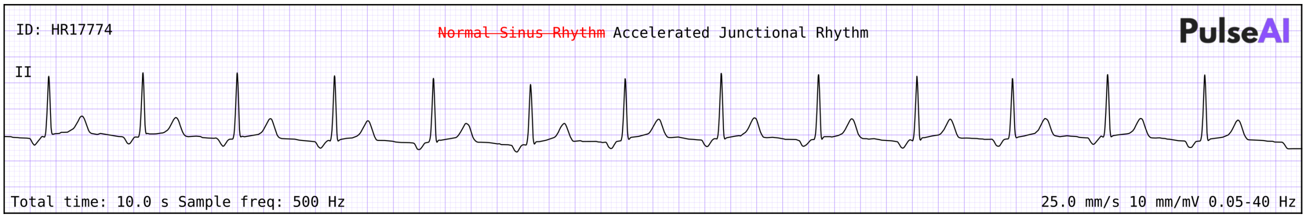
ID: HR17774
~~Normal Sinus Rhythm~~ Accelerated Junctional Rhythm
PulseAI
II
Total time: 10.0 s Sample freq: 500 Hz
25.0 mm/s 10 mm/mV 0.05-40 Hz

Figure 2. An ECG rhythm strip (II) from PTB-XL HR17774 showing inverted retrograde p-waves prior to the QRS complex, indicating accelerated junctional rhythm rather than normal sinus rhythm as labelled.

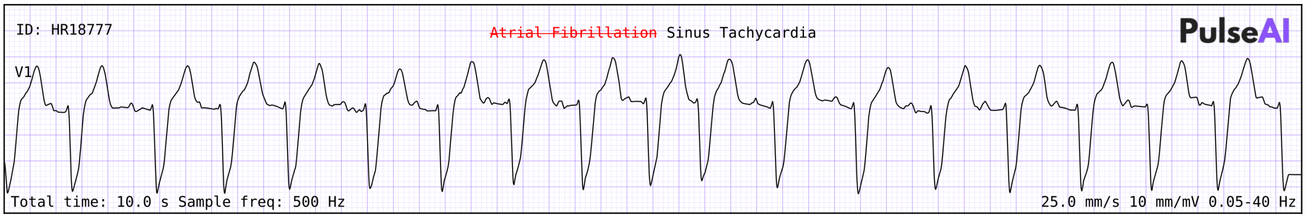
ID: HR18777
~~Atrial Fibrillation~~ Sinus Tachycardia
PulseAI
V1
Total time: 10.0 s Sample freq: 500 Hz
25.0 mm/s 10 mm/mV 0.05-40 Hz

Figure 3. An ECG rhythm strip (V1) from PTB-XL HR18777 showing normal p-wave and AV node conduction, indicating this is not atrial fibrillation but instead is sinus tachycardia with left bundle branch block.

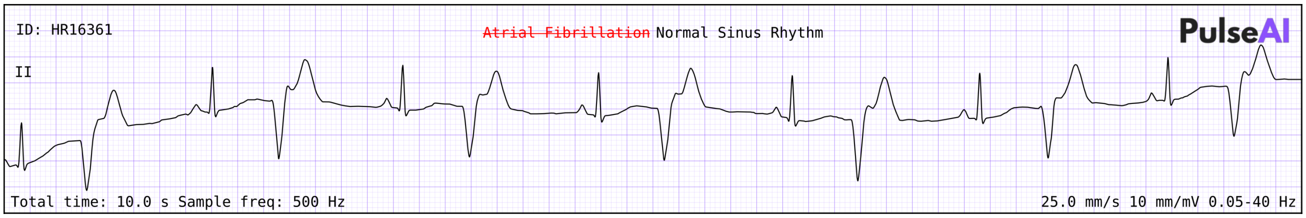
ID: HR16361
~~Atrial Fibrillation~~ Normal Sinus Rhythm
PulseAI
II
Total time: 10.0 s Sample freq: 500 Hz
25.0 mm/s 10 mm/mV 0.05-40 Hz

Figure 4. An ECG rhythm strip (II) from PTB-XL HR16361 showing normal sinus rhythm with ventricular bigeminy which is mislabelled as atrial fibrillation.

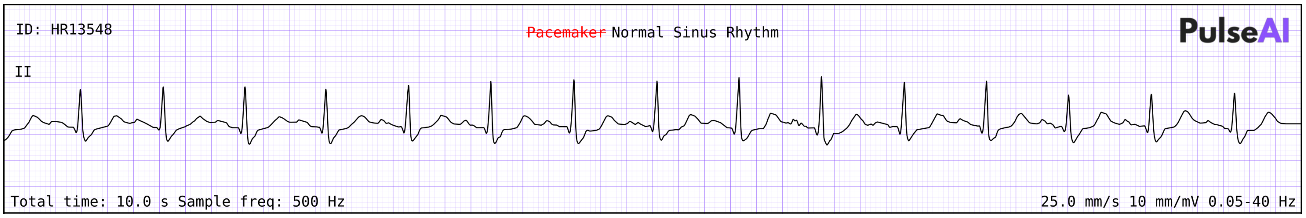
ID: HR13548
~~Pacemaker~~ Normal Sinus Rhythm
PulseAI
II
Total time: 10.0 s Sample freq: 500 Hz
25.0 mm/s 10 mm/mV 0.05-40 Hz

Figure 5. An ECG rhythm strip (II) from PTB-XL HR13548 showing no signs of any pacemaker activity despite containing a pacemaker label.

In this study, we applied confident learning techniques to the Physionet PTB-XL dataset in order to find labelling errors. We found that 158 out of 200 reviewed ECGs (79%) were labelled incorrectly or did not contain a rhythm label at all. This shows that the applied method is capable of identifying labelling errors with a high degree of accuracy. If we assume the same performance across all 515 potential labelling errors, then the estimated label error in rhythm interpretation alone is 1.86%. We would expect this to increase significantly if other, more subtle, ECG abnormalities were included.

The results also show that the same tool was able to select the correct label in 156 out of 200 cases (78%). This gives a high degree of confidence that even though not perfect, applying such methods automatically, without review, would still likely improve overall label quality.

We also reported in Table 3 the total count of each rhythm label in the 200 reviewed ECGs before and after human review. Although this crude measure only shows net changes, it is clear that the largest mislabelled groups are atrial flutter and atrial fibrillation. Misclassifications between these two groups are commonly reported in the literature [9].

It might be assumed by a naïve reader that this mislabelling must be caused by difficult to interpret ECGs. However, that is not the case. To demonstrate this, we present Figures 1-5 which show a single lead rhythm strip extracted from five ECGs in the reviewed set. The figures each show the original dataset label, as well as the automatically corrected label. None of these examples required manual human correction and would be easily classified by anyone with ECG training.

5. Conclusion

Correctly labelled ECG data is important for DNN development and performance reporting, especially for low prevalence classes, which can be significantly impacted by noisy labels. In this study, we demonstrated confident learning techniques can be applied to automatically identify and correct labelling errors in ECG datasets. We also estimated the overall labelling error rate for rhythm classification in the PTB-XL dataset at 1.86%. In future studies, we plan to extend these techniques to full 12-lead ECG interpretation classes in order to get a better picture of overall label quality in this and other publicly available data.

References

[1] Sambasivan N, Kapania S, Highfill H, Akrong D, Paritosh P, Aroyo LM. Everyone wants to do the model work, not the data work: Data cascades in high-stakes ai. In Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems. 2021; 1–15.

[2] Northcutt CG, Athalye A, Mueller J. Pervasive label errors in test sets destabilize machine learning benchmarks. arXiv preprint arXiv210314749 2021;.

[3] Smith SW, Walsh B, Grauer K, Wang K, Rapin J, Li J, Fennell W, Taboulet P. A deep neural network learning algorithm outperforms a conventional algorithm for emergency department electrocardiogram interpretation. Journal of Electrocardiology 2019;52:88–95.

[4] Goldberger AL, Amaral LA, Glass L, Hausdorff JM, Ivanov PC, Mark RG, Mietus JE, Moody GB, Peng CK, Stanley HE. Physiobank, physiotoolkit, and physionet: Components of a new research resource for complex physiologic signals. Circulation 2000;101(23):e215–e220.

[5] Wagner P, Strodthoff N, Bousseljot RD, Kreiseler D, Lunze FI, Samek W, Schaeffter T. Ptb-xl, a large publicly available electrocardiography dataset. Scientific Data 2020;7(1):1–15.

[6] Reyna MA, Sadr N, Alday EAP, Gu A, Shah AJ, Robichaux C, Rad AB, Elola A, Seyedi S, Ansari S, et al. Will two do? varying dimensions in electrocardiography: The physionet/computing in cardiology challenge 2021. In 2021 Computing in Cardiology (CinC), volume 48. IEEE, 2021; 1–4.

[7] Goodfellow SD, Goodwin A, Greer R, Laussen PC, Mazwi M, Eytan D. Towards understanding ecg rhythm classification using convolutional neural networks and attention mappings. In Machine Learning for Healthcare Conference. PMLR, 2018; 83–101.

[8] Northcutt C, Jiang L, Chuang I. Confident learning: Estimating uncertainty in dataset labels. Journal of Artificial Intelligence Research 2021;70:1373–1411.

[9] Lindow T, Kron J, Thulesius H, Ljungström E, Pahlm O. Erroneous computer-based interpretations of atrial fibrillation and atrial flutter in a swedish primary health care setting. Scandinavian Journal of Primary Health Care 2019; 37(4):426–433.

Address for correspondence:

Peter Doggart
PulseAI Ltd, 58 Howard Street, Belfast, BT1 6PL, UK
peter.doggart@pulseai.io