Robustness of Residual Network in Predicting PR Interval Trained Using Noisy Labels

Loc Cao[†1*], Hamid Ghanbari[†2*], Negar Farzaneh[†1], Kevin R Ward[†1], Sardar Ansari[†1]

[1]Department of Emergency Medicine, University of Michigan, Ann Arbor, MI, United States
[2]Division of Internal Medicine, Section of Cardiovascular Medicine, University of Michigan, Ann Arbor, MI, United States
[†]Weil Institute for Critical Care Research and Innovation, University of Michigan, Ann Arbor, MI, United States
[*]Equal contributions

Abstract

The PR interval represents the time required from the electrical impulse to advance from the atrium to AV node and His-Purkinje system until the ventricular myocardium begins to depolarize. PR interval prolongation has been associated with significant increases in atrial fibrillation, heart failure and mortality.

Over the past years, multiple deep learning models have been proposed to interpret electrocardiogram (ECG) signals. Despite initial success, these models are often trained and validated using datasets that contain partially incorrect labels. These "noisy" labels exist because of the way the annotated data was collected and pose challenges for model training and validation.

As a result, a residual neural network (ResNet), trained on noisy data, was proposed to estimate PR intervals. In addition, an electrophysiologist performed a blinded manual adjudication on a stratified sample to validate the accuracy of both the model and the noisy labels.

The conclusion is that a ResNet trained on noisy data can correctly estimate PR intervals and outperforms the noisy labels it was trained on.

1. Introduction

One of the main challenges with training any machine learning model is the cost related to obtaining the annotated data. Fortunately, since ECG is part of the routine patient care, the ECG signal is widely collected. One of the most common ECG collection systems is GE MUSE (GE Healthcare, Chicago, Illinois) which uses the MarquetteTM 12SL analysis program to generate automated annotations before those annotations are overwritten by providers if necessary. However, despite being inaccurate, a small percentage of these automatic annotations fail to get corrected during the process, leading to partially incorrect (i.e "noisy") labels.

A model was trained using 12-lead ECGs and their noisy labels to estimate the PR interval values. An electrophysiologist then performed a blinded manual adjudication to quantify its accuracy.

2. Methods

2.1. MUSE Dataset

The 12-lead ECGs were obtained from the Section of Electrophysiology at the University of Michigan. A total of 1,464,268 ECG signals were collected from 447,270 patients across Michigan Medicine (the University of Michigan Health System) from 1990 to 2018. The signals were sampled at 250Hz or 500Hz, and then transmitted into the MUSE system, where a clinician reviewed the tracings, and modified the automatic annotations as needed before releasing the information into the electronic health record. Almost 45% of all PR values collected were modified by clinicians during this process.

All 10-second long 12-lead ECGs were resampled at 250Hz. By default, only 8 leads (I, II, V1, V2, V3, V4, V5, and V6) were collected by the MUSE system, and the last 4 (III, aVR, aVL, and aVF) were generated using lead I and lead II. The MUSE software also applied a low-pass filter, a high-pass filter and a notch filter to the raw ECGs to remove baseline wanderer, muscle artifacts, power-line interference and other miscellaneous noises [1] though specific filter configurations were customized by technicians per batch of analyzed signals.

All the ECGs with invalid PR values (Figure 1) or with specific arrhythmias and conditions that render the PR values meaningless from a clinical perspective were excluded. For instance, for patients with atrial fibrillation, the PR value cannot be calculated, and should be removed from model training and testing. The arrhythmia labels were generated using a combination of three approaches. The majority of semi-structured diagnosis statements, i.e fragment of text that contains a single or multiple diagnoses [2], embedded in the dataset were assigned to a Unified Medical Language System Concept Unique Identifiers (CUI) [3] and the corresponding Systematized Nomenclature of Medicine Clinical Terms (SNOMED CT) [2]. All the segments of diagnosis statements that did not have a corresponding CUI were then split into two groups sorted by frequency. 91 of these statements appeared in the dataset more than 50 times and were then manually mapped while a word search filter was used to map any of the 71,713 remaining diagnoses. The word search filter helped eliminate any additional cases of atrial fibrillation, atrial flutter, ventricular fibrillation, high degree of atrio-ventricular or sinoatrial block, supraventricular tachycardia, ventricular tachycardia, atrio-ventricular dissociation and any paced signals which both the CUI and manual mapping failed to detect. Finally, since ECG interpretation is drastically different between adult and pediatric patients, the analysis was limited to ECGs from adult patients only (≥ 18 years old). All patients with missing age due to an irretrievable system error in 2013 were removed. In total, 384,349 ECGs were excluded from the original data, the breakdown of which is presented in Figure 2, resulting in a cohort of 1,079,919 ECGs and 399,529 patients.

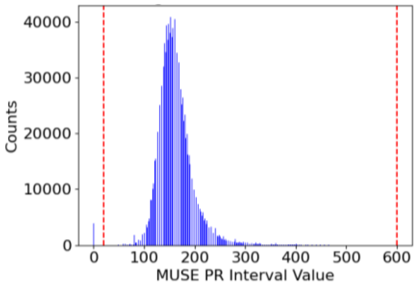

Figure 1: Distribution of PR intervals with clinically sensible upper and lower bounds. The plot does not show 167,434 ECGs with missing PR values, and 1,778 with PR values that exceed 600ms.

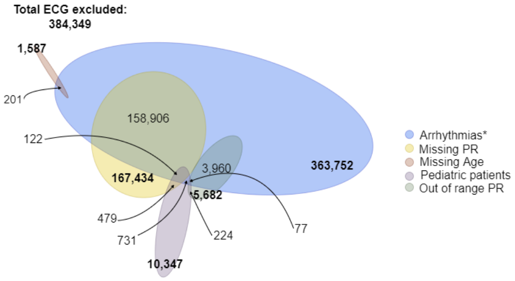

Figure 2: Diagram illustrating the breakdown of the cohort clean-up. About 95% of exclusions are caused by different arrhythmias. The diagram depicts that the main cause for missing or out-of-range PR is due to different arrhythmias that affect the reading of the ECG signals: 94.9% of all the missing PR values (yellow) have some arrhythmias (blue) that require exclusion.

SNOMED CT	Arrhythmias	Frequency
49436004	Atrial fibrillation	109,673
17338001	Ventricular premature beats	99,197
406461004	Ectopic atrial beats	92,318
63593006	Supraventricular premature beats	83,480
5370000	Atrial flutter	23,189

Table 1: Top 5 arrhythmias by counts that need to be excluded. Among the arrhythmias sorted by SNOMED CT Code, atrial fibrillation is the leading cause for excluding an ECG from the analysis.

2.2 Model Architecture

Building on the previous work of Ansari et al. [2], a ResNet model was used to estimate the PR interval. The raw 12-lead ECG signals were fed to a single 1D-convolutional layer with 64 filters and a kernel size of 15, followed by 9 residual blocks. Each block consists of a max pooling layer at the beginning, then 3 1D-convolutional layers with 64 filters and a kernel size of 15, and a residual connection node at the end of each block. Unless otherwise stated, a ReLU activation function is used. Finally, the last layer is flattened before going into a single-output layer with a linear activation (Figure 3) to estimate the PR interval values.

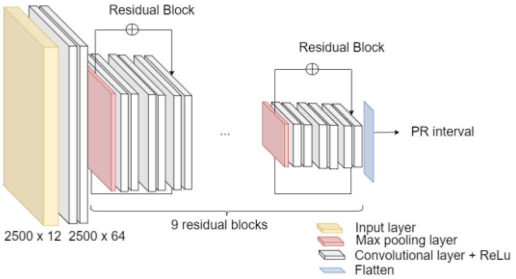

Figure 3: Model architecture.

2.3 Training and Testing

The model is trained on 70% and validated on 10% of all the patients in the cohort. The last 20% is reserved for testing purposes. This approach helps to preserve the independence of the training, validate and test sets by ensuring that a patient only belongs to at most one data subset. This results in 751,886 ECGs for training, 106,609 for validation, and 221,424 for testing, corresponding to 279,403, 39,903 and 80,223 of patients.

As the output is continuous, the network is optimized on a Huber loss function that behaves linearly when the absolute error is greater than 20 samples (80 ms) and quadratically otherwise. In order to avoid overfitting, an Adam optimizer is used with a starting learning rate of 0.01, which decreases by a factor of 10 whenever the validation loss stops dropping after 2 consecutive epochs. The training is terminated if the validation loss does not improve for 3 consecutive epochs.

3. Results

3.1 Model Performance

On the test set, the ResNet achieved a correlation of 94.3%, a bias of 0.269ms, and a mean absolute error of 4.732ms against the noisy test labels. In addition to these more traditional performance metrics, the percentage of observations with an absolute error between the estimated and the observed PR values less than 10ms is included as a metric of precision, as the human eyes are unable to detect differences smaller than 10ms [4]. As a result, if the estimated and observed PR are within 10ms apart, they are considered as identical. The percent of observations with an absolute error of at most 10ms is 91.3%. The high correlation is visualized using the plot of the estimated and observed PR values for the test set. The Bland-Altman indicates an agreement of 96.84% between the labels and the model outputs (Figure 4).

Using SNOMED CT code, the performance stratified by different arrhythmias is calculated and present in Table 2.

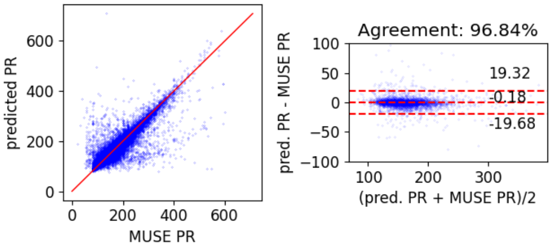

Figure 4: Plot of estimated vs observed PR intervals for the test dataset (left) Bland-Altman plot (right)

SNOMED CT	Cardiac Conditions	Counts (test)	Correlation (r)	%|error| <10ms
64730000	Normal sinus rhythm	143,652	0.94	93.43
49710005	Sinus bradycardia	37,661	0.97	91.97
164934002	T wave abnormal	33,537	0.93	90.28
11092001	Sinus tachycardia	28517	0.84	84.30
414795007	Myocardial ischemia	23,360	0.92	87.29
55827005	Left ventricular hypertrophy	22,924	0.94	91.48
7326005	Inferior myocardial infarction on ECG	21,250	0.93	88.55
428750005	Nonspecific ST-T abnormality on ECG	19,362	0.92	86.15
39732003	Left axis deviation	17,979	0.92	88.05
425623009	Lateral ischemia	16,754	0.92	87.15

Table 2: List of 10 most frequent cardiac conditions

3.2 Stratified Sampling

Stratification is a common sampling technique that ensures the best representation by dividing the population into subgroups called strata, assuming that similar behaviors are observed within the same stratum.

Since the distribution of the absolute errors between the model output and the labels is heavily skewed, different strategies are deployed for the cases with an absolute error of at most 20ms and for those with that of greater than 20ms.

For all the cases with an absolute error of more than 20ms, a representative sample is obtained using stratified sampling with an optimal allocation strategy. More specifically, all these cases are divided into four strata, where the number of observations sampled is proportional to the standard deviation (St.D.) and the size of each stratum, allowing more samples from the stratum with most variation and the larger stratum (Table 3).

On the other hand, a fixed sample size of 40, equally divided into 2 strata using 10ms as the cutoff, is drawn from cases with an absolute error of at most 20ms. Since any difference less than 10ms will fail to be detected during the manual adjudication process, the difference in behavior between the model output and the MUSE label will be characterized by all the cases between 10ms and 20ms.

In addition, to maintain the independence of each stratum, all patients are sampled only once. The resulting 6 strata are presented in Figure 5.

Interval	Size (Counts)	%	St.D.	# samples required
(0,10]	200,658	90.62	2.31	20
(10,20]	14,094	6.37	2.68	20
(20,30]	3,591	1.62	2.86	29
(30,60]	2,194	0.99	7.91	49
(60,120]	550	0.25	17.45	27
(120,600]	337	0.15	57.11	55

Table 3: Description of each stratum

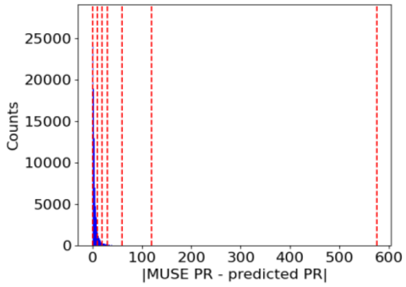

Figure 5: Distribution of the absolute errors

3.3 Manual Adjudication Result

A blinded manual adjudication is performed to confirm the accuracy of both the model output and the label on which it was trained.

The R location and the QRS duration was provided by the MUSE software. The P location for both the model and the MUSE labels was calculated by subtracting the PR interval and half of QRS duration from the R location. Each P location was randomly assigned to two colors of choice. During the adjudication process, the adjudicator needed to decide (1) whether the R location was accurate, (2) whether the QRS duration was correct, and (3) which P location was correct (by color).

The estimated true accuracy of the model in each stratum is shown in Table 4. Similar performances between the model output and the labels are observed for the strata where the absolute difference is relatively small. As the absolute error increases in size, the model gradually overperforms the MUSE labels in terms of accuracy, with the most difference located in the latter stratum.

Interval	Est. Acc (ResNet)	Est. Acc (MUSE)	St.D. (ResNet)	St.D. (MUSE)
(0,10]	1.00	1.00	0.00	0.00
(10,20]	0.70	0.70	0.21	0.21
(20,30]	0.61	0.39	0.24	0.24
(30,60]	0.56	0.32	0.25	0.22
(60,120]	0.74	0.07	0.19	0.07
(120,600]	0.79	0.05	0.17	0.04

Table 4: Table of the estimated accuracy (Est. Acc) across stratum and the corresponding standard deviation (St.D.) for ResNet outputs and MUSE labels.

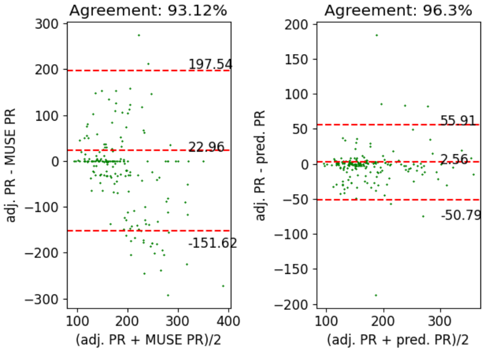

Figure 6: Bland-Altman of MUSE and manual adjudication (left) Bland-Altman of estimated (right) and manual adjudication.

Then, using a pooled variance approach, the 95% confidence interval for the overall accuracy for both the model output and the MUSE label was calculated. The accuracy of the MUSE label is estimated to be 96% [95% CI: 95.7-96.3], implying that approximately 4% of the MUSE labels are incorrect. On the other hand, the ResNet accurately estimates 96.9% [95% CI: 96.59-97.42] of the PR intervals, higher than the estimated accuracy of the noisy labels on which it was trained on. The Bland-Altman plot in Figure 6 also shows a narrower confidence band and higher agreement for the ResNet model than the MUSE labels, and therefore a closer match between the estimated and the manually adjudicated PR values, which were calculated only when at least one of P locations was correct. This results in 189 manually adjudicated PR values in total.

4. Discussion and Conclusions

Using a tailored cohort extracted from the MUSE dataset collected at the University of Michigan, a ResNet model was trained to estimate PR interval values using the noisy labels that result from a laborious process of review and manual adjudication. From a technical standpoint, the model can produce interval values that have a low bias and are highly correlated with the labels on which it was trained.

Taking a step further, the accuracy of the MUSE labels and the model output was estimated by performing a blinded manual adjudication using stratified sampling. The accuracy approximated by this process demonstrated that the model is able to produce more consistent and accurate PR values than the MUSE labels, and this difference in performance is statistically significant.

Although the ResNet model performs better than the noisy labels provided for training, the comparison heavily relies on manual adjudication with variable degrees of accuracy depending on interpreting clinician [5]. As physicians can have their own biases when it comes to interpreting ECGs [5], an adjudication process with multiple electrophysiologists needs to be conducted to ensure fairness. As another potential improvement, since the performance of the algorithm varies based on different arrhythmias (Table 2), adding the diagnosis to the model can improve the estimation. This work can also be incorporated into a more integrated healthcare platform to optimize patient's diagnoses. Finally, since the study is limited to adult patients, the option to conduct the same analysis for pediatric patients can be explored though the model will need to be re-trained for optimal results.

References

[1] S. Chatterjee et al. "Review of noise removal techniques in ECG signals," IET signal process., vol. 14, no. 9, pp. 569–590, Dec. 2020

[2] S. Ansari et al., "Classification of 12-Lead Electrocardiograms Using Residual Neural Networks and Transfer Learning," 2020 Computing in Cardiology, 2020, pp. 1-4

[3] C. Padayachee et al., "Can the computer tell me what's wrong with my heart? Early day lessons from digital hospital and ECG interpretation," Heart Lung and Circulation 2018; 27:S303–S304.

[4] J. J. Bailey et al., "Recommendations for standardization and specifications in automated electrocardiography: bandwidth and digital signal processing. A report for health professionals by an ad hoc writing group of the Committee on Electrocardiography and Cardiac Electrophysiology of the Council on Clinical Cardiology, American Heart Association.," Circulation, vol. 81, no. 2, pp. 730–739, Feb. 1990

[5] S. A. Hicks et al., "Explaining deep neural networks for knowledge discovery in electrocardiogram analysis," Sci Rep, vol. 11, no. 1, p. 10949, Dec. 2021

Address for correspondence:

Loc Cao
Building 10-A103, North Campus Research Complex
2800 Plymouth Road, Ann Arbor, MI 48109-2800
caolq@med.umich.edu