

Electrode Reversal Robustness of Deep Learning Classifiers for ECG Diagnosis

Assad H. Thary Al-Ghraiiri

¹Department of Computer Science, College of Science, Al-Nahrain University, Jadriya, Baghdad, 10001, Iraq,

DOI: 10.37648/ijps.v22i01.008

¹Received: 06/05/2026; Accepted: 19/06/2026; Published: 22/08/2026

Abstract

Swapped limb electrodes are a common technical error in electrocardiography, estimated at about 4% of acute care recordings. Deep learning classifiers now read ECGs automatically, but nobody has measured how often their diagnoses change when the electrodes are swapped, or which defense protects the diagnosis best. We used the public PTB XL dataset (21,388 labeled 12 lead ECGs) with its official training, validation and test folds and five diagnostic super classes. Because every limb lead is a fixed combination of three electrode potentials, all five swaps of the right arm, left arm and left leg electrodes can be computed exactly from real recordings, while the chest leads stay unchanged. We trained a 1D residual network and three defenses: swap augmentation, a swap detector followed by exact correction, and a multitask network that diagnoses and detects the swap together. Each model was trained with three seeds. Three hypotheses were written down before the full run. On correctly recorded test ECGs (2,158 recordings from 1,877 patients), the baseline reached a macro AUROC of 0.923. Averaged over the five swaps, it fell to 0.732 (difference 0.191, 95% CI 0.183 to 0.199) and changed the predicted diagnosis set of 78% of ECGs. After a right arm and left leg swap, 99.7% of ECGs without infarction that had been read correctly were newly called myocardial infarction. Detection followed by correction restored a swap averaged AUROC of 0.923, changed only 0.18% of diagnoses and created no new infarction calls. The detector identified the correct condition in 99.1% of cases, and 92% of its few errors confused conditions that differ only by the left arm and left leg electrodes. Swap augmentation and the multitask network also preserved AUROC (0.919 and 0.922) but still changed 15% and 19% of diagnoses. No defense reduced accuracy on correctly recorded ECGs. All three hypotheses were supported (adjusted $p = 0.003$). Standard deep learning ECG classifiers are highly vulnerable to limb electrode swaps, and the error most often looks like a myocardial infarction. Detecting the swap and undoing it exactly protects the diagnosis almost completely and also tells the user that the recording needs attention.

Keywords: *Electrocardiography; Deep Learning; Electrode Reversal; Limb Lead Swap; Robustness; Myocardial Infarction; PTB XL*

1. Introduction

A 12 lead ECG is only as good as the way it was recorded. When the cables of the limb electrodes are connected to the wrong limbs, the recorder still produces a clean looking tracing, but the limb leads are rearranged and partly inverted. Such errors can mimic conduction disease, ischemia and myocardial infarction [1]. They are not rare: in acute care, electrode swaps are estimated to affect about 4% of recordings [2].

At the same time, deep learning has made automatic ECG diagnosis accurate enough to match or exceed physicians on several tasks [3, 4]. On the public PTB XL dataset, convolutional networks with residual or inception architectures perform best across diagnostic tasks [5, 6]. These networks learn from correctly recorded ECGs and are then applied

¹ How to cite the article: Al-Ghraiiri A.H.T. (August 2026); Electrode Reversal Robustness of Deep Learning Classifiers for ECG Diagnosis; International Journal of Professional Studies; Jul-Dec 2026, Vol 22, 88-101; DOI: <http://doi.org/10.37648/ijps.v22i01.008>

to whatever the recorder produces. Nothing in a standard classifier check whether the electrodes were in the right place, and a swapped recording is passed through like any other.

Detecting electrode swaps has a long history, and machine learning detects most swaps well, with left arm and left leg interchange as the persistent exception [7]. Recent deep learning work uses the physics of the leads to detect and correct limb electrode swaps [2]. What remains unknown is the consequence for the diagnosis itself. How often does a deep learning classifier change its diagnosis when the electrodes are swapped, which swaps matter most, and which defense protects the diagnosis rather than only the swap label?

We answer these questions on PTB XL. Our approach rests on a simple physical fact: all limb leads are linear combinations of three electrode potentials, so every swap of the right arm, left arm and left leg electrodes can be computed exactly from a real recording, with no approximation. The chest leads do not change at all, because their reference is the mean of the same three potentials [1].

Our contributions are the following.

1. A complete measurement of how all five limb electrode swaps affect a standard deep learning ECG classifier, in terms of AUROC, the fraction of diagnoses that change, and newly created infarction calls.
2. A comparison of three defenses with the diagnosis as the outcome: swap augmentation, swap detection followed by exact correction, and a multitask network.
3. Evidence that detection followed by exact correction restores diagnostic performance almost completely without any loss on correctly recorded ECGs, together with an analysis showing that the remaining detection errors concern the left arm and left leg electrodes.
4. Three hypotheses fixed before the full experiment, tested with patient level bootstrap and multiplicity correction.

2. Literature Survey

Deep learning ECG diagnosis. Hannun and et.al. showed that a deep network could classify arrhythmias from single lead recordings at the level of cardiologists [3], and Ribeiro and colleagues extended this to the 12 lead ECG [4]. PTB XL, a public collection of 21,799 clinical 12 lead recordings with diagnostic statements [6], made systematic comparison possible. Its benchmark found residual and inception based convolutional networks strongest across tasks [5]. Deep learning is also widely used in biomedical imaging, where explainable attention mechanisms have been built into segmentation networks so that their decisions can be inspected [8]. Like most work in the field, these ECG studies evaluate classifiers on recordings taken as given.

Missing and misplaced leads. The PhysioNet and Computing in Cardiology Challenge 2021 asked how well diagnosis works with 12, 6, 4, 3 or 2 leads [9]. Deep learning has also been used to detect misplaced chest electrodes, where it outperformed physicians [10]. Both lines of work concern leads that are absent or slightly displaced, not limb electrodes connected to the wrong limb. Deep ECG classifiers can also be misled by small, smooth perturbations that experts cannot see [11], and general benchmarks test robustness to common corruptions [12]. Electrode swaps differ from both: they are common in practice, physically exact, and change the limb leads far more than an imperceptible perturbation does.

Limb electrode swaps. The visual signs of each swap are well described in clinical texts. Batchvarov and colleagues summarized how swaps of the right arm, left arm and left leg cables rearrange and invert the limb leads while leaving the chest leads untouched, and how they can simulate ischemia and infarction [1]. A systematic review of machine learning for swap detection found high accuracy for most interchanges, except left arm and left leg [7]. Most recently, Pandey and colleagues used the lead to electrode transformation to synthesize 180,000 swapped ECGs and trained deep networks to detect and correct limb electrode swaps [2]. Their outcome was the swap itself.

Gap. We found no study that measured how limb electrode swaps change the output of a deep learning diagnostic classifier, or that compared defenses by their effect on the diagnosis. This is the question that matters to a clinician who receives an automatic interpretation.

3. Materials and Methods

3.1 Dataset Description

We used PTB XL version 1.0.3 [6], distributed through PhysioNet [13], at a sampling rate of 100 Hz (10 seconds, 1,000 samples per lead). Diagnostic statements were mapped to the five superclasses of the PTB XL benchmark [5]: normal ECG, myocardial infarction, ST and T change, conduction disturbance and hypertrophy. Each record may carry several superclasses. Records without any superclass were excluded, leaving 21,388 of 21,799 ECGs. We used the recommended folds: folds 1 to 8 for training, fold 9 for validation and fold 10 for testing (Table 1). No patient appears in both the training and test folds, which the program checked.

Leads were reordered by name from each file header, never by position. As a check on the raw data, lead III equaled lead II minus lead I to within a median maximum deviation of 0.001 mV across records.

Table 1. Data used in this study.

Split	Folds	ECGs	Normal	Infarction	ST and T change	Conduction	Hypertrophy
Training	1 to 8	17,084	7,596	4,379	4,186	3,907	2,119
Validation	9	2,146	not used for training				
Test	10	2,158 (1,877 patients)	held out				

Superclass counts are shown for the training set, where they determine what the models learn; an ECG can belong to more than one superclass.

3.2 Exact Simulation of Limb Electrode Swaps

Let Φ denote the electrode potentials at the right arm (RA), left arm (LA) and left leg (LL). The limb leads and the chest leads are

$$\begin{aligned} I &= \Phi_{LA} - \Phi_{RA}, & II &= \Phi_{LL} - \Phi_{RA}, & III &= \Phi_{LL} - \Phi_{LA}, \\ aVR &= \Phi_{RA} - \frac{1}{2}(\Phi_{LA} + \Phi_{LL}), & V_i &= \Phi_i - \frac{1}{3}(\Phi_{RA} + \Phi_{LA} + \Phi_{LL}) \end{aligned} \quad \dots (1)$$

with aVL and aVF defined like aVR. A swap connects the electrode of one limb to the recorder input of another, which permutes the three potentials. The chest leads are unchanged because their reference, the mean of the three potentials, does not depend on their order [1]. Since all limb leads depend only on differences of potentials, we can set the right arm potential to zero, the left arm potential to lead I and the left leg potential to lead II, apply the permutation, and recompute. For each swap σ this gives a fixed 6 by 2 matrices:

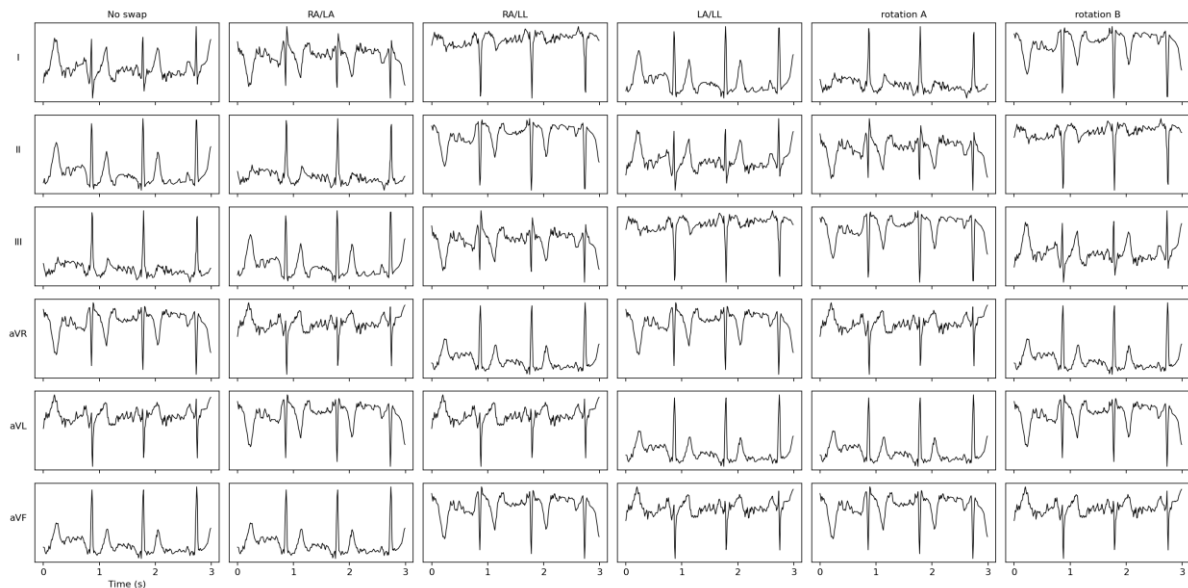
$$[I', II', III', aVR', aVL', aVF']^T = M_{\sigma} [I, II]^T \quad \dots (2)$$

There are five possible swaps of three electrodes: three exchanges of two electrodes and two rotations of all three (Table 2). The rules in Table 2 were derived by the program and agree with the clinical descriptions for the three exchanges [1]. Every swap is undone exactly by its inverse, which was verified in unit tests. To keep clean and swapped inputs strictly comparable, the limb leads of every input, including unmodified recordings, were recomputed from leads I and II in the same way.

Table 2. The five limb electrode swaps and their effect on the limb leads. Chest leads never change.

Swap	Electrodes at the RA, LA and LL inputs	Effect on the limb leads
RA/LA	LA, RA, LL	I inverted; II and III exchanged; aVR and aVL exchanged; aVF unchanged
RA/LL	LL, LA, RA	I becomes minus III; II inverted; III becomes minus I; aVR and aVF exchanged; aVL unchanged
LA/LL	RA, LL, LA	I and II exchanged; III inverted; aVL and aVF exchanged; aVR unchanged
Rotation A	LA, LL, RA	I becomes III; II becomes minus I; III becomes minus II; aVR becomes aVL; aVL becomes aVF; aVF becomes aVR
Rotation B	LL, RA, LA	I becomes minus II; II becomes minus III; III becomes I; aVR becomes aVF; aVL becomes aVR; aVF becomes aVL

Figure 1 shows the six limb leads of one test ECG under each condition. The swapped tracings look like plausible recordings, which is exactly why swaps are easy to miss.

**Figure 1.** The first three seconds of the six limb leads of one test ECG, as recorded and under each of the five simulated swaps. The chest leads, not shown, are identical in all six columns.

3.3 Diagnostic Classifier

The classifier was a one-dimensional residual network [14], following the benchmark finding that residual architectures perform best on PTB XL [5]. Its first layer standardized each lead with fixed means and standard deviations computed on correctly recorded training ECGs, so that swaps always act on signals in millivolts, exactly as on a real recording. A convolutional stem (64 channels, kernel 15, stride 2) was followed by four stages of two residual blocks each (kernel 7; 64, 128, 128 and 256 channels; strides 1, 2, 2 and 2). Global average and global maximum pooling were concatenated, followed by dropout of 0.3 and a linear layer with one sigmoid output per superclass.

3.4 Defenses

We compared the baseline with three defenses built on the same network.

- **Swap augmentation.** During training, each ECG was swapped with probability 0.5, using one of the five swaps chosen at random, and kept its original diagnosis.
- **Detect and correct.** A separate network with the same backbone and a six-way SoftMax output was trained to recognize the condition (no swap or one of the five swaps), with all six conditions equally likely during training. At test time, the inverse of the predicted swap was applied and the baseline classifier diagnosed the corrected recording.
- **Multitask.** One network with a diagnostic output and a six-way swap output, trained on the sum of the binary cross entropy and the swap cross entropy, with swaps applied with probability 0.5.

3.5 Training and Model Selection

All networks were trained for 30 epochs with batches of 128, AdamW [15] (learning rate 0.001, weight decay 0.01) and a one cycle learning rate schedule [16]. The epoch kept for testing was chosen on the validation fold by a criterion fixed in advance: clean macro AUROC for the baseline; macro AUROC averaged over all six conditions for swap augmentation and the multitask network; and accuracy over all six conditions for the detector. Every model type was trained with three seeds, and the predicted probabilities of the three seeds were averaged.

3.6 Evaluation

Every test ECG was evaluated under all six conditions, giving 12,948 test inputs per method. We report the macro AUROC over the five superclasses and the AUROC of each superclass. To measure changes in individual diagnoses, each method's per class thresholds were set to maximize F1 on correctly recorded validation ECGs. The **flip rate** is the fraction of test ECGs whose set of predicted superclasses changes when the swap is applied. The **new false infarction rate** is the fraction of ECGs without infarction, and not called infarction when correctly recorded, that are called infarction after the swap. For the swap detectors we report accuracy per condition and the confusion matrix. We also flagged ECGs as swapped at a threshold that produced 1% false alarms on unmodified validation ECGs.

3.7 Hypotheses Fixed in Advance

Three hypotheses were written to a time stamped file in the results folder before the full run started:

- **H1:** for the baseline, the macro AUROC averaged over the five swaps is lower than on correctly recorded ECGs.
- **H2:** detection followed by correction has a higher swap averaged macro AUROC than the baseline.
- **H3:** swap augmentation has a lower swap averaged flip rate than the baseline.

Secondary analyses covered each swap separately, new false infarction calls, detection per swap, swaps flagged in the unmodified test set, and whether each defense keeps the clean AUROC of the baseline within a margin of 0.01.

3.8 Statistics

Confidence intervals came from 1,000 bootstrap resamples at the patient level, so that all ECGs of a patient were resampled together. Differences were tested with a paired bootstrap in which both quantities were computed on the same resampled patients. The two sided p value was twice the smaller tail fraction and was reported as below 0.001 when no resample crossed zero. Holm correction [17] was applied across H1 to H3. All intervals are percentile bootstrap intervals [18].

3.9 Implementation

The pipeline was written in Python with PyTorch and the wfdb package and run on an NVIDIA T4 GPU in Google Colab. Each trained model was saved, and unit tests verified the swap algebra against the clinical rules, the exactness of every inverse, and the evaluation code on synthetic data.

4. Experimental Results

4.1 Performance on Correctly Recorded ECGs

On the 2,158 correctly recorded test ECGs, the baseline reached a macro AUROC of 0.923 (Table 3). The three single seed models scored 0.912, 0.917 and 0.912, so averaging the seeds added about 0.008. Detection followed by correction scored the same 0.923, because the detector almost never altered an unswapped recording. Swap augmentation reached 0.925 and the multitask network 0.930.

4.2 Swaps Sharply Degraded the Baseline (H1)

Every swap lowered the baseline AUROC (Table 3, Figure 2). Rotation A was the most damaging (0.633), followed by the right arm and left leg exchange (0.677), rotation B (0.719) and the right arm and left arm exchange (0.769). The left arm and left leg exchange did the least harm (0.864). Averaged over the five swaps, the baseline fell from 0.923 to 0.732, a loss of 0.191 (95% CI 0.183 to 0.199), which supports H1 (Section 4.4).

Table 3. Macro AUROC over the five super classes on the test fold under each condition (probabilities averaged over three seeds).

Method	No swap	RA/LA	RA/LL	LA/LL	Rotation A	Rotation B	Mean over swaps (95% CI)
Baseline	0.923	0.769	0.677	0.864	0.633	0.719	0.732 (0.723 to 0.742)
Swap augmentation	0.925	0.922	0.917	0.917	0.919	0.918	0.919 (0.911 to 0.926)
Detect and correct	0.923	0.923	0.923	0.924	0.923	0.923	0.923 (0.916 to 0.930)
Multitask	0.930	0.924	0.922	0.924	0.920	0.920	0.922 (0.914 to 0.929)

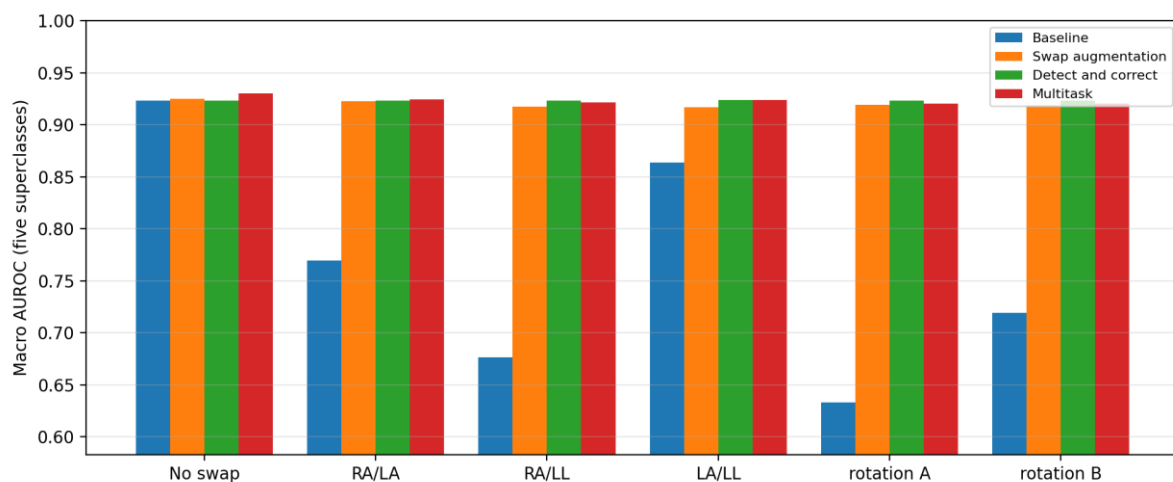


Figure 2. Macro AUROC of each method on correctly recorded ECGs and under each swap. The baseline loses between 0.06 and 0.29; all three defenses stay within 0.01 of their clean performance.

4.3 Swaps Changed Most Diagnoses And Created False Infarctions

At the thresholds chosen on correctly recorded validation ECGs, a swap changed the predicted set of superclasses for 78.2% of test ECGs on average (Table 4, Figure 3). The rate ranged from 38.8% for the left arm and left leg exchange to 89.7% for rotation A.

The errors were not random. Among ECGs without infarction that the baseline had read correctly, the right arm and left leg exchange produced a new infarction call in 99.7% and rotation A in 99.8% (Table 5, Figure 4). The right arm and left arm exchange and rotation B produced new infarction calls in 57.7% and 61.8%, and the left arm and left leg exchange in 8.3%.

Table 4. Diagnosis flip rate: fraction of test ECGs whose set of predicted super classes changes under the swap.

Method	RA/LA	RA/LL	LA/LL	Rotation A	Rotation B	Mean
Baseline	0.854	0.893	0.388	0.897	0.877	0.782
Swap augmentation	0.127	0.142	0.153	0.155	0.165	0.148
Detect and correct	0.002	0.002	0.002	0.001	0.002	0.002
Multitask	0.178	0.195	0.185	0.203	0.213	0.195

Table 5. New false infarction rate: among test ECGs without infarction that were not called infarction when correctly recorded, the fraction called infarction after the swap.

Method	RA/LA	RA/LL	LA/LL	Rotation A	Rotation B
Baseline	0.577	0.997	0.083	0.998	0.618
Swap augmentation	0.037	0.019	0.016	0.028	0.025
Detect and correct	0.000	0.000	0.000	0.000	0.000
Multitask	0.029	0.022	0.025	0.028	0.026

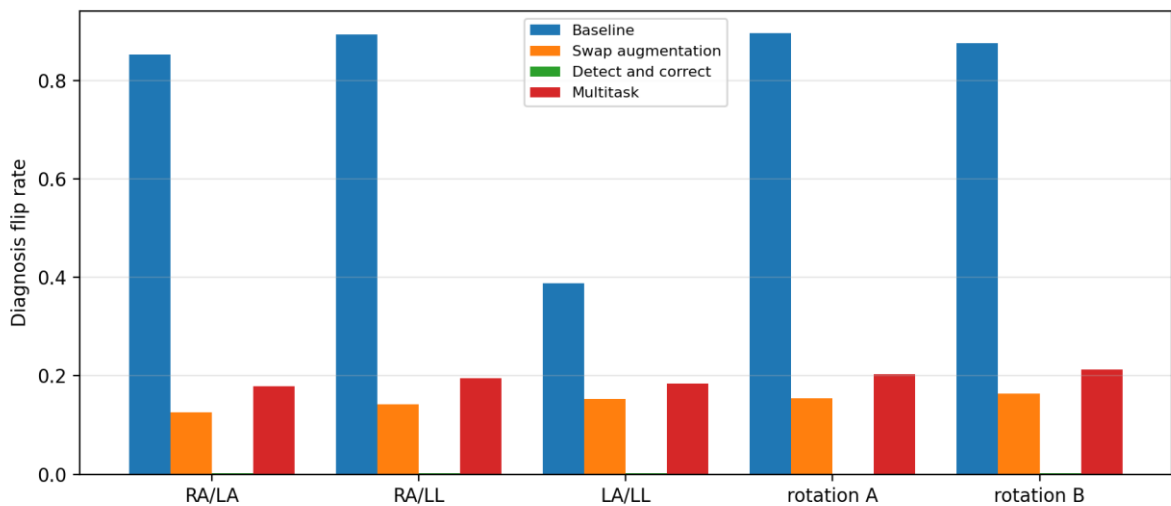


Figure 3. Diagnosis flip rate under each swap. The bars for detect and correct are too small to see (at most 0.2%).

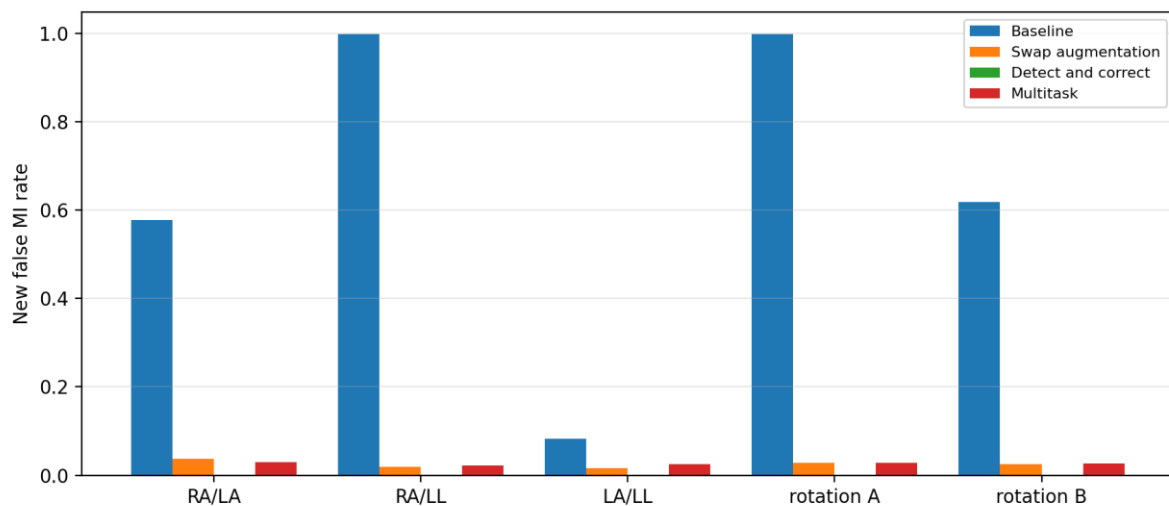


Figure 4. New false infarction rate under each swap. With the baseline, almost every ECG without infarction becomes an infarction after a right arm and left leg exchange or rotation A.

4.4 Defenses Protected The Diagnosis (H2 and H3)

All three hypotheses were supported after Holm correction (Table 6). Detection followed by correction raised the swap averaged AUROC by 0.191 over the baseline (H2), back to 0.923, the value on correctly recorded ECGs. It changed only 0.18% of diagnoses and created no new infarction calls. Swap augmentation lowered the mean flip rate by 0.634 compared with the baseline (H3), to 14.8%, with new false infarction rates of 1.6% to 3.7%. The multitask network preserved AUROC (0.922) but changed 19.5% of diagnoses.

Table 6. Prespecified hypotheses (patient level paired bootstrap, 1,000 resamples; Holm correction over H1 to H3).

Hypothesis	Quantity	Estimate (95% CI)	p	Adjusted p
H1	Baseline: clean minus swap averaged macro AUROC	0.191 (0.183 to 0.199)	below 0.001	0.003
H2	Swap averaged macro AUROC: detect and correct minus baseline	0.191 (0.183 to 0.199)	below 0.001	0.003
H3	Mean flip rate: baseline minus swap augmentation	0.634 (0.618 to 0.649)	below 0.001	0.003

No defense cost accuracy on correctly recorded ECGs (Table 7). All three met the noninferiority margin of 0.01, and the multitask network was slightly better than the baseline.

Table 7. Macro AUROC on correctly recorded test ECGs, defense minus baseline.

Method	Difference (95% CI)	Noninferior at margin 0.01
Swap augmentation	0.002 (−0.002 to 0.005)	yes
Detect and correct	0.000 (−0.0003 to 0.0003)	yes
Multitask	0.007 (0.004 to 0.010)	yes

4.5 Swap Detection

The detector recognized the correct condition in 99.1% of the 12,948 test inputs, with 99.0% to 99.2% for each condition. The swap output of the multitask network reached 99.0% (98.7% to 99.2% per condition). At a threshold

giving 1% false alarms on unmodified validation ECGs, the detector flagged 0.83% of unmodified test ECGs and 99.1% to 100% of swapped ones.

The detector made 118 errors, and they followed a clear pattern (Figure 5). In 108 of them (92%), it confused two conditions that differ only by an exchange of the left arm and left leg electrodes: no swap with the LA/LL exchange, the RA/LA exchange with rotation B, and the RA/LL exchange with rotation A. We confirmed with the swap algebra that each of these pairs is related by exactly one LA/LL exchange.

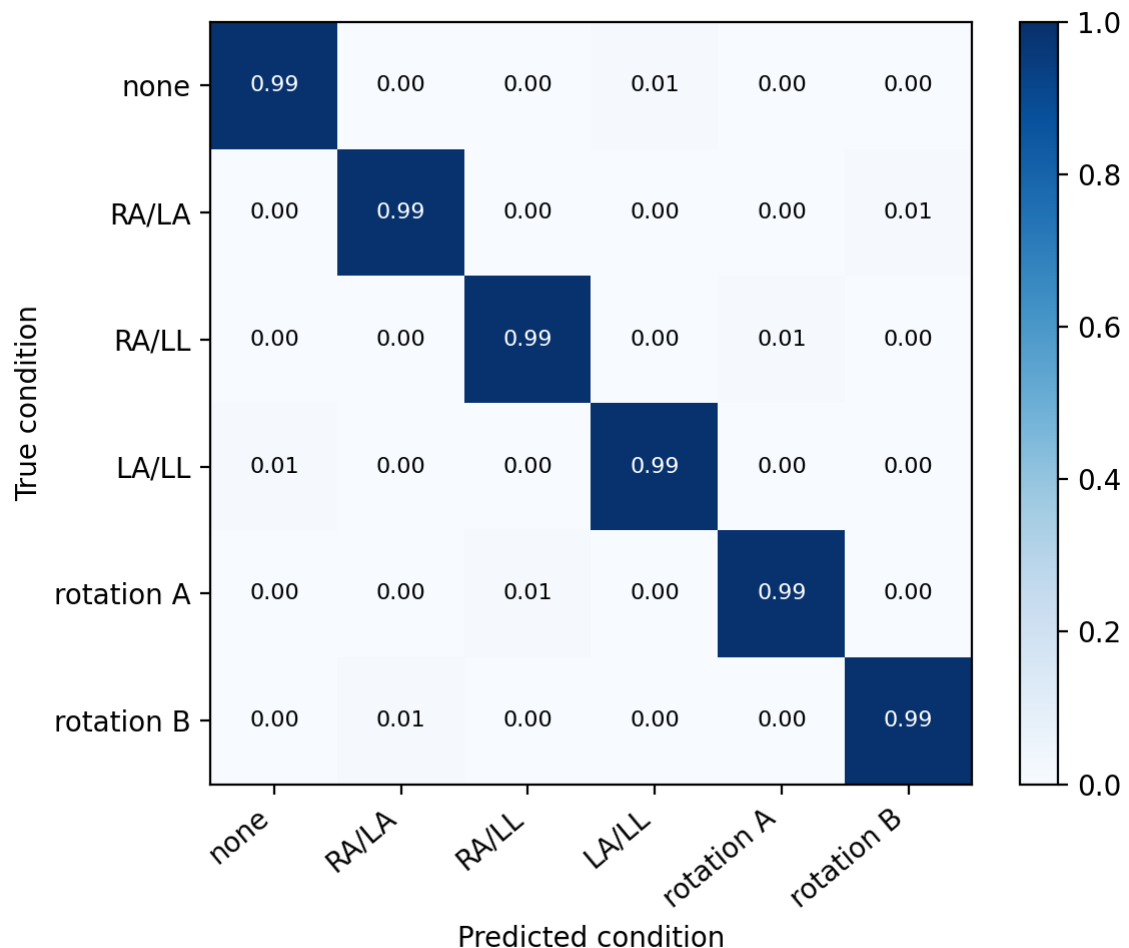


Figure 5. Confusion matrix of the swap detector over all test ECGs under all six conditions, normalized by row. The few errors fall in three symmetric pairs, each pair differing by a left arm and left leg exchange.

5. Discussion and Analysis

A standard deep learning ECG classifier treated swapped recordings as real ones. Its AUROC fell by 0.19 on average, nearly four in five diagnoses changed, and two of the five swaps turned almost every ECG without infarction into an infarction. The best defense, detecting the swap and undoing it, removed nearly all of this damage at no cost on correctly recorded ECGs.

5.1 The failures follow the electrophysiology of each swap

Figure 6 breaks the baseline's AUROC down by superclass and shows that each swap fails in its own way. The pattern matches what each swap does to the limb leads (Table 2).

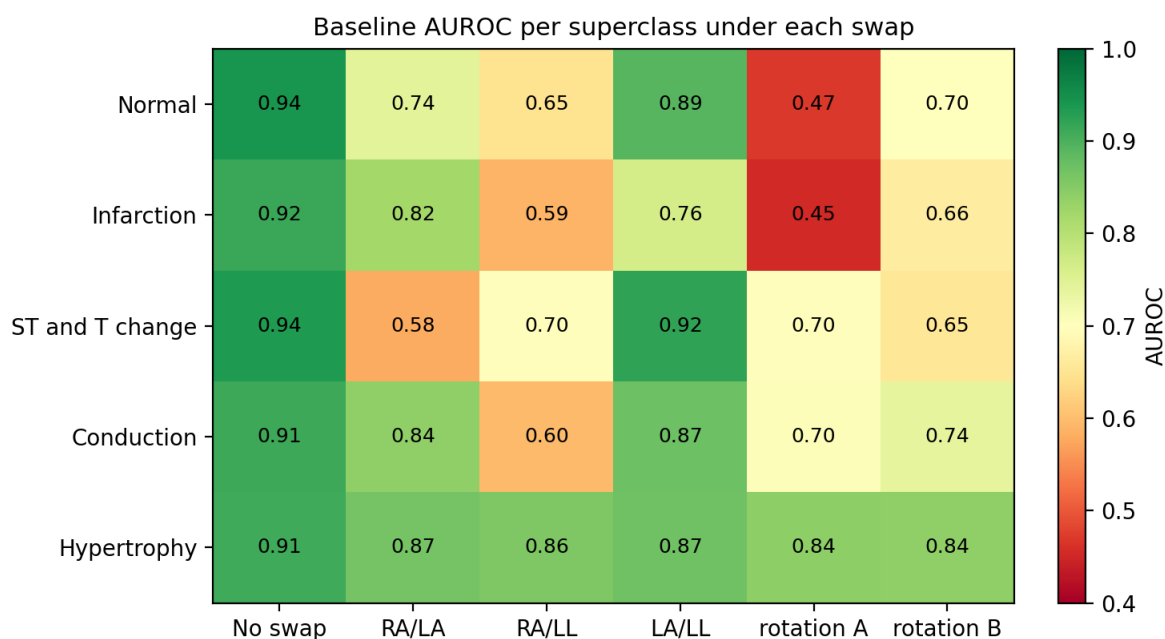


Figure 6. Baseline AUROC for each superclass under each condition. Green cells hold up; red cells collapse. Rotation A pushes the normal and infarction classes below chance.

The right arm and left leg exchange inverts lead II and replaces lead III with an inverted lead I, so the inferior leads show deflections opposite to their usual direction. Inverted inferior complexes resemble the Q waves and T wave changes of inferior infarction, and this is the swap that produced new infarction calls in 99.7% of eligible ECGs (Figure 4). The normal and infarction classes dropped to AUROCs of 0.65 and 0.59. Rotation A also inverts the inferior leads, since lead II becomes inverted lead I and lead III becomes inverted lead II. Here the normal and infarction AUROCs fell below 0.5 (0.47 and 0.45), meaning the classifier ranked these patients systematically in the wrong order.

The right arm and left arm exchange inverts lead I and swaps aVR with aVL, which flips the lateral limb leads. It hit the ST and T change class hardest (0.94 to 0.58), consistent with inverted lateral T waves looking like ischemia. The left arm and left leg exchange only exchanges leads I and II and inverts lead III. Its tracings stay close to normal in shape, it did the least damage (AUROC 0.864), and clinical descriptions likewise note that this swap is easy to overlook and needs specific criteria to recognize [1, 19].

Hypertrophy was the most stable class (AUROC 0.84 to 0.87 under every swap). Many of the voltage criteria for hypertrophy in clinical recommendations use the chest leads [20], and limb swaps leave the chest leads untouched. Taken together, Figure 6 shows that the network reads a swapped tracing exactly as a cardiologist would read the same tracing without suspecting a swap, and makes the errors that electrophysiology predicts.

5.2 Stable AUROC is not the same as stable diagnoses

Swap augmentation and the multitask network kept their AUROC within 0.01 of clean performance under every swap (Figure 2), yet they still changed 14.8% and 19.5% of diagnoses (Figure 3). Figure 7 places each method by both measures.

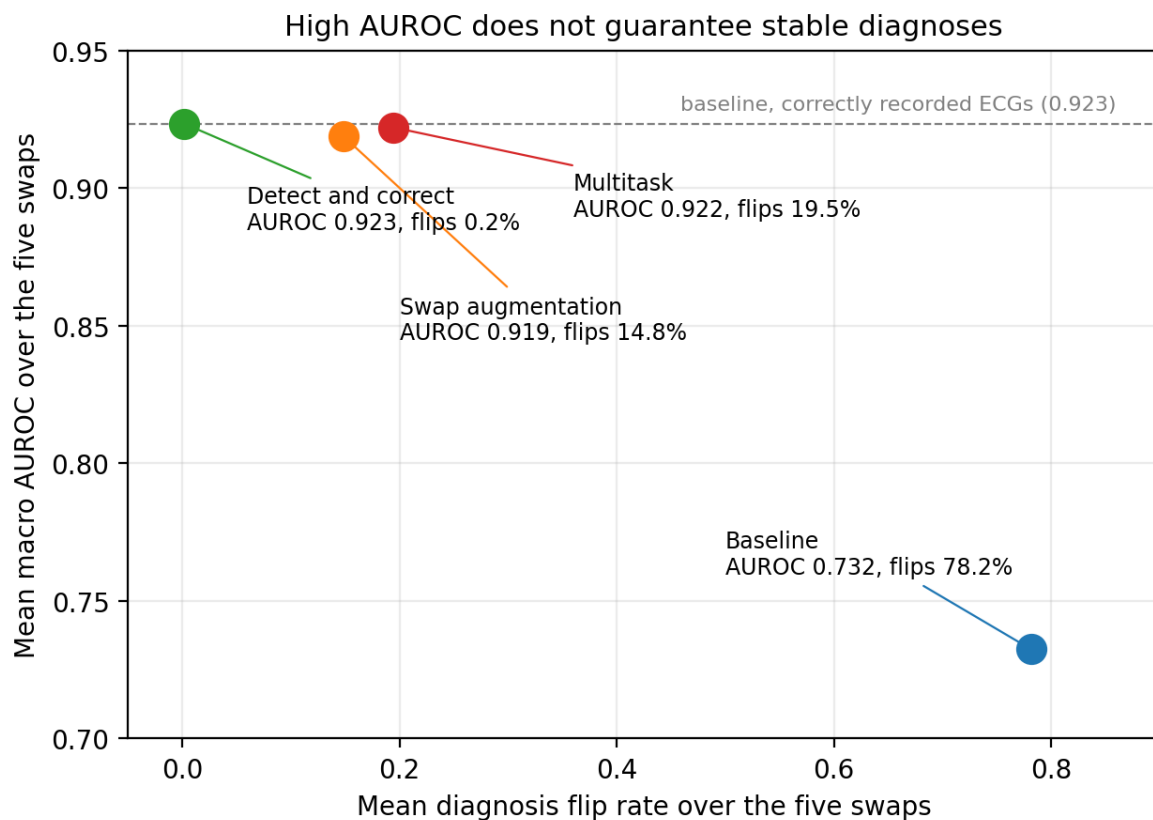


Figure 7. Mean macro AUROC against mean diagnosis flip rate over the five swaps for each method. The three defenses have almost the same AUROC, but only detection followed by correction keeps individual diagnoses stable.

The two measures answer different questions. AUROC asks whether patients with a condition are ranked above patients without it, and a small, consistent shift in every probability barely changes that ranking. A diagnosis, however, is a threshold decision for one patient, and a small shift can carry a probability across the threshold. A network that has learned to be roughly invariant to swaps still produces slightly different probabilities for the swapped and unswapped versions of the same ECG. Detection followed by correction avoids this because, whenever the swap is identified correctly, it feeds the classifier the original recording itself, so the prediction is identical. For clinical use the stability of the individual decision is what matters, so robustness studies should report flip rates and not AUROC alone. Robustness benchmarks in computer vision already report performance under each type of corruption rather than a single clean score [12]; reporting at the level of individual decisions is the natural next step for clinical classifiers.

5.3 Detector's Errors Reveal

The detector was correct in 99.1% of cases, and its errors were concentrated in three pairs of conditions that differ by a left arm and left leg exchange (Figure 5). This agrees with the systematic review, which found left arm and left leg interchange to be the hardest swap for machine learning [7], and with the recent deep learning work that singled out this swap [2]. The difficulty is long standing. Neural networks were applied to lead reversal detection in the 1990s [21]. A correlation based method that reached a sensitivity of at least 93% for 11 of 14 interchanges performed poorly for left arm and left leg reversal [22], and a widely used interpretation program needed specific enhancements to raise its detection of arm lead connection errors from 64% to 88% [23]. Our result adds a structural explanation: the detector's remaining uncertainty is almost entirely about which of those two electrodes is which, whatever else happened to the recording.

There is a reassuring side to this. The swap that is hardest to detect was also the least harmful to the diagnosis in our data (Figures 2 to 4). It is not harmless, though: it still changed 38.8% of the baseline's diagnoses.

At a threshold set for 1% false alarms on unmodified validation ECGs, the detector flagged 0.83% of unmodified test ECGs. This number is close to the false alarm rate we chose, so it cannot tell how many PTB XL recordings were truly swapped. The estimate of about 4% swapped recordings refers to acute care data [2], and PTB XL was not designed to measure it.

5.4 Implications

As deep learning ECG models move toward clinical use [3, 4, 24], these results argue for placing swap detection in front of any automatic ECG diagnosis. Detection followed by correction gave the most stable diagnoses, created no false infarctions and also produced a flag that can prompt a check of the electrodes or a new recording. In practice the correction should be shown to the reader rather than applied silently, so that a clinician can confirm it. Where a separate detector is not practical, swap augmentation is a cheap alternative that kept AUROC high and cut false infarction calls from as high as 99.8% to below 4%, at the price of more changed diagnoses. The multitask network gave both outputs from a single network and the best clean AUROC (0.930), which suggests that learning about swaps did not distract from diagnosis.

Compared with earlier work, our study shifts the outcome from the swap to the diagnosis. Batchvarov and colleagues described how swaps can mimic infarction [1]; we show that a deep learning classifier falls for these patterns almost every time. Pandey and colleagues showed that limb electrode swaps can be detected and corrected [2]; we show that doing so restores the diagnostic performance of a downstream classifier almost exactly.

6. Limitations and Gaps

First, the swaps were simulated. The simulation is exact for exchanges among the right arm, left arm and left leg electrodes, because it follows directly from how the leads are defined, but real recording errors also include swaps involving the right leg (neutral) electrode, swapped chest electrodes, electrodes placed in the wrong position rather than connected to the wrong cable, and poor skin contact. None of these were studied.

Second, PTB XL itself probably contains some recordings that were truly swapped. These act as label noise in training and sit unrecognized in the no swap condition of the test set, so the true detection accuracy on clean recordings may be slightly higher or lower than reported.

Third, all results come from one dataset at 100 Hz with five diagnostic superclasses and one network architecture. Other datasets, sampling rates, finer diagnostic labels or other architectures, including large pretrained models, may behave differently. External validation on recordings with documented swaps would be the strongest test.

Fourth, the detector was trained with all six conditions equally likely, whereas real swaps occur in a few percent of recordings. In deployment, the balance between missed swaps and false alarms would have to be set for the expected prevalence. Flip rates also depend on the thresholds, which we chose to maximize F1 on correctly recorded validation ECGs.

Finally, the hypotheses were written after the code was designed and tested on a small subset, but before the full run. H1 in particular was expected from clinical knowledge; the confirmatory value of the study lies mostly in the size of the effects and in the comparison of the defenses.

7. Conclusion

Limb electrode swaps break standard deep learning ECG classifiers in predictable and dangerous ways. On PTB XL, the swap averaged AUROC of a residual network fell from 0.923 to 0.732, most diagnoses changed, and two swaps turned nearly every ECG without infarction into an infarction call. Detecting the swap and undoing it exactly restored performance to 0.923, kept 99.8% of diagnoses unchanged and also flagged the recording. Swap augmentation and multitask training preserved AUROC but left 15% to 20% of diagnoses unstable. Automatic ECG interpretation should check the electrodes before it reads the heart, and robustness studies should report how often individual diagnoses change, not only AUROC.

Declarations

Competing Interests

The authors declare that they have no known competing financial interests (e.g., funding, employment, consultancy, patents, stock ownership) or non-financial interests (e.g., personal, business, or family relationships) that could have appeared to influence the work reported in this article.

Data Availability. PTB XL version 1.0.3 is openly available on PhysioNet [6, 13]. The results tables, per condition predictions and the time stamped hypothesis file underlying this article are available from the corresponding author.

Ethics Statement. This study used only a public, anonymized dataset. No new data were collected from human participants.

Funding Statement

The author(s) declare that no financial support was received for the research, authorship, and/or publication of this article.

References

- [1]. Batchvarov VN, Malik M, Camm AJ. Incorrect electrode cable connection during electrocardiographic recording. *Europace*. 2007;9(11):1081–1090. doi:10.1093/europace/eum198.
- [2]. Pandey R, Reyna MA, Koscova Z, Clifford GD, Sameni R. Detecting and correcting electrocardiogram limb-lead electrode swapping via deep learning. In: *Computing in Cardiology (CinC)*; 2026. Accepted abstract. <https://cinc.org/2026/Program/accepted/280.html>
- [3]. Hannun AY, Rajpurkar P, Haghpanahi M, Tison GH, Bourn C, Turakhia MP, Ng AY. Cardiologist-level arrhythmia detection and classification in ambulatory electrocardiograms using a deep neural network. *Nat Med*. 2019;25(1):65–69. doi:10.1038/s41591-018-0268-3. PMID: 30617320.
- [4]. Ribeiro AH, Ribeiro MH, Paixão GMM, Oliveira DM, Gomes PR, Canazart JA, Ferreira MPS, Andersson CR, Macfarlane PW, Meira W Jr, Schön TB, Ribeiro ALP. Automatic diagnosis of the 12-lead ECG using a deep neural network. *Nat Commun*. 2020;11:1760. doi:10.1038/s41467-020-15432-4. PMID: 32273514.
- [5]. Strodthoff N, Wagner P, Schaeffter T, Samek W. Deep learning for ECG analysis: benchmarks and insights from PTB-XL. *IEEE J Biomed Health Inform*. 2021;25(5):1519–1528. doi:10.1109/JBHI.2020.3022989.
- [6]. Wagner P, Strodthoff N, Bousseljot RD, Kreiseler D, Lunze FI, Samek W, Schaeffter T. PTB-XL, a large publicly available electrocardiography dataset. *Sci Data*. 2020;7:154. doi:10.1038/s41597-020-0495-6. PMID: 32451379.
- [7]. Rjoob K, Bond R, Finlay D, McGilligan V, Leslie SJ, Rababah A, Iftikhar A, Guldenring D, Knoery C, McShane A, Peace A. Machine learning techniques for detecting electrode misplacement and interchanges when recording ECGs: a systematic review and meta-analysis. *J Electrocardiol*. 2020;62:116–123. doi:10.1016/j.jelectrocard.2020.08.011. PMID: 32209826.
- [8]. Mushgil HM, Al-Mukhtar FS, Ahmed EQ, Abduljabbar KS. Deep image segmentation using explainable attention mechanisms: applications in biomedical imaging. *Al-Nahrain Journal of Science*. 2025;28(4):241–258. doi:10.22401/ANJS.28.4.18.
- [9]. Reyna MA, Sadr N, Perez Alday EA, Gu A, Shah AJ, Robichaux C, Rad AB, Elola A, Seyedi S, Ansari S, Ghanbari H, Li Q, Sharma A, Clifford GD. Will two do? Varying dimensions in electrocardiography: the PhysioNet/Computing in Cardiology Challenge 2021. In: *Computing in Cardiology (CinC)*. 2021;48. doi:10.23919/CinC53138.2021.9662687.
- [10]. Rjoob K, Bond R, Finlay D, McGilligan V, Leslie SJ, Rababah A, Iftikhar A, Guldenring D, Knoery C, McShane A, Peace A. Reliable deep learning-based detection of misplaced chest electrodes during electrocardiogram recording: algorithm development and validation. *JMIR Med Inform*. 2021;9(4):e25347. doi:10.2196/25347. PMID: 33861203.

- [11]. Han X, Hu Y, Foschini L, Chinitz L, Jankelson L, Ranganath R. Deep learning models for electrocardiograms are susceptible to adversarial attack. *Nat Med.* 2020;26(3):360–363. doi:10.1038/s41591-020-0791-x. PMID: 32152582.
- [12]. Hendrycks D, Dietterich T. Benchmarking neural network robustness to common corruptions and perturbations. In: *International Conference on Learning Representations (ICLR)*; 2019. arXiv:1903.12261.
- [13]. Goldberger AL, Amaral LAN, 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. doi:10.1161/01.CIR.101.23.e215. PMID: 10852186.
- [14]. He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*; 2016:770–778. doi:10.1109/CVPR.2016.90.
- [15]. Loshchilov I, Hutter F. Decoupled weight decay regularization. In: *International Conference on Learning Representations (ICLR)*; 2019. arXiv:1711.05101.
- [16]. Smith LN, Topin N. Super-convergence: very fast training of neural networks using large learning rates. In: *Proc SPIE 11006, Artificial Intelligence and Machine Learning for Multi-Domain Operations Applications*; 2019:1100612. doi:10.1117/12.2520589.
- [17]. Holm S. A simple sequentially rejective multiple test procedure. *Scand J Stat.* 1979;6(2):65–70.
- [18]. Efron B, Tibshirani RJ. *An Introduction to the Bootstrap.* New York: Chapman and Hall; 1993:436 p.
- [19]. Abdollah H, Milliken JA. Recognition of electrocardiographic left arm/left leg lead reversal. *Am J Cardiol.* 1997;80(9):1247–1249. doi:10.1016/S0002-9149(97)00656-5.
- [20]. Hancock EW, Deal BJ, Mirvis DM, Okin P, Kligfield P, Gettes LS; AHA/ACCF/HRS. AHA/ACCF/HRS recommendations for the standardization and interpretation of the electrocardiogram: part V: electrocardiogram changes associated with cardiac chamber hypertrophy. *Circulation.* 2009;119(10):e251–e261. doi:10.1161/CIRCULATIONAHA.108.191097. PMID: 19281911.
- [21]. Hedén B, Ohlsson M, Holst H, Mjöman M, Rittner R, Pahlm O, Peterson C, Edenbrandt L. Detection of frequently overlooked electrocardiographic lead reversals using artificial neural networks. *Am J Cardiol.* 1996;78(5):600–604. doi:10.1016/S0002-9149(96)00377-3. PMID: 8806356.
- [22]. Kors JA, van Herpen G. Accurate automatic detection of electrode interchange in the electrocardiogram. *Am J Cardiol.* 2001;88(4):396–399. doi:10.1016/S0002-9149(01)01686-1. PMID: 11545755.
- [23]. Cooper C, Clark E, Macfarlane PW. Enhanced detection of electrode placement/connection errors. In: *Computers in Cardiology*; 2008;35:89–92. <https://www.cinc.org/old/Proceedings/2008/pdf/0089.pdf>
- [24]. Attia ZI, Kapa S, Lopez-Jimenez F, McKie PM, Ladewig DJ, Satam G, Pellicka PA, Enriquez-Sarano M, Noseworthy PA, Munger TM, Asirvatham SJ, Scott CG, Carter RE, Friedman PA. Screening for cardiac contractile dysfunction using an artificial intelligence-enabled electrocardiogram. *Nat Med.* 2019;25(1):70–74. doi:10.1038/s41591-018-0240-2. PMID: 30617318.