Continuity Contrastive Representations of ECG for Heart Block Detection from Only Lead-I

Teya S. Bergamaschi TEYA@MIT.EDU
Massachusetts Institute of Technology, Cambridge, MA, USA

Collin M. Stultz CMSTULTZ@MIT.EDU
Massachusetts Institute of Technology, Cambridge, MA, USA
Massachusetts General Hospital, Boston, MA, USA

Ridwan Alam RIDWAN@CSAIL.MIT.EDU
Massachusetts Institute of Technology, Cambridge, MA, USA

Abstract

Early detection of heart block can prevent life-threatening outcomes in patients with cardiac conduction disorders. While 12-lead ECG interpretation is the clinical standard apparatus, this work investigates detecting heart block from the lead-I ECG signals, the lead available on commercial smartwatches. We evaluate two state-of-the-art architectures: residual neural network and transformer encoder, both trained in a self-supervised contrastive learning manner with a novel signal-continuity-based ECG view definition on a dataset of 3.6 million ECGs from Massachusetts General Hospital. These models learn efficient ECG representations, which are used for heart block detection via linear probing on the PTB-XL dataset, a public ECG resource. To provide performance benchmarks, we compare our self-supervised models to supervised adaptations of both models trained on 10.6 thousand single-lead PTB-XL ECGs. Our analysis evaluates the performance of each model using the area under the receiver-operating curve (AUC), sensitivity, and specificity. We observe improved performance from the self-supervised pretraining. Additionally, we demonstrate the robust generalizability of these models in scarce-data scenarios, maintaining consistent performance with a reduced number of labeled training examples. This study highlights the potential of self-supervised learning in lead-I ECG diagnostics, offering promising implications for clinical applications where labeled data is scarce.

Keywords: Cardiac Electrophysiology, Deep Learning, ECG, Heart, Lead-I, PTB-XL, Single-Lead, Wearables, Resnet, Transformer.

Data and Code Availability ECG data used for heart block detection are available online (Physionet) at http://physionet.org/content/ptb-xl/1.0.3/. Access to data from the Massachusetts General Hospital is contingent on IRB approval. All code is available online at https://github.com/teyaberg/continuity-contrastive-ecg.

Institutional Review Board (IRB) Retrospective analyses for this study were approved by the IRB at the Massachusetts General Hospital (protocol #2020P000132). Given that the analyses are conducted on retrospectively collected data, no consent was required or mandated.

1. Introduction

Heart block is defined as the interruption of the regular flow of electrical impulses, also known as action potentials, across the heart chambers via the conduction pathways of the heart. This electric propagation triggers and regulates the contractions of the atria and the ventricles, keeping the heart pumping. Blockage along the conduction pathway impedes the rhythm and order of contractions, causing irregular heartbeats and arrhythmia, and can lead to syncope, cardiac arrest, and stroke (Tsao et al. (2023)). Depending on the location and severity of the blockage, the symptoms and treatments vary. For example, 1st degree atrioventricular block (AVB) is prevalent in 6% of the population older than 60 years, though often not severe. However, a 3rd degree AVB (prevalence of 0.03%) can collapse the ventricular contraction, leading to sudden cardiac arrest—a life-threatening condition (Kashou et al. (2023)). Heart block are reflective of the myocardial pathology or structural disorders; for example, left and right bundle branch block (LBBB, RBBB) can develop from ischemia (Tan et al. (2020)). Hence, early diagnosis and intervention can prevent severe outcomes and reduce the risk of long-term damage and heart failure.

The standard 12-lead ECG is clinically used for diagnosing patients. It records the time-varying magnitude of the net cardiac dipole moment along 12 axes; i.e., leads I, II, III, aVL, aVR, aVF, V1-V6. Each ECG lead represents a different view of the heart's conduction system, and the appearance of the ECG signal varies across the leads (Goldberger et al. (2017)). A cardiologist examines these views to identify and characterize the heart block. As 12-lead ECGs are typically ordered only when a heart condition is suspected or after symptoms manifest, most monitoring occurs post-symptomatically. This creates a substantial gap in proactive outpatient monitoring. The recent rise in the adoption of wearable and pocket ECG devices raises the possibility of outpatient cardiac monitoring (Bayoumy et al. (2021); Sana et al. (2020)). Since many wearable ambulatory ECG devices acquire only the lead-I and not all 12 leads, identifying heart block from lead-I ECG would enable early intervention and prevent severe outcomes and associated costs and burdens.

Deep learning (DL) approaches have been shown to be performant for predicting a variety of cardiovascular outcomes from the 12-lead ECG, and several investigators have proposed solutions for various single-lead cardiac disease detection tasks (Cao et al. (2022); Giudicessi et al. (2021); Maille et al. (2021)). Many of these DL methods have shown robust performance against data variation and noise artifacts. Yet, scarcity of labeled data, especially for rare and less prevalent health conditions, remains a major challenge in developing data-driven AI solutions for healthcare applications. For example, the PTB-XL dataset (Wagner et al. (2020)), a large public ECG resource, contains more than 8,000 "normal" ECGs in comparison to only 800 ECGs with AVB diagnosis. The generalizability and utility of DL solutions built from such imbalanced data require robust evaluation across various training and test scenarios.

In this work, we address these challenges as we explore solutions for identifying and characterizing heart block from lead-I ECG alone. We pretrain multiple contrastive representation learning pipelines on a large dataset of 3.6 million ECGs, which was acquired from about 800 thousand patients at the Massachusetts General Hospital. We propose a contrastive view with the hypothesis that ECG segments that are temporally subsequent possess more similarity than those that belong to other patients or the same patient at other time points. Then, we define our classification task on PTB-XL, which contains 14 thousand ECGs from patients with various heart block as well as patients with normal ECGs, and we aim to discriminate a lead-I ECG with evident heart block from the normal ones. Using the pretrained models, we acquire the representations for the PTB-XL ECG and compare simple classifiers trained on those representations against (a) corresponding large supervised models, and (b) classifiers trained on representations from a publicly available pretrained ECG model. We compare performance using the area under the receiver-operating curve (AUC), sensitivity, specificity, and predictive values. Finally, we demonstrate the generalizability of this approach by comparing the performance of the models across different training set sizes. To demonstrate clinical utility, we also use the self-supervised pipelines to build classifiers that can diagnose a patient with potential categories of heart block, including AVB, LBBB, and RBBB, from their lead-I ECG. Our work highlights the potential of deep learning-based lead-I ECG representations as a performant and generalizable apparatus in outpatient cardiac diagnostic applications.

2. Related Works

Deep learning (DL) has notably advanced ECG analyses in recent years from identifying cardiac arrhythmias (Hannun et al. (2019)) to inferring hemodynamic features (Alam et al. (2024a,b)). Supervised classifiers built using convolutional neural networks (CNN), including residual neural networks (Resnet), have been proposed to predict various arrhythmias (Alday et al. (2020)), atrial fibrillation (Attia et al. (2019)), cardiac arrest (Kwon et al. (2020)), heart defects (Ko et al. (2020)), and conduction disorders (Ribeiro et al. (2020); Bergamaschi et al. (2024)). Recent advances in transformer architectures have led to proposed ECG solutions for similar cardiac abnormality detection tasks (Natarajan et al. (2020); Strodthoff et al. (2020)). Motivated by the ability to learn from 12-lead ECG, similar models have been built and evaluated on single or four-lead ECG inputs from wearable or ambulatory devices (Reyna et al. (2021)). However, the scarcity of clinically validated annotations, especially for rare cardiac conditions and diseases, poses a perpetual challenge for supervised learning-based methods. Moreover, data quality, reliability, inter-rater variability, and confounding factors associated with health conditions add to the list of challenges.

Self-supervised learning has been investigated to bypass the need for large datasets with clinically validated annotations. For ECG, representations learned this way have proven useful for downstream classification tasks, e.g., arrhythmia classification (Kiyasseh et al. (2021)), atrial fibrillation and ventricular hypertrophy prediction (Diamant et al. (2022)), and heart block detection on 12-lead ECG (Bergamaschi et al. (2024)). Such representations are applied to build decision support systems in hypertension diagnosis (Schlesinger et al. (2022)). Adapting contrastive learning methods, Mehari et al. (2022) explores the utility of such methods in identifying different cardiac disorders on the PTB-XL dataset. Strodthoff et al. (2020) benchmarked several approaches for multilabel classification using the PTB-XL dataset, including convolutional neural networks with 1D ECG input, LSTMs, wavelet-based neural networks, and ensemble methods. Furthermore, various self-supervised methods, such as (Khunte et al. (2023); Sangha et al. (2024)), have been developed to detect atrial fibrillation, gender, and left ventricular systolic dysfunction from ECG images.

For deep learning in healthcare, generalization remains a significant obstacle toward real-world impact. In our earlier work (Bergamaschi et al. (2024)), we have shown that self-supervised representations remain performant when transferred to new datasets, following general transfer learning results (Kolesnikov et al. (2020)). However, many proposed performant models use conspicuously small test datasets: for classification tasks, Ko et al. (2020) tested their supervised model only on 328 ECG from 328 patients, while their model was trained on 91 thousand samples from 54 thousand patients; Ribeiro et al. (2020) had a test set of 827 samples in contrast to 2.3 million training samples. Self-supervised models are expected to learn and generate useful representations even for external datasets, yet generalizability remains under active exploration. Mehari et al. (2022) evaluated the pretrained model on about 2000 ECG split across 71 labels (average 30 samples per class) and kept the evaluation set as part of the pretraining dataset. The likelihood that a model will be generalizable is increased when the model is performant across several different datasets and training settings, representing data acquired from different patient populations. In this work, we test our models on varying proportions of the labeled dataset, each representing smaller and smaller patient populations. We propose this work as a baseline for future research, especially on publicly available ECG datasets.

Table 1: Demographics of ECG Datasets

Dataset	MGH	PTB-XL
Patients	770,615	18,885
ECG	3,617,253	21,837
Age (yr)	61 ± 18.7	61 ± 29.5
Female (%)	43	48.2

3. Data Description

To develop a self-supervised ECG representation learning pipeline, we utilize our "internal" ECG dataset from the Massachusetts General Hospital (MGH). We evaluate the utility of those representations on the ECG data from an independent "external" healthcare institution, namely PTB-XL. Though these datasets contain 12-lead clinical ECG, we use only the lead-I ECG for this study.

3.1. MGH dataset

The MGH dataset contains 3,617,253 clinical ECG records from 770,615 patients at the Massachusetts General Hospital (MGH) in Boston, MA, USA, acquired between 1981 and 2020. MGH is one of the largest full-service healthcare networks in the US providing primary, secondary, and tertiary care. The data were acquired by the in-hospital ECG acquisition machines (GE and Philips) and reviewed by attending cardiologists. The ECG recordings were acquired in millivolts (mV) of voltages with 12-bit quantization and at sampling rates of 250 Hz or 500 Hz. For the contrastive representation learning task, no labels are required, though the ECG metadata contains relevant information from ECG intervals to diagnostic statements.

3.2. PTB-XL

PTB-XL (Wagner et al. (2020)) is a large publicly available ECG dataset that contains 21,837 ECG from 18,885 patients with a variety of cardiovascular diagnoses including conduction disorders, myocardial infarctions, ischemia, and hypertrophic cardiomyopathy, as well as those without any cardiac disorders. The ECG data were acquired using devices by Schiller AG and are available at a sampling rate of 500 Hz. 67.13% of ECG annotations were performed by at least one cardiologist, 31.2% were generated via automatic device interpretation, and 1.67% had no initial annotation; following initial annotation, a random subset of the data was secondarily annotated by an independent cardiologist. The normal patient cohort has a mean±std age of 52±22 years (median 53 years), with 55% being female. The conduction disorder cohort is 72±42 years old (median 68 years) with 39% female patients. The WPW subset has an age distribution of 48±17 years, and 46% of patients are female. For our experiments, we curate a subset of 13,432 ECGs whose labels are validated by at least one cardiologist with confidence scores of at least 80%; included are 4,827 ECGs labeled as positive for one or more conduction disorder (CD) categories: 1,774 with fascicular block (FB), 1,639 RBBB, 609 LBBB, 808 with AV block (AVB), and only 70 ECG with evident Wolff-Parkinson-White (WPW) syndromes, or non-specific conduction disorders (IVCD, 780). 8,605 included ECGs are labeled as "normal," or without any cardiac disorders.

The demographic properties of the datasets are presented in Table 1. Moreover, the prevalence of each disease category in the PTB-XL dataset is presented in Table 2. Since the ECGs from MGH do not have labels, no similar summary statistics can be presented.

Table 2: Prevalence of Heart Block Categories

Block Type	**Prevalence**
CD	24.53%
WPW	0.31%
RBBB	8.42%
LBBB	3.14%
FB	9.05%
IVCD	3.45%
AVB	4.66%

4. Methods

4.1. Lead-I ECG Preprocessing

Our objective is to identify the presence of heart block from a lead-I ECG recording. The sampling rates for the input lead-I ECG vary across the datasets; hence, for our proposed models, we resample the signal at 250 Hz to ensure uniformity. Then, we remove baseline wander and high-frequency noise using a bandpass filter that only allows frequency components within a 0.05-to-40 Hz band. Following that, we exclude any 10-second ECG signal that has an absolute voltage amplitude larger than 5 mV, as such signal is physiologically nonsensical. We do not perform any signal normalization to preserve any comparative amplitude information. We use similar preprocessing methods to prepare the input signals for the baseline algorithms we use for comparison.

4.2. Continuity Contrastive ECG Views

To train the self-supervised models, we acquire two views of the same sample by splitting each 10-second ECG into two consecutive non-overlapping 5-second ECG views; that is, one of the views is a temporal continuation of the other view. In the concepts of similarity (Chen et al. (2020)), we hypothesize that two 5-second ECG signals are more similar when they belong to the same patient and are an immediate continuation in time. Considering the physiological and electrical characteristics of the heart, this definition is clinically meaningful. This sampling differs from that in many previously proposed methods, such as (Diamant et al. (2022); Kiyasseh et al. (2021)), which hypothesize ECG segments from the same patient at different time points should be similar. The contrastive learning, ideally, closely aligns the representations of the 5-second segments from the same 10-second ECG in the latent space, while driving other samples away.

4.3. Model Architectures

We adapt, implement, and compare two state-of-the-art architectures for our self-supervised learning pipeline; residual neural network (Resnet) (He et al. (2016)) and transformer (Baevski et al. (2020)).

ResNet We implement a Resnet architecture, REI, a single-channel Resnet for ECG lead-I consisting of four residual blocks, that we use for both our supervised and self-supervised learning pipelines. Each residual block in REI contains two convolutional layers and a skip connection. We add a single 1D convolutional layer before the residual blocks to ingest the lead-I ECG signal as input. We use a kernel size of 16 units for all convolutional filters. The input channels are of 1x1250 sample length, as we resample all 5-second ECG signals to a 250 Hz sampling rate to get the input tensor. The ingest layer convolves this tensor with a single sample stride over 64 filters. We learn convolutional layers with 128, 196, 256, and 128 filters for the four consecutive residual blocks. For each block, the skip connections are implemented with max pooling and a 1-to-1 convolutional layer. Batch normalization and rectified linear unit (ReLU) activation layers follow each convolutional layer of the model. Using average pooling, we get a 1x128 feature representation from the output of the last residual block.

Transformer Our implementation of the transformer architecture, XEI, a single-channel Xformer for ECG lead-I, is an adaptation of the Wav2Vec model (Baevski et al. (2020)), that we use for both our supervised and self-supervised learning pipelines. This architecture features a multi-layer CNN as the feature extractor and a transformer encoder. The feature extractor, which consists of multiple blocks, each containing a temporal convolutional layer followed by layer normalization and a GELU activation function, processes segments of the raw signal into representations, which are passed to the transformer encoder to generate the final representations. We closely follow the base implementation originally proposed in Wav2Vec; however, we use a shallower feature extractor with convolutional blocks of 512 channels with kernel widths of (10, 3, 3, 3) and strides of (5, 2, 2, 2), respectively. The internal dimension is 2048 and we choose a 128-dimensional space for our final representation.

4.4. Learning Methods

Supervised Learning We adopt the REI and XEI backbones in building a lead-I ECG supervised classifier model. As presented above, both backbones learn a 1x128 representation vector from the lead-I ECG input. The representation from the backbone encoders is passed on to two fully connected dense layers, followed by a ReLU layer that outputs the probability of the presence of CD in comparison to normal ECG, completing the end-to-end supervised classification architecture.

These supervised models are used as the performance baselines for our proposed self-supervised models, REI and XEI.

Self-supervised Learning As our core contribution, we implement two self-supervised architectures following the contrastive learning standard method, SimCLR (Chen et al. (2020)). We contrast two non-overlapping 5-second segments of the same 10-second ECG, thus ensuring both ECG views belong to the same patient and continuous instances. The goal of contrastive learning is to learn useful representations of the input data by maximizing agreement between embeddings from the same data. This process involves using a backbone to generate representations, a projection head to produce embeddings, and a loss function to compare the embeddings. We use the REI and XEI models, described above, as the backbone with an added linear layer to achieve a 128-dimensional representation. The projection head consists of two linear layers with a ReLU activation between them to get a 64-dimensional embedding.

We employ normalized temperature-scaled cross-entropy loss (NT-Xent), as described in Chen et al. (2020). The loss is applied to the embeddings, obtained after the projection head. We utilize the representations for downstream task classification, discarding the projection head after training.

The 5-second ECG segments are shuffled before being fed into the model so that the model does not inherently know which segment precedes the other. This shuffling introduces variability, helping the model learn robust and invariant representations of the ECG data. The contrastive model is trained using these ECG segments until the minimum validation loss is achieved, ensuring optimal learning of the ECG representations. This optimal point refers to the setting when the representations of ECGs from the same sample are closely aligned, while the representations from other samples are further away in the representation space. Once the training for this contrastive model is completed, we conduct linear probing experiments for the tasks of interest.

PCLR As an external baseline for ECG representations, we compare our proposed REI and XEI representations to PCLR, a widely used and publicly available ECG pretrained model (Diamant et al. (2022)). PCLR was built in a similar contrastive learning manner using SimCLR, but the ECG views were restricted only to belong to the same patient, irrespective of the time of the acquisition. We use the pretrained model to get the ECG representations for the PTB-XL data and conduct the linear probing to evaluate in the same manner as our REI and XEI models.

Finally, we compare the performances of PCLR representations with our implemented models.

4.5. Training Details

Pretraining From the 3.6 million ECG from the MGH dataset, we use an 85-15 split: 3.1 million ECG for training and 500 thousand for validation of the contrastive pretraining pipelines. We optimize the learning rate and temperature hyperparameters to achieve the best performance.

Linear Probing For linear probing, the weights of the model are frozen, and the model is then used to obtain 128-dimensional representations of the data from the PTB-XL dataset. A logistic regression model is trained on these representations to classify positive and negative samples. For the PTB-XL dataset, we use the published splits, sub-selecting the 15,133 patients within our cohort criteria.

Supervised Training In the supervised setup, we train the models with an input of a 5-second 12-lead ECG to a classification prediction and a corresponding binary label (1 = presence of CD, 0 = normal) with binary cross-entropy loss until a minimum validation loss is achieved. We conducted a hyperparameter search over the learning rate for optimal results. We use the same PTB-XL training, validation, and holdout test sets as were used in the linear probing.

For supervised training, the best performance is achieved with a learning rate of 0.002 for REI and 0.0004 for XEI. In contrast, self-supervised training benefits most from a lower learning rate of 0.0001 and a temperature of 0.23 for XEI and 0.24 for REI.

All models are implemented using Pytorch and Pytorch Lightning, and hyperparameter sweeps are conducted using Hydra.

4.6. Evaluation

The discriminatory ability of the modeling approach is evaluated with the Area Under the receiver-operating Curve (AUC), sensitivity, specificity, and the predictive values on the holdout test set from the PTB-XL dataset.

5. Results

5.1. Heart Block Identification

Notably, the self-supervised pipelines on the lead-I ECG perform significantly better than their supervised counterparts, as shown in Table 3. The calibration of the models for CD classification can be seen in Figure 1, and the calibration of the models for individual heart block classification can be seen in Appendix A. The statistical significance of this difference is reported in Appendix B. Moreover, in comparison to the baseline PCLR representations, our proposed continuity-based representations improve the models' discriminatory abilities by a large scale, 0.91 AUC for XEI compared to 0.64 for PCLR. Between the two backbone architectures, REI and XEI, we observe very small improvements for the transformer model in comparison to the residual network pipeline. The improvement in AUC in the models with contrastive pretraining further emphasizes the advantage of self-supervised learning in scenarios where balanced performance is critical, especially when sensitivity is a key clinical requirement. Moreover, we also implement and compare a 12-lead pipeline to quantify the performance degradation due to the reduction in input dimensions from 12 channels to single channels. We highlight that only 5% improvement is observed in response to changing the input to 12-lead ECG, which is a reasonable bound considering the widespread potential of outpatient monitoring.

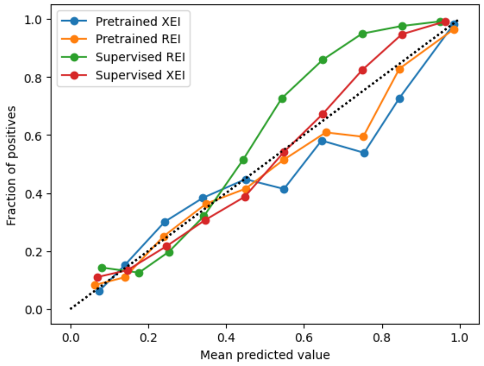

Figure 1: Calibration of models for CD Classification

5.2. Clinical Application

To simulate a clinical use case, we design a multi-task classification objective where the models try to discriminate the subcategories of heart block. From the dataset, we leverage the labels for LBBB, RBBB, AVB, FB, WPW, and IVCD subcategories of heart block against the normal ECG. We compare both REI and XEI pipelines and evaluate their performance for

Table 3: AUROC Across Learning Approaches on CD Classification

Model	Training	AUC[1]	Sensitivity	Specificity
REI	pretrain & linear probe	0.90 ± 0.01	0.78 ± 0.04	0.88 ± 0.02
XEI	pretrain & linear probe	0.91 ± 0.01	0.79 ± 0.04	0.86 ± 0.02
REI	supervised	0.87 ± 0.01	0.80 ± 0.01	0.51 ± 0.01
XEI	supervised	0.88 ± 0.02	0.80 ± 0.01	0.71 ± 0.01
PCLR	pretrain & linear probe	0.64 ± 0.01	0.74 ± 0.04	0.25 ± 0.03
12-lead Resnet	pretrain & linear probe	0.96 ± 0.00	0.90 ± 0.03	0.88 ± 0.03

[1] Bootstrapping was used to determine CI. Data was sampled with replacement for 100 iterations.

each task independently. The AUC for individual heart clock classification tasks is shown in Table 4 and sensitivity, specificity, and predictive values can be found in Appendix C.

5.3. Resilience against data scarcity

To provide further insights into the resilience and adaptability of these methods, we conduct experiments where the amount of training data is progressively reduced, while the size of the holdout test set remains constant. The results for REI and XEI are summarized in Figure 2. In these experiments, the AUC for the self-supervised model with linear probing remains consistent, whereas the performance of the supervised model declines significantly as the training set shrinks. This analysis is crucial for understanding which methods are more resilient to limited data, a frequent obstacle in clinical environments.

By examining model performance under constrained training conditions, we aim to underscore the potential of self-supervised learning, particularly self-supervised training with linear probing, in harnessing large-scale datasets to build more generalizable ECG classifiers. These findings are especially relevant for improving diagnostic capabilities in real-world clinical settings, where labeled data is often scarce, further emphasizing the robustness of self-supervised approaches in such contexts.

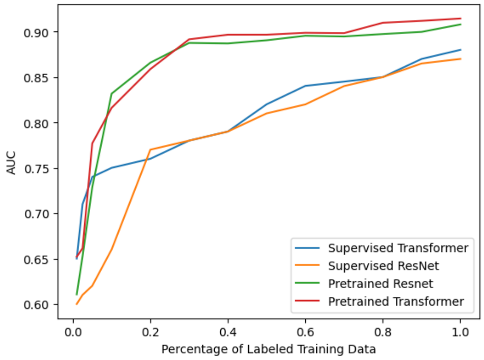

Figure 2: Holdout performance of REI and XEI with self-supervised pretraining and linear probing compared to supervised training across varying training set sizes.

5.4. Identifying Heart Block Categories

In this comparison of learning approaches across various heart block classification tasks, we observe an improved AUC for models pretrained with self-supervised, contrastive learning methods. Table 4 presents the performance metrics on individual heart block classification tasks. Notably, AUC remains higher for both the REI and XEI models when pretrained and followed by linear probing, showing superior performance compared to the supervised learning approach. The pretrained models achieved AUC values close to 91%, a reasonable improvement over the supervised versions.

Further observation of sensitivity and specificity illustrates that the linear probing methods show a higher sensitivity than their supervised counterparts. Results can be seen in Table 3 for general heart block identification and results for individual heart block categories can be found in Appendix C. The sensitivity for both REI and XEI remains high, around 89% and 88%, respectively, while specificity is around 72%. In contrast, the supervised models exhibit a significant drop in sensitivity, particularly for REI, which falls to 65%, though they achieve higher specificity, reaching 94%. Full sensitivity, specificity, and predictive value results can be found in Appendix C.

Table 4: AUROC on Individual Heart Block Categories

Model	Training	WPW	RBBB	LBBB	IVCD	AVB	FB
XEI	pretrained	0.97 ± 0.02	0.89 ± 0.02	0.99 ± 0.00	0.84 ± 0.02	0.94 ± 0.01	0.83 ± 0.02
REI	pretrained	0.96 ± 0.02	0.88 ± 0.02	0.99 ± 0.00	0.84 ± 0.02	0.93 ± 0.01	0.82 ± 0.02
REI	supervised	0.82 ± 0.00	0.75 ± 0.01	0.93 ± 0.01	0.78 ± 0.01	0.77 ± 0.00	0.80 ± 0.03
XEI	supervised	0.87 ± 0.02	0.78 ± 0.01	0.98 ± 0.00	0.82 ± 0.01	0.84 ± 0.00	0.83 ± 0.02

6. Insights

This work provides several key insights into the application of self-supervised learning for ECG analysis, particularly in the context of heart block detection:

Representation Quality: The superior performance of models pretrained with self-supervised, contrastive learning methods suggests that these approaches learn more robust and generalizable ECG representations. This is evidenced by the consistently higher AUC values achieved by the pretrained models across various heart block classification tasks.

Data Efficiency: As illustrated in Figure 2, the self-supervised models maintain consistent performance even with reduced training data, while the supervised models' performance declines significantly. This resilience to data scarcity is a significant advantage in medical contexts where large labeled datasets are often challenging to obtain.

Model Architecture Influence: Both REI and XEI architectures show improvements with self-supervised pretraining, indicating that the benefits of this approach are not limited to a specific model structure. However, the extent of improvement varies between architectures, suggesting that some models may be better suited to leverage self-supervised learning.

Task-Specific Performance: The improvements in AUC are not uniform across all types of heart block. For instance, the improvement is particularly notable for WPW detection, while less pronounced for the RBBB. This variability suggests that the effectiveness of self-supervised learning may depend on the specific characteristics of each heart block type.

7. Conclusion

The improved AUC values and sensitivity of self-supervised models, coupled with their resilience to reduced training data, address critical challenges in clinical ECG analysis. These properties are particularly important for rare disease detection and resource-constrained healthcare settings where labeled data is scarce or costly to obtain. This approach opens avenues for developing comprehensive multi-task models capable of detecting various cardiac abnormalities simultaneously. While our results are encouraging, they also highlight areas for future research, such as integrating these models into existing clinical workflows to maximize their practical impact and investigating the predictive capabilities on a wider range of cardiac abnormalities.

By addressing these challenges, we can further refine and validate these techniques across diverse clinical settings and patient populations to work towards more accurate, efficient, and accessible cardiac care.

References

Ridwan Alam et al. Detecting QT prolongation from a single-lead ECG with deep learning. *PLOS Digital Health*, 3(6), 2024a.

Ridwan Alam et al. Estimating ecg intervals from lead-i alone: External validation of supervised models. *medRxiv*, 2024b.

Erick Alday et al. Classification of 12-lead ECGs: The Physionet/Computing in Cardiology challenge 2020. *Physiological Measurement*, 41(12), 2020.

Zachi I Attia et al. An artificial intelligence-enabled ecg algorithm for the identification of patients with atrial fibrillation during sinus rhythm: a retrospective analysis of outcome prediction. *The Lancet*, 394(10201), 2019.

Alexei Baevski et al. wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in neural information processing systems*, 33, 2020.

Karim Bayoumy et al. Smart wearable devices in cardiovascular care: where we are and how to move forward. *Nature Reviews Cardiology*, 18(8), 2021.

Teya Bergamaschi et al. Heart block identification from 12-lead ecg: Exploring the generalizability of self-supervised AI. *medRxiv*, pages 2024–10, 2024.

Loc Cao et al. Robustness of residual network in predicting PR interval trained using noisy labels. In *CinC*, volume 498. IEEE, 2022.

Ting Chen et al. A simple framework for contrastive learning of visual representations. In *Int Conf on Machine Learning (ICML)*, pages 1597–1607. PMLR, 2020.

Nathaniel Diamant et al. Patient contrastive learning: A performant, expressive, and practical approach to electrocardiogram modeling. *PLoS computational biology*, 18(2):e1009862, 2022.

John R Giudicessi et al. Artificial intelligence–enabled assessment of the heart rate corrected qt interval using a mobile electrocardiogram device. *Circulation*, 143(13), 2021.

Ary Goldberger et al. *Clinical Electrocardiography: A Simplified Approach*. Elsevier Health Sciences, 2017.

Awni Hannun et al. Cardiologist-level arrhythmia detection and classification in ambulatory electrocardiograms using a deep neural network. *Nature Medicine*, 25(1), 2019.

Kaiming He et al. Deep residual learning for image recognition. *IEEE CVPR*, pages 770–778, 2016.

Anthony Kashou et al. *Atrioventricular Block*. StatPearls, 2023.

Akshay Khunte et al. Detection of left ventricular systolic dysfunction from single-lead electrocardiography adapted for portable and wearable devices. *npj Digital Medicine*, 6(1):124, 2023.

Dani Kiyasseh et al. CLOCS: Contrastive learning of cardiac signals across space, time, and patients. In *ICML*. PMLR, 2021.

Wei-Yin Ko et al. Detection of hypertrophic cardiomyopathy using a convolutional neural network-enabled electrocardiogram. *JACC*, 75(7), 2020.

Alexander Kolesnikov et al. Big transfer (BiT): General visual representation learning, 2020. URL https://arxiv.org/abs/1912.11370.

Joon-myoung Kwon et al. Artificial intelligence algorithm for predicting cardiac arrest using electrocardiography. *Scandinavian journal of trauma, resuscitation and emergency medicine*, 28:1–10, 2020.

Baptiste Maille et al. Smartwatch electrocardiogram and artificial intelligence for assessing cardiac-rhythm safety of drug therapy in the covid-19 pandemic. the qt-logs study. *International Journal of Cardiology*, 331, 2021.

Temesgen Mehari et al. Self-supervised representation learning from 12-lead ecg data. *Computers in Biology and Medicine*, 141, 2022.

Annamalai Natarajan et al. A wide and deep transformer neural network for 12-lead ecg classification. In *Computing in Cardiology (CinC)*, pages 1–4. IEEE, 2020.

Matthew Reyna et al. Will two do? varying dimensions in electrocardiography: the physionet/computing in cardiology challenge 2021. In *CinC*, volume 48. IEEE, 2021.

Antônio Ribeiro et al. Automatic diagnosis of the 12-lead ECG using a deep neural network. *Nature Comm*, 11(1), 2020.

Furrukh Sana et al. Wearable devices for ambulatory cardiac monitoring: Jacc state-of-the-art review. *JACC*, 75(13), 2020.

Veer Sangha et al. Biometric contrastive learning for data-efficient deep learning from electrocardiographic images. *JAMIA*, 31(4), 2024.

Daphne E Schlesinger et al. A deep learning model for inferring elevated pulmonary capillary wedge pressures from the 12-lead electrocardiogram. *JACC: Advances*, 1(1), 2022.

Nils Strodthoff et al. Deep learning for ECG analysis: Benchmarks and insights from PTB-XL. *JBHI*, 25 (5), 2020.

Nicholas Y Tan et al. Left bundle branch block: current and future perspectives. *Circulation: Arrhythmia and Electrophysiology*, 13(4):e008239, 2020.

Connie Tsao et al. Heart disease and stroke statistics—2023 update: A report from the american heart association. *Circulation*, 147, 2023.

Patrick Wagner et al. PTB-XL, a large publicly available electrocardiography dataset. *Scientific Data*, 7(1), 2020.

Appendix A. Calibration Plots

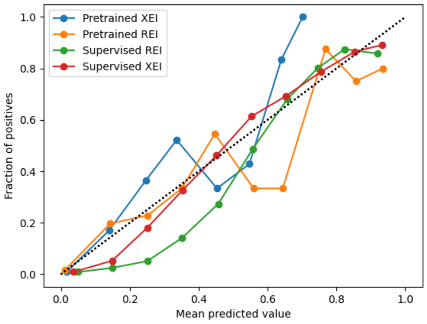

Figure 3: Calibration plot for AVB.

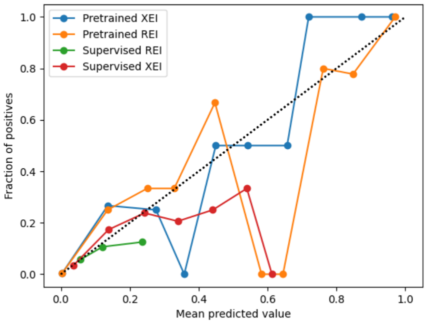

Figure 4: Calibration plot for RBBB.

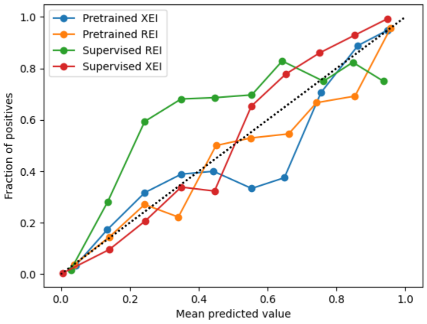

Figure 5: Calibration plot for LBBB.

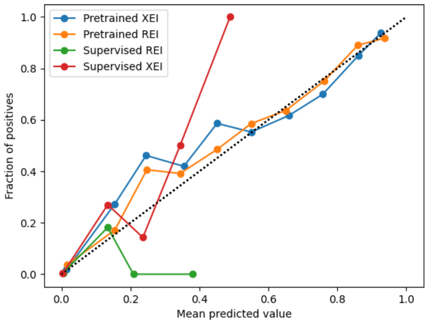

Figure 6: Calibration plot for FB.

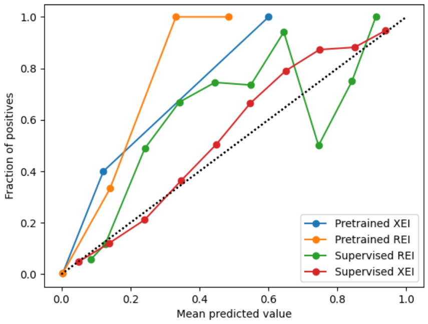

Figure 7: Calibration plot for WPW.

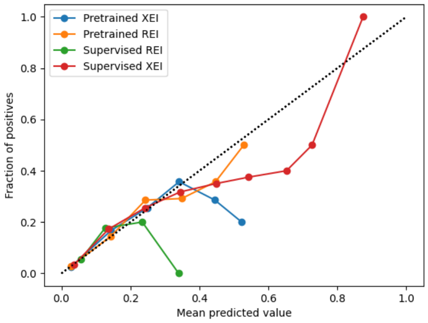

Figure 8: Calibration plot for IVCD.

Appendix B. Statistical Testing

Model Comparison	CD	FB	WPW	RBBB	LBBB	IVCD	AVB
Pretrained REI vs. Pretrained XEI	0.307	0.538	0.707	0.269	0.726	0.138	0.018
Supervised REI vs. Supervised XEI	0.001	0.001	0.001	0.000	0.000	0.001	0.001
Pretrained REI vs. Supervised XEI	0.001	0.044	0.001	0.000	0.001	0.057	0.041
Supervised REI vs. Pretrained XEI	0.000	0.001	0.001	0.000	0.000	0.001	0.001
Pretrained REI vs. Supervised REI	0.001	0.001	0.001	0.002	0.000	0.001	0.001
Pretrained XEI vs. Supervised XEI	0.002	0.016	0.001	0.000	0.000	0.025	0.032

Table 5: P-value for paired t-tests for model comparisons.

Appendix C. Performance on Individual Heart Block Categories

Model	Training	Sensitivity	Specificity	PPV	NPV
XEI	pretrained	0.79 ± 0.04	0.86 ± 0.02	0.75 ± 0.04	0.88 ± 0.02
REI	pretrained	0.78 ± 0.04	0.88 ± 0.02	0.78 ± 0.04	0.88 ± 0.02
REI	supervised	0.80 ± 0.01	0.51 ± 0.01	0.48 ± 0.01	0.82 ± 0.01
XEI	supervised	0.80 ± 0.01	0.71 ± 0.01	0.60 ± 0.01	0.86 ± 0.01

Table 6: Sensitivity, Specificity, PPV, and NPV for CD with confidence intervals.

Model	Training	Sensitivity	Specificity	PPV	NPV
XEI	pretrained	0.78 ± 0.03	0.84 ± 0.03	0.90 ± 0.02	0.67 ± 0.04
REI	pretrained	0.78 ± 0.03	0.86 ± 0.03	0.91 ± 0.02	0.67 ± 0.04
REI	supervised	0.73 ± 0.11	0.68 ± 0.01	0.01 ± 0.00	1.00 ± 0.00
XEI	supervised	0.68 ± 0.11	0.83 ± 0.01	0.02 ± 0.01	1.00 ± 0.00

Table 7: Sensitivity, Specificity, PPV, and NPV for FB with confidence intervals.

Model	Training	Sensitivity	Specificity	PPV	NPV
XEI	pretrained	1.00 ± 0.16	0.82 ± 0.02	0.03 ± 0.02	1.00 ± 0.00
REI	pretrained	0.88 ± 0.22	0.91 ± 0.02	0.05 ± 0.04	1.00 ± 0.00
REI	supervised	0.79 ± 0.02	0.46 ± 0.01	0.17 ± 0.01	0.94 ± 0.01
XEI	supervised	0.79 ± 0.02	0.65 ± 0.01	0.24 ± 0.01	0.96 ± 0.00

Table 8: Sensitivity, Specificity, PPV, and NPV for WPW with confidence intervals.

Model	Training	Sensitivity	Specificity	PPV	NPV
XEI	pretrained	0.83 ± 0.06	0.75 ± 0.02	0.32 ± 0.04	0.97 ± 0.01
REI	pretrained	0.80 ± 0.06	0.80 ± 0.02	0.35 ± 0.05	0.97 ± 0.01
REI	supervised	0.80 ± 0.03	0.91 ± 0.01	0.29 ± 0.02	0.99 ± 0.00
XEI	supervised	0.79 ± 0.03	0.99 ± 0.00	0.82 ± 0.03	0.99 ± 0.00

Table 9: Sensitivity, Specificity, PPV, and NPV for RBBB with confidence intervals.

Model	Training	Sensitivity	Specificity	PPV	NPV
XEI	pretrained	0.81 ± 0.10	1.00 ± 0.00	0.98 ± 0.05	0.99 ± 0.01
REI	pretrained	0.81 ± 0.10	1.00 ± 0.00	0.93 ± 0.07	0.99 ± 0.01
REI	supervised	0.80 ± 0.03	0.46 ± 0.01	0.08 ± 0.01	0.97 ± 0.00
XEI	supervised	0.77 ± 0.03	0.59 ± 0.01	0.10 ± 0.01	0.98 ± 0.00

Table 10: Sensitivity, Specificity, PPV, and NPV for LBBB with confidence intervals.

Model	Training	Sensitivity	Specificity	PPV	NPV
XEI	pretrained	0.83 ± 0.08	0.70 ± 0.03	0.14 ± 0.03	0.99 ± 0.01
REI	pretrained	0.82 ± 0.09	0.70 ± 0.02	0.14 ± 0.03	0.98 ± 0.01
REI	supervised	0.79 ± 0.03	0.46 ± 0.01	0.08 ± 0.01	0.97 ± 0.00
XEI	supervised	0.78 ± 0.03	0.71 ± 0.01	0.14 ± 0.01	0.98 ± 0.00

Table 11: Sensitivity, Specificity, PPV, and NPV for IVCD with confidence intervals.

Model	Training	Sensitivity	Specificity	PPV	NPV
XEI	pretrained	0.81 ± 0.09	0.90 ± 0.02	0.32 ± 0.06	0.99 ± 0.01
REI	pretrained	0.67 ± 0.10	0.94 ± 0.01	0.40 ± 0.08	0.98 ± 0.01
REI	supervised	0.80 ± 0.01	0.61 ± 0.01	0.79 ± 0.01	0.63 ± 0.01
XEI	supervised	0.79 ± 0.01	0.74 ± 0.01	0.85 ± 0.01	0.67 ± 0.01

Table 12: Sensitivity, Specificity, PPV, and NPV for AVB with confidence intervals.