ELECTROCARDIOGRAM FOUNDATION MODEL USING TEMPORALLY AUGMENTED PATIENT CONTRASTIVE LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

Electrocardiograms ($ECGs$) capture the electrical activity of the heart, offering rich diagnostic and prognostic insights. Traditionally, electrocardiograms are interpreted by human experts, but deep learning is now encroaching on this domain and combining human-like intelligence with machine precision for a deeper insight. Self-supervised pretraining is essential for maximising the potential of scarce medical data. Applied to $ECGs$, patient-contrastive learning has shown promising results, by utilising the natural variations in the cardiac signals. In this study, we introduce **T**emporally **A**ugmented **P**atient **C**ontrastive **L**earning of **R**epresentations (TA-$PCLR$), a novel approach that incorporates temporal augmentations into a patient contrastive self-supervised foundation model. Trained on one of the largest diverse cohorts of more than six million unlabelled electrocardiograms from three continents, we demonstrate the efficacy of our approach and show its value as a feature extraction tool for small and medium-sized labeled datasets. We also validate the performance on an open-source external cohort, surpassing other pretraining approaches while outperforming an ensemble of fully supervised deep networks on some labels. Additionally, we conduct a detailed exploration of how the pretraining and labeled electrocardiogram dataset distributions impact supervised task performance.

1 INTRODUCTION

Electrocardiograms ($ECGs$) record cardiac electrical activity as a graph of voltage versus time. Electrodes are placed on the body surface to detect the electrical changes resulting from cardiac muscle depolarization and repolarization. The standard electrode positions define twelve different leads (signals between two specific electrodes) representing cardiac activity along twelve axes, essential for localising the underlying processes. The heart's electrical activity has been recorded since 1887 (Waller, 1887) and has been an important source of information for cardiologists and physicians. Human understanding has since evolved to interpret ECG patterns as manifestations of different health conditions. Being a simple non-invasive investigation, it is part of routine medical care although the expertise for accurate interpretation is not so readily available.

Deep learning has excellent pattern recognition capabilities, surpassing any other methodology, and holds great potential for medical sciences (Esteva et al., 2019). The traditional ECG interpretation depends on the distinct patterns of the waveforms combined with an understanding of the heart function and clinical observations. Human perception is limited by low visual accuracy, gaps in theoretical knowledge, and the complexity of the diverse, non-linear, interrelations (Strodthoff et al., 2021). Deep learning has demonstrated high precision in predicting cardiac diseases (Sau et al., 2023; 2024; Pastika et al., 2024) while opening the possibility of predicting novel labels like age and sex that are beyond human capabilities (Attia et al., 2019). Artificial intelligence has the potential to surpass human capabilities while allowing for accurate diagnosis and risk stratification.

Contrastive learning, a self-supervised pretraining approach, can greatly improve the accuracy for subsequent supervised tasks, especially where the labeled datasets are quite small (Chen et al., 2020). Contrastive learning extracts meaningful representation using a notion of positive and negative instances (Chen et al., 2020). Representations of positive instances are trained to be more similar

and distinct from representations of negative instances. There are diverse contrastive learning approaches, mostly differing in their definitions of the positive and negative instances and the loss computations. In the visual image domain, the positives are usually augmentations (transformations) of the same image while the negatives are augmentations from others. The training retains the meaningful features shared between the positives while discarding the extraneous features.

Contrastive learning has been leveraged for feature extraction from biomedical data with various definitions of data augmentations (Mohsenvand et al., 2020). The augmentations applied to medical data should preserve clinically significant information. Data from the same patient across time is a robust way to encode trivial transformations (Jamaludin et al., 2017; Diamant et al., 2022), thereby allowing the model to reject any artifacts arising from instrument noise, different lead locations, patient movement, etc. Combining different types of random augmentations has remarkably enhanced the performance for contrastive learning (Chen et al., 2020). Incorporating temporal augmentations with patient-based contrastive learning is a key novel feature of our work. We hypothesise that multiple transformations enhance the quality of representations and the selection of clinically insignificant augmentations is essential for the generalisation of the approach. We thus improve on existing patient contrastive methods, by adding temporal augmentation like zero-masking (Soltanieh et al., 2022) and random cropping for our Temporally Augmented Patient Contrastive Learning for ECG Representations ($TA\text{-}PCLR$). In the medical domain, labeled data is often highly scarce, and thus pretraining approaches are essential for maximum utilisation of available data. Foundation models can be pretrained on large unlabeled datasets, to learn general features that can be leveraged for a range of subsequent supervised tasks (Zhang & Metaxas, 2024). We train a foundation model on $6,174,025$ $ECGs$ to improve generalisation for ECG feature extraction and explore the effect of pretraining data on the quality of the learned representations. The main contribution of our work is to present a new foundation model for ECG interpretation. Our model achieves state-of-the-art results due to:

- a new method: Temporally Augmented Patient Contrastive learning that incorporates augmentations along the temporal axis.
- a new multi-center dataset we constructed for training with over six million $ECGs$ with four cohorts from three different continents.

Our augmentation techniques improve on previous contrastive learning techniques based on contrasting between exams from the same patient (Diamant et al., 2022). Our newly constructed dataset allows us to showcase, the effect of training data demographics, on the quality of the learned representations. We present a foundation model that outperforms other, much larger, self-supervised foundation models (Song et al., 2024) and is close to highly optimised fully supervised benchmarks (Strodthoff et al., 2021). Section 2 presents a brief overview of the past efforts conducted for the contrastive learning of electrocardiograms. Section 3 describes the implementation details of our approach, followed by the performance analysis in Section 4. Section 5 concludes with the highlights of our research findings and presents some suggestions for future research.

2 Literature Background

Contrastive learning Contrastive learning has shown remarkable improvement for image classification tasks, by incorporating a self-supervised pretraining (Hadsell et al., 2006). In the computer vision domain, pretraining enhances the similarity of the augmented versions of the same image (positives) while reducing that from distinct images (negatives). Efficient techniques such as $InfoNCE$ (van den Oord et al., 2018) present a robust encoding of the contrastive loss, and the $SimCLR$ (Chen et al., 2020) introduces the idea of a non-linear projection layer for performance improvement.

Data augmentations Data augmentations are essential for contrastive learning to enhance meaningful context and reject spurious information. The augmentations of visual images can be performed by using different segments, views, and coloring of the original image, retaining the meaningful context (Chen et al., 2020). The dimensions for augmentation are more limited for the time series data and mainly involve noise addition, time-masking, cropping, shuffling, inverting, etc (Wen et al., 2021). Previous work has shown that good augmentations are crucial to preserving only meaningful features and greatly affect the generalisation for subsequent supervised tasks. The diversity

in negative and positive examples is also essential for learning meaningful features. Combining multiple augmentations has been proved to reinforce the learning process (Chen et al., 2020; Gopal et al., 2021). For medical data, augmentations are also restricted by possible clinical implications, therefore synthetic augmentations have to be taken with care. Past work has shown that while some augmentations might improve the performance on one task, they may have adverse effects on others (Lee et al., 2022; Raghu et al., 2022).

Contrastive learning for ECGs Contrastive learning has been employed for ECG representation learning. $CLOCS$ (Kiyasseh et al., 2021) presents the idea of using ECG from the same patient as a meaningful context, with the different leads and non-overlapping slices from the same ECG as positives, thereby incorporating multiple positive ECGs in the batch, thus improving performance over the $SimCLR$ baseline (Chen et al., 2020). $PCLR$ (Diamant et al., 2022) takes the concept further defining ECG from the same patient over time as positive instances and demonstrates superior performance compared to previous approaches. The contrastive heartbeats (CT-HB) (Wei et al., 2022) splits the individual heartbeats from an ECG recording and defines heartbeats from the same ECG as positives and implements a variant of triplet loss (Wang et al., 2019). Soltanieh et al. (2022) systematically explores a spectrum of time series augmentations for $ECGs$ including time-warping, permutation (slice and shuffle), inverting, and scaling, which nonetheless could have clinical implications. Physiologically-inspired spatial and temporal augmentations including axis rotation, scaling, and zero-masking, are combined by the $3KG$ (Gopal et al., 2021) for self-supervised pretraining, improving ECG classification performance. $ECG - FM$ (McKeen et al., 2024) employs a multi-layer convolutional feature extractor and a transformer encoder for feature extraction with random-lead-masking augmentation. The joint cross-dimensional contrastive learning approach (Liu et al., 2023) is based on learning ECG representation by contrasting ECG signals against images incorporating several modes of augmentations. $MERL$ (Liu et al., 2024) contrasting $ECGs$ with clinical reports to provide the possibility of zero-shot inference.

Generative pretraining: Self-supervised pretraining has been implemented following generative approaches, such as masked autoencoders (MAE) reconstructing random masked ECG segments (Gedon et al., 2021; Na et al., 2024). Hybrid techniques combining contrastive learning and generative pretraining based on transformer architecture have been implemented for ECG feature extraction (Song et al., 2024).

Foundation models Foundation models are defined by the flexibility to facilitate generic downstream tasks by exploiting huge pre-training cohorts (Zhang & Metaxas, 2024). All of the self-pretraining methodologies have the potential to adapt to any generic task and large cohorts can further enhance the model capabilities. HeartBeiT (Vaid et al., 2023) exploits vision transformer architecture to present an ECG-based foundational model. (Song et al., 2024) trained a foundation ECG model on more than a million $ECGs$ exploiting a hybrid approach combining contrastive learning with generative pretraining involving vision transformers. $ECG - FM$ (McKeen et al., 2024) employs $wav2vec$ architecture with a convolutional feature extractor with a BERT-like transformer encoder trained on 1.6 million ECGs. Foundation models have also been developed following a supervised approach (Li et al., 2024) where the generalization capabilities may be limited.

Our approach Our work is a natural extension and improvement of these efforts. We combine patient-based augmentation with simple temporal augmentations based on random zero-masking and cropping without impacting the underlying clinical information, thus making the approach robust. We further train on a large diverse cohort for improved generalization and explore the impact of the pretraining data demographics on the learned representations. We also demonstrate that label-based performance comparisons for diverse datasets may not reflect the true merits of an approach.

3 Materials and Methods

3.1 Cohorts

The study employs a range of large, diverse ECG cohorts from three continents: Beth Israel Deaconess Medical Center ($BIDMC$) (Pastika et al., 2024) from the United States, Clinical Outcomes in Digital Electrocardiography ($CODE$) (Ribeiro et al., 2019) from Brazil, Shanghai Zhong-

shan Hospital cohort dataset ($SHZS$) from China, Vanderbilt University Medical Center cohort ($VUMC$) (Aras et al., 2023) from United States, UK Biobank (UKB) (Sudlow et al., 2015) from United Kingdom and Physikalisch-Technische Bundesanstalt (PTB-XL) dataset (Wagner et al., 2020) from Germany. Table 1 presents the data used in current research in terms of the number of unique patients with more than one ECG and the corresponding number of $ECGs$ for the contrastive learning pretraining cohorts: $BIDMC$, $CODE$, $VUMC$, and $SHZS$, and the total $ECGs$ for datasets used in performance validation: UKB and PTB-XL. We denote the combined pretraining cohort as $BCSV$ consisting of more than six million individual ECGs, with each ECG comprising eight leads. Appendix A provides important information about the dataset demographics in Table 6, while further details can be obtained from the corresponding references.

Table 1: Datasets

No.	Cohorts	Patients*	ECGs
1	$BIDMC$ (United States) (Pastika et al., 2024)	$127,055$	$1,106,900$
2	$CODE$ (Brazil) (Ribeiro et al., 2019)	$424,577$	$1,123,903$
3	$SHZS$ (China)	$420,957$	$2,257,485$
4	$VUMC$ (United States) (Aras et al., 2023)	$252,306$	$1,685,737$
5	$BIDMC$+$CODE$+$SHZS$+$VUMC$ ($BCSV$)	$1,224,895$	$6,174,025$
6	UKB (United Kingdom) (Sudlow et al., 2015)	-	$70,655$
7	PTB-XL (Germany) (Wagner et al., 2020)	-	$21,800$

*Unique patients with more than one ECG

3.2 Contrastive Loss

The contrastive loss employed for the current work is the $InfoNCE$ loss (van den Oord et al., 2018) applied to the non-linear projections similar to $SimCLR$ (Chen et al., 2020). Given that z_i and z_j are the non-linear projections of representations from two different augmented $ECGs$ belonging to the same patient, the similarity between z_i and z_j is enhanced over all other instances in the batch by applying a $softmax$ (Bridle, 1989) over the similarity values. The loss function then implements the following equation 1 where τ is a temperature coefficient defining how soft or hard the $softmax$ constrains the similarity distributions and N denotes the number of pairs in the batch. The $\mathbb{I}_{[k \neq i]} \in 0, 1$ is an indicator function evaluating to 1 only if $k \neq i$.

$$\ell_{i,j} = -\log \frac{\exp(\text{sim}(z_i, z_j)/\tau)}{\sum_{k=1}^{2N} \mathbb{I}_{[k \neq i]} \exp(\text{sim}(z_i, z_k)/\tau)} \tag{1}$$

Where sim is the cosine similarity defined as:

$$\text{sim}(z_i, z_j) = \frac{z_i^\top \cdot z_j}{||z_i|| \cdot ||z_j||} \tag{2}$$

3.3 Augmentation

Generalization is an important aspect of a foundation model, we refrain from scaling, rotating, or frequency-warping employed in some past pre-training approaches (Soltanieh et al., 2022) that may potentially impact the performance for any unforeseen future supervised tasks where the scale, axis, or rhythm may be important information (Raghu et al., 2022). We reason that contrasting for patient identity (Diamant et al., 2022) is a natural way of encoding trivial transformation and we combine it with temporal augmentations, such as zero-masking (Gopal et al., 2021; Soltanieh et al., 2022) and random cropping thus ensuring that the pretraining will be relevant for any potential future supervised task. We define the positive views as augmentations of random slices from different $ECGs$

of the same patient. The number of $ECGs$ from each patient greatly differs thus the training epoch is defined as one complete iteration for all unique patients with the positive views randomly sampled at training time. In this way, the training is not biased by patients having more $ECGs$, while exploiting the available data diversity. In contrast to Diamant et al. (2022), $ECGs$ in a positive pair are always unique instances. The input window size of seven seconds allows random cropping (RC) by using a different patch from the same ten-second ECG for each epoch. The additional temporal augmentation includes zero-masking (Soltanieh et al., 2022; Raghu et al., 2022) that is unimpactful of any intrinsic clinical information. The transformations are applied at the training time so for the same ECG, the slices and masks are unique for each epoch, increasing the diversity of the positive samples. We experimented with applying zero-masking to the same random segments for each lead (RZM), different random segments for each lead ($RLZM$), masking random leads (RLM) (McKeen et al., 2024), and a novel notion of using the raw and filtered ECGs as augmentations (RF)[1]. We retained the simpler configuration of RZM, which showed the best performance.

3.4 Preprocessing

The standard procedure for ECG recording involves measurements from 12 leads recorded for 10 seconds with sampling rates typically at 400 to 500 samples per second (Hz). For model development, we used eight ECG leads as four leads are linear combinations of other leads and thus do not impart additional information (Eem et al., 2020). We apply a bandpass filter (0.5 to 100 Hz) and a notch filter relevant to the mains frequency and interpolate $ECGs$ from different sources to a standard sampling frequency of 400 Hz. We retain the original scale of the ECGs in millivolts. The final input shape to the contrastive learning model is 2800×8 (7 second signal).

3.5 Architecture

Figure 1 presents an overview of the $TA\text{-}PCLR$ architecture. The $ECGs$ for the same patient are treated as positive views while all other $ECGs$ in the batch are negative views. For a fair comparison, we use the backbone architecture from Ribeiro et al. (2020) with non-linear projections (similar to $PCLR$ (Diamant et al., 2022)). The contrastive loss is applied to the non-linear projections of the $ECGs$. The output of the model is 256 features or embeddings learned from the $ECGs$ that can be exploited for any downstream supervised training. The model is implemented in Tensorflow (Abadi et al., 2015) (2.10.1). The total number of parameters is less than 6 million and thus the model is much more compact, as compared to the transformer-based architectures like $ECG - FM$, having more than 300 million parameters (McKeen et al., 2024).

3.6 Training

The contrastive loss performs best with larger batches due to the larger variation of the negative instances (Chen et al., 2020), but the computation resources limit the batch size. We use a batch size of 1024 (512 patients) and train the model for 200 epochs using the Adam optimizer (Kingma & Ba, 2014). The initial learning rate is 0.1 and then decayed according to a half-period cosine schedule (Loshchilov & Hutter, 2016), similar to previous approaches (Chen et al., 2020; Diamant et al., 2022). The training time in minutes per epoch on NVIDIA GeForce RTX 3090 is approximately 5 for the $BIDMC$ dataset and 50 for the $BCSV$ dataset. This is notably shorter than comparative approaches in the literature taking weeks on multi-gpu configurations (Cheng et al., 2021; Song et al., 2024; McKeen et al., 2024).

4 Performance Analysis

The $TA\text{-}PCLR$ is a pretraining paradigm for unlabelled data, thus the performance of the learned embeddings is evaluated by subsequent supervised training similar to previous works (Chen et al., 2020). Following are the different performance evaluation configurations employed in the current work, depending on the experiment at hand:

[1]For RZM and $RLZM$ a 20% segment is masked while a 10% masking probability is applied for RLM. The RF configurations contrast raw vs. raw with 40%, filtered vs. filtered with 40%, and raw vs. filtered with 40% probability.

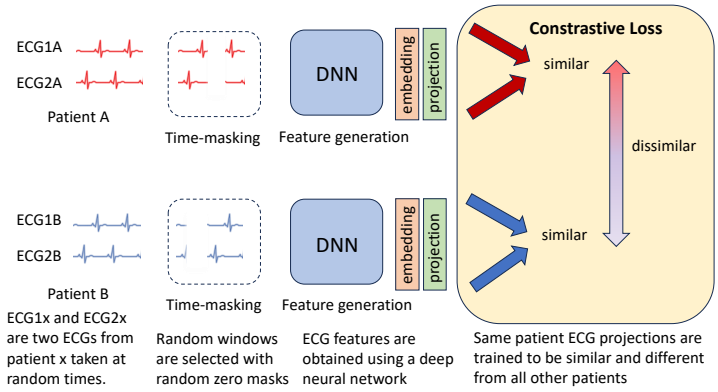

Figure 1: The $TA\text{-}PCLR$ pretraining overview. Patient A has $ECG1A$ and $ECG2A$ while Patient B is another patient in the same batch with $ECG1B$ and $ECG2B$. The $ECGs$ are transformed by temporal augmentations and converted by the network to a 256-feature vector. The contrastive loss brings the projections from the same patient $ECGs$ closer and apart from others in the batch.

- Linear evaluation: A linear probe is a standard methodology to compare the expressiveness of the features learned during unsupervised pretraining approaches (Chen et al., 2020). The feature-generating model is frozen while a single neuron is trained to predict each label.
- Multi-layer perceptron (MLP): We also train a two-layer MLP for the supervised task to allow the model to learn a non-linear mapping of the $TA\text{-}PCLR$ generated features.
- Fine-tuning: After training a linear model the feature-generating model can also be allowed to update and further improve the performance.

Validation is performed along two main dimensions: proof of concept by internal validation of the approach for different experimental configurations, and external validation by comparing to previous benchmarks. The test evaluation involves a single prediction from a random crop of each test ECG. We mainly report macro-averaged area under the receiver-operating characteristic curve $AUROC$ or AUC as a threshold-free assessment of classification performance and mean absolute error MAE in years for the age regression performance, as suggested previously (Wagner et al., 2020; Strodthoff et al., 2021). Unless otherwise specified, we report the mean of ten independent runs for each task.

4.1 Proof of concept

Experimental setup The $BIDMC$ dataset is primarily employed for model development and we demonstrate the superiority of our pretraining paradigm on this cohort. The supervised tasks for $BIDMC$ involve predicting an individual's age, sex, and probability of five-year mortality from an ECG. The self-supervised pretraining is performed on the $ECGs$ of the patients having more than two $ECGs$ (details in Table 1) with 80% of the patients included in the training and the rest in the validation set. The supervised tasks are then implemented using a train, validation, and test split of 50%, 10%, and 40% for $1,169,387$ labeled $ECGs$. The $TA\text{-}PCLR$ supervised tasks in this section use an MLP for the label prediction (details in Appendix D). The extracted features are standardized and the learning rate is manually optimised within the range of $[0.00001, 0.01]$.

Ablation study Our approach consists of patient-based (Diamant et al., 2022b) and temporal augmentations, together with our unique diverse dataset. Table 2 represents an ablation study highlighting the contribution of each component. The top row shows the performance of the patient-based contrastive learning ($PCLR$) pretrained on the $BIDMC$. The combined patient and temporal augmentations significantly improve the performance. We experiment with several augmentations including random cropping and zero-masking, as well as a novel raw-filtered augmentation with RZM resulting in the best performance. Finally, training with the combined $BCSV$ data further

Table 2: Ablation study exploring the contribution from each component of our approach on the supervised task for $BIDMC$ age, sex, and five-year mortality predictions. All the implementations use patient-based augmentations and all temporal augmentations include random cropping (RC). The additional augmentations tested include random lead zero masking (RLZ), raw-filtered (RF), random lead masking (RLM), and random zero masking (RZ). The best performance is indicated in bold while the second best is underlined.

No.	Augmentation	Pretraining cohort	Age (MAE)	Sex (AUC)	Mortality (AUC)
1	Patient-based	BIDMC	8.3061	0.9145	0.7729
2	Patient + RC + $RLZM$		7.8860	0.9338	0.7864
3	Patient + RC + RZM		<u>7.8498</u>	<u>0.9351</u>	0.7883
4	Patient + RC + RZM + RF		7.9018	0.9335	0.7873
5	Patient+ RC + RZM + RLM		7.9353	0.9318	<u>0.7900</u>
6	Patient + RC + RZM	$BCSV$	**7.7849**	**0.9393**	**0.7926**

enhances the performance, providing 6.27%, 2.71%, and 2.55% improvement for age, sex, and mortality, respectively. Hence, we demonstrate that temporal augmentations (on top of $PCLR$) enhance the model performance and an increased dataset size further improves task-specific efficacy.

How performance compares to supervised training for different train data sizes? Figure 2 compares the performance of the TA-$PCLR$ pretraining versus a randomly initialized $ResNet$ network with similar backbone used previously for ECG classification (Ribeiro et al., 2020). The test is conducted on a frozen feature-generating backbone with an MLP head, for different sizes of the labeled training data. The TA-$PCLR$ performance is reported as a mean of ten independent runs while the $ResNet$ is trained for a single run. The TA-$PCLR$ outperforms the $ResNet$ for smaller training data sizes up to 200 k when it converges. For age, sex, and mortality label prediction with 1000 training samples, the performance improvement is 15.84%, 16.94%, and 8.18%, respectively. It should be noted that TA-$PCLR$ performance can be further improved by fine-tuning. We thus demonstrate that the performance of TA-$PCLR$ is superior to fully-supervised $ResNet$ specifically for smaller datasets, which are often prevalent in the medical domain.

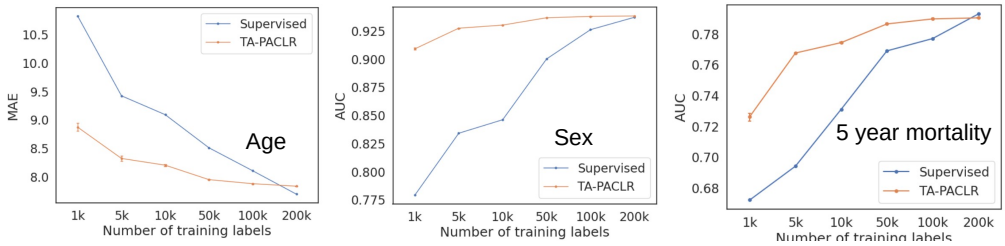

Figure 2: Comparison of TA-$PCLR$ (orange) with supervised learning (blue) for labels from the BIDMC dataset shows the remarkable improvement achieved through the proposed approach. Left) Age regression; center) Sex classification; right) Five-year mortality prediction.

4.2 Multiple Pretraining Cohorts Comparison

An important aspect of the work is to explore how training can benefit from the huge corpus of available ECG data in terms of data size and diversity. Prior work hints that the results for a supervised task also depend on the underlying distribution of the dataset. For example, age and sex prediction shows higher performance for healthy subjects compared to unhealthy (Strodthoff et al.,

2021). The cardiac signal has been known to have ethnicity signatures (Mansi & Nash, 2004), thus the ethnic distribution of the training cohort can also affect the performance.

Experimental setup The following experiment is designed to study the effect of data demographics on the performance of pretraining approaches. Pretraining is accomplished on the four cohorts $BIDMC$ (USA), $CODE$ (Brazil), $SHZS$ (China), and $VUMC$ (USA), having different ethnic distributions, patient count, and ECG numbers. The $BIDMC$, $CODE$, $SHZS$, $PTB\text{-}XL$, and UKB datasets contain the age and sex of the patients for each ECG. The pretrained models are used to extract features from these cohorts. The features are then leveraged for age and sex prediction with the train/val/test splits for all datasets consisting of $10k$/$2k$/$2k$ labeled instances, to remove any dataset size bias from the supervised training. Figure 3 compares the performance of the age and sex prediction across multiple pretrained feature extraction models and labeled datasets, using a similar configuration for the supervised setup. The right-hand panel presents the AUC for the sex prediction while the left-hand panel shows the MAE for age prediction, as the mean of six runs.

How does the pretraining cohort affect performance? The following observations and insights can be obtained from the test that we believe to be essential for future work. The model pretrained on $BCSV$ (i.e. the combination of $BIDMC$, $CODE$, $SHZS$, and $VUMC$ datasets) outperforms for all labels and has the best generalisation capabilities for a foundation model. Apart from the $BCSV$, the model pretrained on the same labeled dataset generally performs the best, thus highlighting the importance of external open-source cohorts for performance comparison. It is interesting to note, that while the $CODE$ and $SHZS$ have more patients and ECGs, the performance is generally decreased compared to the secondary care cohorts of $BIDMC$ and $VUMC$. A plausible explanation can be that the learned features are more expressive when pretrained from a more diseased population as there is more diversity in ECG patterns, which is the case for the $BIDMC$ and $VUMC$ datasets. Moreover, the performance does not only depend on the number of unique patients but also the diversity of the positive examples i.e., the $ECGs$ per patient. Looking back at Table 1, the $BIDMC$ has far fewer unique patients and ECGs than $VUMC$, but has a higher number of $ECGs$ per patient that may help it achieve comparable performance.

Does the labeled dataset impact performance? The performance metric for a labeled dataset also depends on the distribution of the underlying population health. The UKB supervised tasks show the highest AUC and lowest MAE for the same pretrained model, being a healthy volunteer cohort. Therefore, a metric for a particular label cannot be used for performance comparison across diverse datasets. Several previous works compare their results based on a specific training task (McKeen et al., 2024), which may be meaningful only when model evaluation is compared on the same dataset and ideally the same train/test splits. The model, pretrained on the $CODE$ and $SHZS$ datasets, shows lower performance compared to much smaller datasets. One reason can be that these datasets come from a healthier population and consequently reduced diversity. Additionally, the population of $SHZS$ being more ethnically distinct can also impact the performance on diverse datasets. The performance of the patient-based approach is highly dependent on the diversity, number, and health of the underlying population.

4.3 External Validation

Standard benchmarks for computer vision approaches based on open source datasets like Imagenet (Deng et al., 2009), CIFAR10 (Krizhevsky, 2009), and COCO (Lin et al., 2014) have greatly facilitated unbiased performance comparison. Section 4.2 exploring the effect of pretraining cohorts, shows that the performance can vary due to data demographics, thus a fair comparison can be established by exploitation of open-source cohorts that can be easily accessible for future works.

Experimental setup We use the open-source ECG dataset $PTB\text{-}XL$ (Wagner et al., 2020) for external validation. The dataset is well explored in the literature and provides a range of benchmarks for performance comparison including fully supervised and pretraining strategies. The comparison here is limited to published results from literature where the authors may have fully optimsed all aspects of their approach. The values for all metrics may not be available as indicated by '-' in the tables. Appendix B provides a more detailed comparison with other pretraining approaches in Table 7 and detailed classification metrics in Table 8. The model is pretrained on the $BCSV$

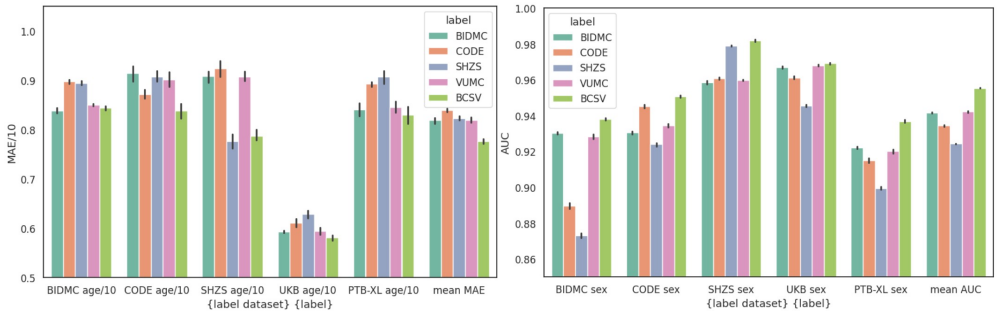

Figure 3: Performance comparison for the different pretraining and labeled datasets: right) Sex prediction, left) Age prediction. The different colors indicate different pretraining cohorts. The x-axis denotes the label and the labeled dataset while the y-axis denotes the metric under observation. The performance for $BCSV$ is the best across all labels and datasets.

cohort while the supervised setup consists of a linear classification head. Detailed hyperparameters optimisation is not performed; only learning rates are optimised from a range of $[0.00001, 0.1]$. A simple fine-tuning procedure is implemented without exploiting advanced techniques. The details of the experimental hyperparameters are provided in Appendix D. We explore two levels of diagnostic labels from $PTB\text{-}XL$ denoting cardiac abnormalities: Sub-classes refer to 23 morphological labels, and Super-classes are 5 overarching diagnostic classes. The dataset is divided into ten stratified folds, with fold 10 suggested for the test, fold 9 for the validation, and the remaining folds for the training (as per Wagner et al. (2020)). Further details about the dataset distribution and the labels can be obtained from Wagner et al. (2020).

How the TA-PCLR compares with supervised models? Table 3 compares the results with previously published benchmarks for the $PTB\text{-}XL$ employing fully supervised training. The training for super and sub classes is accomplished as multi-label classification where an instance can belong to more than one class, thus we report macro AUC. We compare with the best-performing benchmark from a study exploring deep neural networks (Strodthoff et al., 2021), which combines the predictions generated for different slices of the ECG signal, with a highly optimised ensemble of state-of-the-art deep neural networks. Bickmann et al. (2024) implements fully supervised learning for the prediction of $PTB\text{-}XL$ super-classes, utilizing advanced deep learning architecture (InceptionTime (Ismail Fawaz et al., 2020)). The $TA\text{-}PCLR$ with the linear probe has slightly lower performance than the fully supervised approaches as the feature extraction backbone is frozen. The performance improves significantly when the model is further fine-tuned allowing all weights to update. The age and sex predictions outperform the aggregated predictions from the ensemble network while diagnostic classes surpass the InceptionTime network pereformance (Bickmann et al., 2024). We set new benchmarks for age and sex prediction with 5.2% and 1.9% improvement, respectively.

How does the TA-PCLR compare with other pretraining methodologies? We now compare the $TA\text{-}PCLR$ approach to several representative pretraining approaches in Table 4. The linear classification head is trained individually for each label in super-classes: Myocardial Infarction (MI), ST/T change ($STTC$), Conduction Disturbance (CD), and Hypertrophy (HYP). We compare with a state-of-the-art masked autoencoder-based foundational model by Song et al. (2024) and a transformer-based cross-dimensional approach Liu et al. (2023) (reporting a superior performance as compared to the $3KG$ (Gopal et al., 2021), and $PCLR$), although noting that they removed instances with multiple labels to implement the task as multi-class. We present their result while noting that the data will be a subset of the dataset with a simpler task. The performance of $TA\text{-}PCLR$ with linear probe significantly outperforms these more complex approaches with an improvement of 5 to 8 percent over Song et al. (2024) for individual classes and 2.97% above Liu et al. (2023). Additional comparisons in Appendix B Table 7 further substantiate the superiority of the proposed approach while Table 8 provide more detailed metrics for the the results presented in this section.

Table 3: Performance Comparison with Supervised training for $PTBXL$: Age, sex, diagnosis super/sub classes

No.	Method	MAE		AUC	
		Age	Sex	Super-classes	Sub-classes
1	Strodthoff et al. (2021) (ensemble)	7.12	0.928	**0.934**	**0.933**
2	Bickmann et al. (2024) (Inception)	-	-	0.902	-
3	TA-$PCLR$ linear	7.57	0.938	0.889	0.919
4	TA-$PCLR$ fine-tuning	**6.75**	**0.946**	0.919	0.928

Table 4: Comparing with pretraining approaches for the PTB-XL Super-classes: Myocardial Infarction (MI), ST/T change ($STTC$), Conduction Disturbance (CD), and Hypertrophy (HYP).

No.	Method	MI	$STTC$	CD	HYP	$NORM$	Mean
1	Song et al. (2024)	0.8318	0.8165	0.8411	0.8135	-	-
2	Liu et al. (2023)	-	-	-	-	-	0.8648
3	TA-$PCLR$ linear	**0.8948**	**0.8848**	**0.8885**	**0.8669**	**0.9173**	**0.8905**

5 Conclusions and Future Work

In this work we present TA-$PCLR$, applying a novel combination of temporal augmentations and patient-based ECG contrastive learning, enhancing performance in downstream supervised tasks. We demonstrate that TA-$PCLR$ is superior to fully supervised training methods on small to medium-sized datasets, proving to be especially valuable in scenarios where labeled data is limited.

Pretraining with combined datasets from three continents, forming one of the largest and most diverse, multi-site ECG cohorts, further improves the performance for downstream supervised tasks and sets new benchmarks when evaluated in external validation using the $PTB-XL$ dataset. Looking ahead, we plan to explore the impact of additional data augmentations on the learned representations and focus on enhancing the interpretability of the features learned by TA-$PCLR$ (some work on interpretability is presented in Appendix C). The current work demonstrates the capabilities of our foundation model for a range of generic tasks without increasing the model's complexity. An interesting research direction for future work is exploring model scalability while noting that the rule of ten times more data than model parameters from prior research (Alwosheel et al., 2018) and our remarkable results attest to the fact that the model size maybe adequate. The strong performance, generalizability, and efficiency of TA-$PCLR$ in terms of training time and network size, positions it as a powerful foundation model.

5.1 Ethics statement

Our research complies with all relevant ethical regulations and details of the ethics approval are provided in Table 5.

5.2 Reproducibility Statement

The pretraining data consist of private cohorts thus the trained network cannot be released but the details provided in Section 3 are sufficient to reproduce the methodology.

Table 5: Datasets

No.	Cohorts	Ethics Approval
1	$BIDMC$	Beth Israel Deaconess Medical Center Committee on Clinical Investigations (IRB protocol # 2023P000042).
2	$CODE$	Research Ethics Committee of the Universidade Federal de Minas Gerais (protocol 49368496317.7.0000.5149)
3	$SHZS$	Institutional Research Board of Zhongshan Hospital (No. B2023-253R) with a waiver of patient consent
4	$VUMC$	The Vanderbilt component of this study was reviewed and approved by the Institutional Review Board (#212147)
5	UKB	North West Multi-Centre Research Ethics Committee application ID 48666
6	$PTB\text{-}XL$	The Institutional Ethics Committee approved the publication of the anonymous data in an open-access database (PTB-2020-1).

AUTHOR CONTRIBUTIONS

ACKNOWLEDGMENTS

REFERENCES

Martín Abadi, Ashish Agarwal, et al. TensorFlow: Large-scale machine learning on heterogeneous systems, 2015. URL https://www.tensorflow.org/. Software available from tensorflow.org.

Ahmad Alwosheel, Sander van Cranenburgh, and Caspar G. Chorus. Is your dataset big enough? sample size requirements when using artificial neural networks for discrete choice analysis. *Journal of Choice Modelling*, 28:167–182, 2018. ISSN 1755-5345. doi: https://doi.org/10.1016/j.jocm.2018.07.002. URL https://www.sciencedirect.com/science/article/pii/S1755534518300058.

Mandar A. Aras, Sean Abreau, et al. Electrocardiogram detection of pulmonary hypertension using deep learning. *Journal of Cardiac Failure*, 29(7):1017–1028, 2023. ISSN 1071-9164. doi: https://doi.org/10.1016/j.cardfail.2022.12.016. URL https://www.sciencedirect.com/science/article/pii/S107191642300012X.

Zachi I. Attia, Paul A. Friedman, Peter A. Noseworthy, Francisco Lopez-Jimenez, Dorothy J. Ladewig, Gaurav Satam, Patricia A. Pellikka, Thomas M. Munger, Samuel J. Asirvatham, Christopher G. Scott, Rickey E. Carter, and Suraj Kapa. Age and sex estimation using artificial intelligence from standard 12-lead ecgs. *Circulation: Arrhythmia and Electrophysiology*, 12 (9):e007284, 2019. doi: 10.1161/CIRCEP.119.007284. URL https://www.ahajournals.org/doi/abs/10.1161/CIRCEP.119.007284.

Lucas Bickmann, Lucas Plagwitz, and Julian Varghese. Benchmarking approaches: Time series versus Feature-Based machine learning in ECG analysis on the PTB-XL dataset. *Stud Health Technol Inform*, 316:589–593, August 2024.

John Scott Bridle. Probabilistic interpretation of feedforward classification network outputs, with relationships to statistical pattern recognition. In *NATO Neurocomputing*, 1989. URL https://api.semanticscholar.org/CorpusID:59636530.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *Proceedings of the 37th International Conference on Machine Learning*, ICML'20. JMLR.org, 2020.

Xinlei Chen and Kaiming He. Exploring simple siamese representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 15750–15758, June 2021.

Joseph Y. Cheng, Hanlin Goh, Kaan Dogrusoz, Oncel Tuzel, and Erdrin Azemi. Subject-aware contrastive learning for biosignals, 2021. URL https://arxiv.org/pdf/2007.04871.pdf.

Yuanzheng Ci, Chen Lin, Lei Bai, and Wanli Ouyang. Fast-moco: Boost momentum-based contrastive learning with combinatorial patches. In Shai Avidan, Gabriel Brostow, Moustapha Cissé, Giovanni Maria Farinella, and Tal Hassner (eds.), *Computer Vision – ECCV 2022*, pp. 290–306, Cham, 2022. Springer Nature Switzerland. ISBN 978-3-031-19809-0.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 248–255, 2009. doi: 10.1109/CVPR.2009.5206848.

Nathaniel Diamant, Erik Reinertsen, Steven Song, Aaron D. Aguirre, Collin M. Stultz, and Puneet Batra. Patient contrastive learning: A performant, expressive, and practical approach to electrocardiogram modeling. *PLOS Computational Biology*, 18(2):1–16, 02 2022. doi: 10.1371/journal.pcbi.1009862. URL https://doi.org/10.1371/journal.pcbi.1009862.

Changkyoung Eem, Hyunki Hong, and Yoohun Noh. Deep-learning model to predict coronary artery calcium scores in humans from electrocardiogram data. *Applied Sciences*, 10:8746, 12 2020. doi: 10.3390/app10238746.

Emadeldeen Eldele, Mohamed Ragab, Zhenghua Chen, Min Wu, Chee-Keong Kwoh, Xiaoli Li, and Cuntai Guan. Self-supervised contrastive representation learning for semi-supervised time-series classification. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(12): 15604–15618, 2023. doi: 10.1109/TPAMI.2023.3308189.

Andre Esteva, Alexandre Robicquet, et al. A guide to deep learning in healthcare. *Nature Medicine*, 25(1):24–29, Jan 2019. ISSN 1546-170X. doi: 10.1038/s41591-018-0316-z. URL https://doi.org/10.1038/s41591-018-0316-z.

Daniel Gedon, Antônio H. Ribeiro, Niklas Wahlström, and Thomas B. Schön. First steps towards self-supervised pretraining of the 12-lead ecg. In *2021 Computing in Cardiology (CinC)*, volume 48, pp. 1–4, 2021. doi: 10.23919/CinC53138.2021.9662748.

Bryan Gopal, Ryan Han, Gautham Raghupathi, Andrew Ng, Geoff Tison, and Pranav Rajpurkar. 3kg: Contrastive learning of 12-lead electrocardiograms using physiologically-inspired augmentations. In Subhrajit Roy, Stephen Pfohl, Emma Rocheteau, Girmaw Abebe Tadesse, Luis Oala, Fabian Falck, Yuyin Zhou, Liyue Shen, Ghada Zamzmi, Purity Mugambi, Ayah Zirikly, Matthew B. A. McDermott, and Emily Alsentzer (eds.), *Proceedings of Machine Learning for Health*, volume 158 of *Proceedings of Machine Learning Research*, pp. 156–167. PMLR, 04 Dec 2021. URL https://proceedings.mlr.press/v158/gopal21a.html.

Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre H. Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Daniel Guo, Mohammad Gheshlaghi Azar, Bilal Piot, Koray Kavukcuoglu, Rémi Munos, and Michal Valko. Bootstrap your own latent a new approach to self-supervised learning. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS '20, Red Hook, NY, USA, 2020. Curran Associates Inc. ISBN 9781713829546.

R. Hadsell, S. Chopra, and Y. LeCun. Dimensionality reduction by learning an invariant mapping. In *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*, volume 2, pp. 1735–1742, 2006. doi: 10.1109/CVPR.2006.100.

Hassan Ismail Fawaz, Benjamin Lucas, Germain Forestier, Charlotte Pelletier, Daniel F. Schmidt, Jonathan Weber, Geoffrey I. Webb, Lhassane Idoumghar, Pierre-Alain Muller, and François Petitjean. Inceptiontime: Finding alexnet for time series classification. *Data Mining and Knowledge Discovery*, 34(6):1936–1962, Nov 2020. ISSN 1573-756X. doi: 10.1007/s10618-020-00710-y. URL https://doi.org/10.1007/s10618-020-00710-y.

Amir Jamaludin, Timor Kadir, and Andrew Zisserman. Self-supervised learning for spinal mris. In M. Jorge Cardoso, Tal Arbel, Gustavo Carneiro, Tanveer Syeda-Mahmood, João Manuel R.S. Tavares, Mehdi Moradi, Andrew Bradley, Hayit Greenspan, João Paulo Papa, Anant Madabhushi, Jacinto C. Nascimento, Jaime S. Cardoso, Vasileios Belagiannis, and Zhi Lu (eds.), *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support*, pp. 294–302, Cham, 2017. Springer International Publishing. ISBN 978-3-319-67558-9.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *CoRR*, abs/1412.6980, 2014. URL https://api.semanticscholar.org/CorpusID:6628106.

Dani Kiyasseh, Tingting Zhu, and David A Clifton. Clocs: Contrastive learning of cardiac signals across space, time, and patients. In Marina Meila and Tong Zhang (eds.), *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 5606–5615. PMLR, 18–24 Jul 2021. URL https://proceedings.mlr.press/v139/kiyasseh21a.html.

Alex Krizhevsky. Learning multiple layers of features from tiny images. 2009. URL https://api.semanticscholar.org/CorpusID:18268744.

Byeong Tak Lee, Yong-Yeon Jo, Seon-Yu Lim, Youngjae Song, and Joon-myoung Kwon. Efficient data augmentation policy for electrocardiograms. In *Proceedings of the 31st ACM International Conference on Information & Knowledge Management*, CIKM '22, pp. 4153–4157, New York, NY, USA, 2022. Association for Computing Machinery. ISBN 9781450392365. doi: 10.1145/3511808.3557591. URL https://doi.org/10.1145/3511808.3557591.

Jun Li, Aaron Aguirre, Junior Moura, Che Liu, Lanhai Zhong, Chenxi Sun, Gari Clifford, Brandon Westover, and Shenda Hong. An electrocardiogram foundation model built on over 10 million recordings with external evaluation across multiple domains, 2024. URL https://arxiv.org/abs/2410.04133.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft coco: Common objects in context. In David Fleet, Tomas Pajdla, Bernt Schiele, and Tinne Tuytelaars (eds.), *Computer Vision – ECCV 2014*, pp. 740–755, Cham, 2014. Springer International Publishing.

Che Liu, Zhongwei Wan, Cheng Ouyang, Anand Shah, Wenjia Bai, and Rossella Arcucci. Zero-shot ecg classification with multimodal learning and test-time clinical knowledge enhancement. In *Forty-first International Conference on Machine Learning*, 2024. URL https://openreview.net/pdf/51fc8d05af466d8ea0c8798d5603726c7b643ae.pdf.

Wenhan Liu, Huaicheng Zhang, Sheng Chang, Hao Wang, Jin He, and Qijun Huang. A joint cross-dimensional contrastive learning framework for 12-lead ecgs and its heterogeneous deployment on soc. *Computers in Biology and Medicine*, 152:106390, 2023. ISSN 0010-4825. doi: https://doi.org/10.1016/j.compbiomed.2022.106390. URL https://www.sciencedirect.com/science/article/pii/S0010482522010988.

Ilya Loshchilov and Frank Hutter. Sgdr: Stochastic gradient descent with warm restarts. 08 2016.

Ishak A Mansi and Ira S Nash. Ethnic differences in electrocardiographic amplitude measurements. *Ann Saudi Med*, 24(6):459–464, November 2004.

Kaden McKeen, Laura Oliva, Sameer Masood, Augustin Toma, Barry Rubin, and Bo Wang. Ecg-fm: An open electrocardiogram foundation model, 2024. URL https://arxiv.org/abs/2408.05178.

Mostafa Neo Mohsenvand, Mohammad Rasool Izadi, and Pattie Maes. Contrastive representation learning for electroencephalogram classification. In Emily Alsentzer, Matthew B. A. McDermott, Fabian Falck, Suproteem K. Sarkar, Subhrajit Roy, and Stephanie L. Hyland (eds.), *Proceedings of the Machine Learning for Health NeurIPS Workshop*, volume 136 of *Proceedings of Machine Learning Research*, pp. 238–253. PMLR, 11 Dec 2020. URL https://proceedings.mlr.press/v136/mohsenvand20a.html.

Yeongyeon Na, Minje Park, and Yunwon Tae. Guiding masked representation learning to capture spatio-temporal relationship of electrocardiogram. https://synthical.com/article/2479071f-19d4-4d1b-95d7-a85a7909abfb, 1 2024.

Libor Pastika, Arunashis Sau, Konstantinos Patlatzoglou, Ewa Sieliwonczyk, Antônio H. Ribeiro, Kathryn A. McGurk, Sadia Khan, Danilo Mandic, William R. Scott, James S. Ware, Nicholas S. Peters, Antonio Luiz P. Ribeiro, Daniel B. Kramer, Jonathan W. Waks, and Fu Siong Ng. Artificial intelligence-enhanced electrocardiography derived body mass index as a predictor of future cardiometabolic disease. *npj Digital Medicine*, 7(1):167, Jun 2024. ISSN 2398-6352. doi: 10.1038/s41746-024-01170-0. URL https://doi.org/10.1038/s41746-024-01170-0.

Aniruddh Raghu, Divya Shanmugam, Eugene Pomerantsev, John Guttag, and Collin M. Stultz. Data augmentation for electrocardiograms, 2022. URL https://arxiv.org/abs/2204.04360.

Antônio H. Ribeiro, Manoel Horta Ribeiro, Gabriela M. M. Paixão, Derick M. Oliveira, Paulo R. Gomes, Jéssica A. Canazart, Milton P. S. Ferreira, Carl R. Andersson, Peter W. Macfarlane, Wagner Meira Jr., Thomas B. Schön, and Antonio Luiz P. Ribeiro. Automatic diagnosis of the 12-lead ecg using a deep neural network. *Nature Communications*, 11(1):1760, Apr 2020. ISSN 2041-1723. doi: 10.1038/s41467-020-15432-4. URL https://doi.org/10.1038/s41467-020-15432-4.

Antonio Luiz P. Ribeiro, Gabriela M.M. Paixão, Paulo R. Gomes, Manoel Horta Ribeiro, Antônio H. Ribeiro, Jéssica A. Canazart, Derick M. Oliveira, Milton P. Ferreira, Emilly M. Lima, Jermana Lopes de Moraes, Nathalia Castro, Leonardo B. Ribeiro, and Peter W. Macfarlane. Tele-electrocardiography and bigdata: The code (clinical outcomes in digital electrocardiography) study. *Journal of Electrocardiology*, 57:S75–S78, 2019. ISSN 0022-0736. doi: https://doi.org/10.1016/j.jelectrocard.2019.09.008. URL https://www.sciencedirect.com/science/article/pii/S0022073619304984.

Arunashis Sau, Antonio H. Ribeiro, Kathryn A. McGurk, Libor Pastika, Nikesh Bajaj, Maddalena Ardissino, Jun Yu Chen, Huiyi Wu, Xili Shi, Katerina Hnatkova, Sean Zheng, Annie Britton, Martin Shipley, Irena Andršová, Tomáš Novotný, Ester Sabino, Luana Giatti, Sandhi M Barreto, Jonathan W. Waks, Daniel B. Kramer, Danilo Mandic, Nicholas S. Peters, Declan P. O'Regan, Marek Malik, James S. Ware, Antonio Luiz P. Ribeiro, and Fu Siong Ng. Neural network-derived electrocardiographic features have prognostic significance and important phenotypic and genotypic associations. *Circulation: Cardiovascular Quality and Outcomes (in press)*, 2023. doi: 10.1101/2023.06.15.23291428. URL https://www.medrxiv.org/content/early/2023/06/16/2023.06.15.23291428.

Arunashis Sau, Libor Pastika, Ewa Sieliwonczyk, Konstantinos Patlatzoglou, Antonio H. Ribeiro, Kathryn A. McGurk, Boroumand Zeidaabadi, Henry Zhang, Krzysztof Macierzanka, Danilo Mandic, Ester Sabino, Luana Giatti, Sandhi M Barreto, Lidyane do Valle Camelo, Ioanna Tzoulaki, Declan P. O'Regan, Nicholas S. Peters, James S. Ware, Antonio Luiz P. Ribeiro, Daniel B. Kramer, Jonathan W. Waks, and Fu Siong Ng. Artificial intelligence–enabled electrocardiogram for mortality and cardiovascular risk estimation: An actionable, explainable and biologically plausible platform. *Lancet Digital Health (in press)*, 2024. doi: 10.1101/2024.01.13.24301267. URL https://www.medrxiv.org/content/early/2024/01/15/2024.01.13.24301267.

Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pp. 618–626, 2017. doi: 10.1109/ICCV.2017.74.

Sahar Soltanieh, Ali Etemad, and Javad Hashemi. Analysis of augmentations for contrastive ecg representation learning. In *2022 International Joint Conference on Neural Networks (IJCNN)*, pp. 1–10, 2022. doi: 10.1109/IJCNN55064.2022.9892600.

Junho Song, Jong-Hwan Jang, Byeong Tak Lee, DongGyun Hong, Joon myoung Kwon, and Yong-Yeon Jo. Foundation models for electrocardiograms, 2024. URL https://arxiv.org/abs/2407.07110.

Nils Strodthoff, Patrick Wagner, Tobias Schaeffter, and Wojciech Samek. Deep learning for ecg analysis: Benchmarks and insights from ptb-xl. *IEEE Journal of Biomedical and Health Informatics*, 25(5):1519–1528, 2021. doi: 10.1109/JBHI.2020.3022989.

Cathie Sudlow, John Gallacher, Naomi Allen, Valerie Beral, Paul Burton, John Danesh, Paul Downey, Paul Elliott, Jane Green, Martin Landray, Bette Liu, Paul Matthews, Giok Ong, Jill Pell, Alan Silman, Alan Young, Tim Sprosen, Tim Peakman, and Rory Collins. Uk biobank: An open access resource for identifying the causes of a wide range of complex diseases of middle and old age. *PLOS Medicine*, 12(3):1–10, 03 2015. doi: 10.1371/journal.pmed.1001779. URL https://doi.org/10.1371/journal.pmed.1001779.

Akhil Vaid, Joy Jiang, Ashwin Sawant, Stamatios Lerakis, Edgar Argulian, Yuri Ahuja, Joshua Lampert, Alexander Charney, Hayit Greenspan, Jagat Narula, Benjamin Glicksberg, and Girish N. Nadkarni. A foundational vision transformer improves diagnostic performance for electrocardiograms. *npj Digital Medicine*, 6(1):108, Jun 2023. ISSN 2398-6352. doi: 10.1038/s41746-023-00840-9. URL https://doi.org/10.1038/s41746-023-00840-9.

Aäron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *ArXiv*, abs/1807.03748, 2018. URL https://api.semanticscholar.org/CorpusID:49670925.

Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of Machine Learning Research*, 9(86):2579–2605, 2008. URL http://jmlr.org/papers/v9/vandermaaten08a.html.

Patrick Wagner, Nils Strodthoff, Ralf-Dieter Bousseljot, Dieter Kreiseler, Fatima I. Lunze, Wojciech Samek, and Tobias Schaeffter. Ptb-xl, a large publicly available electrocardiography dataset. *Scientific Data*, 7(1):154, May 2020. ISSN 2052-4463. doi: 10.1038/s41597-020-0495-6. URL https://doi.org/10.1038/s41597-020-0495-6.

A D Waller. A demonstration on man of electromotive changes accompanying the heart's beat. *J Physiol*, 8(5):229–234, October 1887.

Ning Wang, Panpan Feng, Zhaoyang Ge, Yanjie Zhou, Bing Zhou, and Zongmin Wang. Adversarial spatiotemporal contrastive learning for electrocardiogram signals. *IEEE Transactions on Neural Networks and Learning Systems*, 35(10):13845–13859, 2024. doi: 10.1109/TNNLS.2023.3272153.

Xun Wang, Xintong Han, Weilin Huang, Dengke Dong, and Matthew R. Scott. Multi-similarity loss with general pair weighting for deep metric learning. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5017–5025, 2019. doi: 10.1109/CVPR.2019.00516.

Crystal T. Wei, Ming-En Hsieh, Chien-Liang Liu, and Vincent S. Tseng. Contrastive heartbeats: Contrastive learning for self-supervised ecg representation and phenotyping. In *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1126–1130, 2022. doi: 10.1109/ICASSP43922.2022.9746887.

Qingsong Wen, Liang Sun, Fan Yang, Xiaomin Song, Jingkun Gao, Xue Wang, and Huan Xu. Time series data augmentation for deep learning: A survey. In Zhi-Hua Zhou (ed.), *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI-21*, pp. 4653–4660. International Joint Conferences on Artificial Intelligence Organization, 8 2021. doi: 10.24963/ijcai.2021/631. URL https://doi.org/10.24963/ijcai.2021/631. Survey Track.

Jure Zbontar, Li Jing, Ishan Misra, Yann LeCun, and Stéphane Deny. Barlow twins: Self-supervised learning via redundancy reduction. In *International conference on machine learning*, pp. 12310–12320. PMLR, 2021.

Shaoting Zhang and Dimitris Metaxas. On the challenges and perspectives of foundation models for medical image analysis. *Medical Image Analysis*, 91:102996, 2024. ISSN 1361-8415. doi: https://doi.org/10.1016/j.media.2023.102996. URL https://www.sciencedirect.com/science/article/pii/S1361841523002566.

Wenrui Zhang, Ling Yang, Shijia Geng, and Shenda Hong. Self-supervised time series representation learning via cross reconstruction transformer. *IEEE Transactions on Neural Networks and Learning Systems*, 35(11):16129–16138, 2024. doi: 10.1109/TNNLS.2023.3292066.

A Dataset analysis

Table 6 presents the demographics of the population included in the study. Further details can be obtained from the corresponding references. Although the ethnicity information for most cohorts is not available, depending on the geographical location the predominant ethnicity can be inferred.

Table 6: Baseline characteristics of the population used in the current research.

	$BIDMC$	$CODE$	$SHZS$	$VUMC$	UKB	$PTB\text{-}XL$
Patients*	$127,041$	$424,577$	$420,956$	$252,306$	$66,402$	$18,869$
Ecgs	$1,106,886$	$1,123,903$	$1,560,551$	$1,412,012$	$70,655$	$21,799$
Age mean	57.99	56.00	52.08	52.08	65.35	62.36
Age IQR	23.02	23.00	27.00	27.00	12.00	23.00
Male	$63,006$	$165,285$	$233,808$	$233,808$	$32,191$	$9,640$
Female	$640,35$	$259,292$	$187,148$	$187,148$	$34,211$	$9,229$
Hispanic	$7,077$	-	-	-	-	-
White	$84,265$	-	-	-	-	-
Black	$17,778$	-	-	-	-	-
Asian	$5,315$	-	-	-	-	-
Other	$12,606$	-	-	-	-	-
Mortality!	21%	-	-	-	-	-

* Patients with more than one ECG.
! Five-year mortality.

B Additional results for PTB-XL

Table 7 compares the macro AUC values reported by Liu et al. (2024) for state-of-the-art pre-training approaches, with our $TA\text{-}PCLR$ approach. We repeat the tests for ten independent runs with 1%, 10%, and 100% random splits of the training data while the validation and test splits remain the same. The learning rates are optimized from the range of [0.0001, 0.01]. The parameters for 100% split are provided in Table 9, while learning rates used for 10% split are 0.001 for the sub classes and 0.01 for the super classes. Similarly, the learning rates for 1% split are 0.05 for sub classes and 0.01 for the super classes. It should be noted here that $MERL$ has some auxiliary supervision during training, in the form of diagnostic ecg-report alignment that enables the model to perform zero-shot prediction. The $TA\text{-}PCLR$ outperforms for all except for the 1% split for the super classes, where it has the second-best performance.

Table 8 provides detailed metrics for the classification tasks performed in Tables 3 and 4. The results are obtained from ten random runs and presented as macro AUC mean with 95% confidence interval. The metrics like precision, recall, $F1$, and accuracy greatly depend on the threshold for the binary output. A threshold of 0.5 is used for the reported results but the metrics will be significantly improved by optimizing the threshold.

Table 7: Linear probe result comparison with SOTA approaches for the PTB-XL Super and sub classes classification.

No.	Method	Super-classes 1%	Super-classes 10%	Super-classes 100%	Sub-classes 1%	Sub-classes 10%	Sub-classes 100%
1	SimCLR Chen et al. (2020)	0.634	0.698	0.735	0.608	0.683	0.734
2	BYOL Grill et al. (2020)	0.717	0.738	0.764	0.572	0.674	0.716
3	BarlowTwins Zbontar et al. (2021)	0.729	0.760	0.784	0.626	0.708	0.743
4	MoCo-v3 Ci et al. (2022)	0.732	0.766	0.783	0.559	0.692	0.767
5	SimSiam Chen & He (2021)	0.731	0.727	0.756	0.625	0.693	0.764
6	TS-TCC Eldele et al. (2023)	0.707	0.759	0.789	0.535	0.670	0.779
7	CLOCS Kiyasseh et al. (2021)	0.689	0.734	0.763	0.579	0.725	0.762
8	ASTCL Wang et al. (2024)	0.725	0.773	0.810	0.619	0.688	0.765
9	CRT Zhang et al. (2024)	0.697	0.782	0.772	0.620	0.708	0.787
10	ST-MEM Na et al. (2024)	0.611	0.669	0.713	0.541	0.579	0.636
11	MERL Liu et al. (2024)	**0.824**	0.862	0.887	0.649	0.806	0.847
12	TA-$PCLR$	0.788	**0.870**	**0.889**	**0.685**	**0.849**	**0.918**

Table 8: Additional results for the PTB-XL Super and sub classes classification for TA-$PCLR$.

No.	label	precision	recall	F1	accuracy	AUROC
			Linear Probe			
1	Super	0.752±0.0027	0.578±0.0035	0.643±0.0030	0.860±0.0006	0.889±0.0003
2	Sub	0.531±0.0193	0.316±0.0067	0.369±0.0063	0.961±0.0002	0.912±0.0015
3	MI	0.783±0.0039	0.811±0.0043	0.794±0.0020	0.837±0.0035	0.895±0.0013
4	STTC	0.761±0.0016	0.808±0.0024	0.777±0.0014	0.823±0.0020	0.885±0.0004
5	CD	0.754±0.0070	0.801±0.0046	0.770±0.0057	0.823±0.0073	0.888±0.0018
6	HYP	0.681±0.0031	0.800±0.0044	0.711±0.0037	0.835±0.0054	0.867±0.0016
7	NORM	0.832±0.0014	0.835±0.0014	0.825±0.0022	0.825±0.0022	0.917±0.0002
			Fine tuning			
8	Super	0.777±0.0062	0.681±0.0080	0.720±0.0041	0.884±0.0010	0.919±0.0007
9	Sub	0.534±0.0218	0.376±0.0218	0.420±0.0082	0.964±0.0005	0.928±0.0029

C INTERPRETABILITY

The t-SNE (t-distributed Stochastic Neighbor Embedding) (van der Maaten & Hinton, 2008) is a non-linear dimensionality reduction algorithm that can help to visualize high-dimensional data in two or three dimensions. The pretraining is not supervised so it encodes the generic ECG features. Figure 4 right panel shows the correlations between the different features (features are standardized

and the features with zero standard deviations are removed from the further study). The correlations show that features are not very correlated and thus more expressive. The left panel shows the two principle components obtained by the t-SNE, with five super classes in different colors. The classes are multi-label and not mutually exclusive thus mostly expressed as gradients instead of clustering. The classes are also composite so the same classes can be observed to be located in the different but nearby regions of the embedding space.

Figure 5 shows principle t-SNE components for other labels like sex, NORM, and age. The components are obtained for features with greater than 0.3 correlation with the corresponding label. The plots for sex and NORM show that the embedding space can separate the genders and normal versus abnormal in opposite directions. The right panel shows the different age groups that show a gradient in the t-SNE representation.

Figure 6 further explores the interpretability of the model using $GradCam$ (Selvaraju et al., 2017). The top two ECGs are examples with positive class $STTC$, while the bottom two ECGs are negatives. The gradients are superimposed on the input signal and are represented in red color where the darker color indicates higher importance. The $STTC$ class is related to abnormalities in the ST segment of the ECG. The plots show that the model can recognize the normal samples using the region around the ST segments probably by learning the specific shape in a normal person. The positive samples with abnormalities do not give any importance to this region.

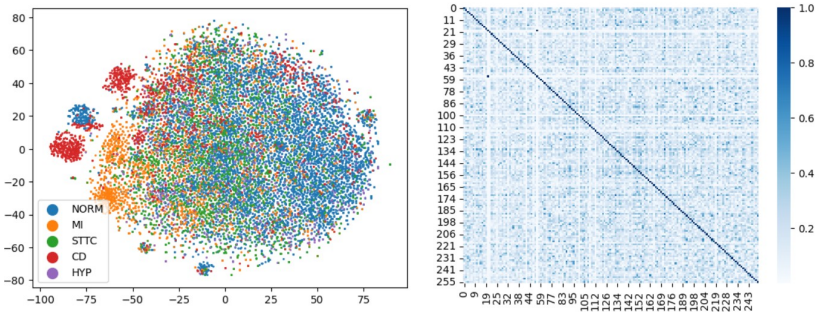

Figure 4: The $TA\text{-}PCLR$ features using self supervised pretraining: Left) The t-SNE plots for the PTB-XL super classes. Right) Correlations between features.

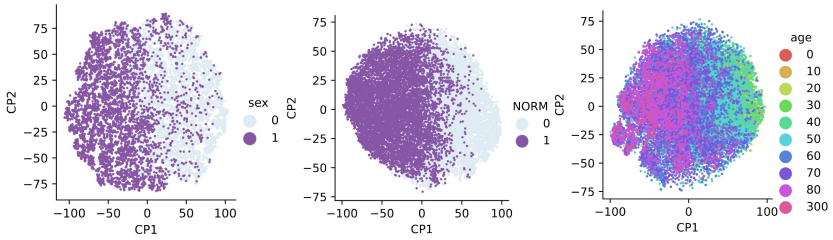

Figure 5: The GradCam is used to highlight the regions contributing to the prediction of a label. The ECGs are superimposed on the gradient maps with darker colors indicating more important regions. It can be observed that the model gives high importance to the ST segment for normal samples while failing to do so in the abnormal samples, thus probably using the absence of the pattern as the indication of the diagnosis.

D SUPERVISED TRAINING HYPER-PARAMETER

Details of the supervised training hyper-parameters are presented in Table 9

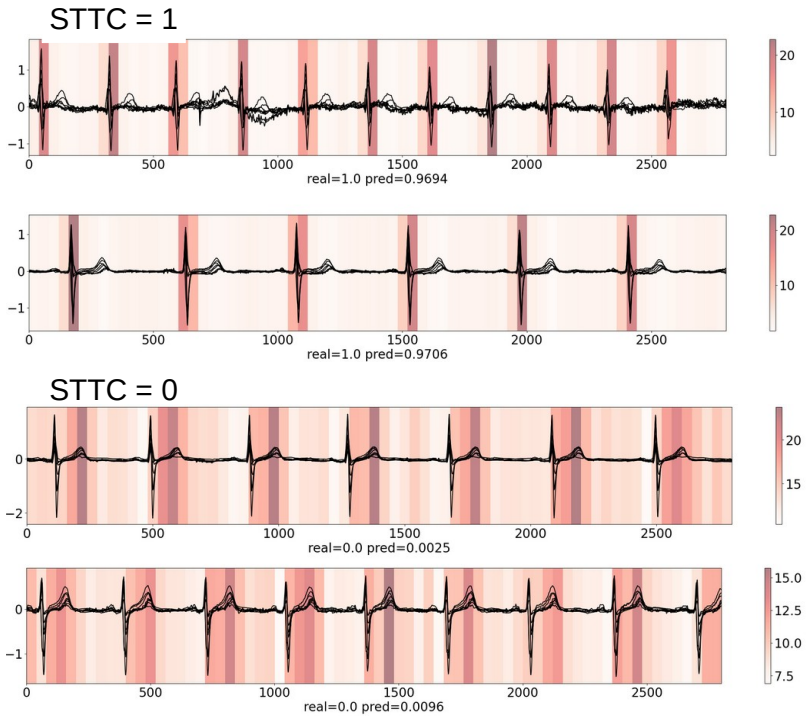

Figure 6: The TA-$PCLR$ features using self supervised pretraining: Left) The t-SNE plots for the PTB-XL super classes. Right) Correlations between features.

Table 9: Training hyperparameters

No.	task	Parameter	Value
1	All	optimizer	Adam with lr schedule: reduce-on-plateau and early-stopping
2	ResNet (all)	learning rate	0.0005
		Section 4.1 and 4.2	
2	Age regression	learning rate	0.0001
3		prediction head	MLP hidden = [256, 128]
4	Sex classification	learning rate	0.0001
5		prediction head	MLP hidden = [256, 256]
6	Mortality (5y) classification	learning rate	0.0001
7		prediction head	MLP hidden = [256, 256]
		Section 4.2 Table 3	
8	Age regression	learning rate	0.005
9		prediction head	single neuron
10	Sex classification	learning rate	0.005
11		prediction head	single neuron
12	Super classes classification	learning rate	0.0001
13		prediction head	single layer with neurons equal to the number of outputs
14	Sub classes classification	learning rate	0.0001
15		prediction head	single layer with neurons equal to the number of outputs
		Section 4.3 Table 4	
16	Super classes regression	learning rate	0.005
17		prediction head	single neuron