ECG Representation Learning with Multi-Modal EHR Data

Anonymous authors
Paper under double-blind review

Abstract

Electronic health records (EHRs) provide a rich source of medical information across different modalities such as electrocardiograms (ECG), structured EHRs (sEHR), and unstructured EHRs (text). Inspired by the fact that many cardiac and non-cardiac diseases influence the behavior of the ECG, we leverage structured EHRs and unstructured EHRs from multiple sources by pairing with ECGs and propose a set of three new multi-modal contrastive learning models that combine ECG, sEHR, and text modalities and a supervised large scale multi-task learning model trained to perform both classification and regression tasks on a large number of cardiovascular diseases and lab test measurements to produce robust representations of ECGs that can subsequently be used for a variety of downstream tasks. The performance of these models is compared against different baseline models such as supervised learning models trained from scratch with random weights initialization, and self-supervised learning models trained only on ECGs. We pre-train the models on a large proprietary dataset of about 9 *million* ECGs from around 2.4 *million* patients and evaluate the pre-trained models on various downstream tasks such as classification, zero-shot retrieval, and out-of-distribution detection involving the prediction of various heart conditions using ECG waveforms as input, and demonstrate that the models presented in this work show significant improvements compared to all baseline modes.

1 Introduction

Electronic health records (EHRs) are generated for every patient encounter or event and are becoming increasingly available in recent years. These are multi-modal in nature and capture rich phenotypic information of the patients over time in the form of structured EHRs and unstructured EHRs. Structured EHRs (denoted sEHR) contain information about diagnoses, procedures, medication prescriptions, lab tests, vitals, and more, while unstructured EHRs encompass clinical notes, radiology images, pathology images, echocardiogram videos, time series ECG signals, etc. Recently multi-modal contrastive learning methods applied to radiology and pathology images by pairing with the corresponding medical reports to learn medical image representations (Zhang et al., 2022; Huang et al., 2021; Boecking et al., 2022; Bannur et al., 2023; Lu et al., 2023) have shown promising results on downstream tasks such as classification, image-text retrieval, etc. These methods generally have two stages: i) In stage I, the model is pre-trained on large unlabelled data to learn generic representations by maximizing the alignment between embeddings of different modalities in latent space; ii) In stage II, the model is fine-tuned on a task-specific labeled dataset by transferring the knowledge from the pre-trained model. However, ECG representation learning by pairing with EHRs via multi-modal contrastive learning is less explored. Uni-modal contrastive learning similar to Chen et al. (2020a) has been applied to the ECG domain to learn ECG representations (Kiyasseh et al., 2021; Diamant et al., 2022; Gopal et al., 2021; Mehari & Strodthoff, 2022; Oh et al., 2022), but they lack the ability to compare different modalities in latent space using similarity metrics like cosine similarity for use in zero-shot transfer learning. Also, contrastive learning using multi-modal data produces high-quality representations as they exploit information from multiple sources and extract semantics by aligning with various modalities.

ECG is a simple, non-invasive test that records the electrical activity of the heart and is helpful in diagnosing heart conditions and patient monitoring. In recent years, deep learning techniques have been employed on ECG data to predict various heart conditions, even in cases where diagnostic criteria using ECGs have not

been firmly established in clinical practice (Tison et al., 2019; Hannun et al., 2019; Galloway et al., 2019; Attia et al., 2019a;b;c; Ko et al., 2020; Adedinsewo et al., 2020; Christopoulos et al., 2020; Yao et al., 2021; Siontis et al., 2021; Cohen-Shelly et al., 2021; Bos et al., 2021; Grogan et al., 2021; Ahn et al., 2022). Supervised learning models of this nature demand extensive, high-quality datasets with precise annotations to achieve robust generalization on real-world data. Unfortunately, within the healthcare domain, acquiring such labeled datasets is challenging, as they are scarce, expensive, and time-consuming to obtain due to the necessity of trained physicians for the annotations. Motivated by the facts that, i) multi-modal contrastive learning applied to the general domain images (Radford et al., 2021; Jia et al., 2021; Goel et al., 2022) as well as the medical domain images (Zhang et al., 2022; Huang et al., 2021; Boecking et al., 2022; Bannur et al., 2023; Lu et al., 2023) by pairing images with text data has demonstrated promising results; ii) ECG signals contain information related to both cardiovascular and non-cardiovascular diseases (Venn et al., 2022); and iii) EHR data capture rich phenotypic information of the patients over time, we address the challenges described previously by leveraging EHRs. We pair structured EHR and unstructured EHR data with ECGs to learn ECG representations via multi-modal contrastive learning and multi-task learning. In particular, we utilize International Classification of Diseases (ICD) diagnosis codes, ICD procedure codes, medication prescriptions, lab test results, and echocardiogram measurements from the structured EHR category and text data from various sources such as ECG reports, ECHO reports, radiology reports, pathology reports, microbiology reports, clinical notes and surgical notes from the unstructured EHR category. Our contributions are summarised as follows:

1. We propose **sEHR-BERT**, a BERT model pre-trained to encode sEHR modality for use in multi-modal contrastive learning models.

2. We propose a set of three multi-modal contrastive learning models that combine sEHR, ECG, and text modalities to learn ECG representations:

   - **ECG-sEHR**: A model that combines ECG and sEHR modalities,
   - **ECG-Text**: A model that combines ECG and text modalities,
   - **sEHR-ECG-Text**: A model that combines sEHR, ECG, and text modalities.

3. We propose a large-scale multi-task learning model, **ECG-MTL**, which performs several classification and regression tasks simultaneously on a large number of diseases and lab test measurements, designed to play in the ECG domain akin to ImageNet's in the general image domain.

4. We then compare the effectiveness of the pre-trained models on downstream tasks such as linear classification, fine-tuning, zero-shot retrieval, and out-of-distribution detection with different baseline models including supervised learning models trained from scratch with random initialization and current state-of-the-art ECG-only self-supervised learning models.

2 Related Work

2.1 Contrastive Learning for General Domain Images

Self-supervised learning (SSL) using contrastive learning methods has emerged as a powerful pre-training technique to learn generic representations of the data. These methods learn representations either i) by pulling the embeddings of similar pairs (positive pairs) together and pushing the embeddings of dissimilar pairs (negative pairs) apart in the latent embedding space; or ii) by contrasting cluster assignments. Some of the notable works in computer vision include InfoNCE (Oord et al., 2018), SimCLR (Chen et al., 2020a), SimCLRv2 (Chen et al., 2020b), MoCo (He et al., 2020), SupCon (Khosla et al., 2020), SEER (Goyal et al., 2021), PIRL (Misra & Maaten, 2020), SwAV (Caron et al., 2020), and PCL (Li et al., 2021). These methods come under the category of uni-modal contrastive learning as they utilize only one type of data modality, i.e., images. Multi-modal contrastive learning by pairing general domain images with the corresponding image captions to learn image-text embeddings jointly in the shared space (Radford et al., 2021; Jia et al., 2021; Goel et al., 2022) has shown impressive results on downstream tasks such as zero-/few-shot learning.

2.2 Contrastive Learning for Medical Domain Images

Uni-modal contrastive learning based on SimCLR has been applied to medical domain images (Azizi et al., 2021; 2022; Ciga et al., 2022; Wang et al., 2022; Srinidhi & Martel, 2021; Sowrirajan et al., 2021) to learn medical image representations. Motivated by some of the initial works in uni-modal contrastive learning, ConVIRT (Zhang et al., 2022) proposed a multi-modal contrastive learning method by pairing chest radiology images with the corresponding radiology reports. Huang et al. (2021) extended ConVIRT for learning local and global representations by contrasting image sub-regions with words in the medical report. Boecking et al. (2022); Bannur et al. (2023) made improvements in the radiology domain by leveraging longitudinal medical images, building a domain-specific language model for radiology reports, and adding Masked Language Modeling (MLM) loss to contrastive loss during joint vision-language pre-training. Lu et al. (2023) applied multi-modal contrastive learning by pairing histopathology whole slide images with pathology reports. Our multi-modal contrastive learning models are largely inspired by ConVIRT (Zhang et al., 2022).

2.3 Contrastive Learning for Time Series ECG signals

SimCLR and the other aforementioned uni-modal contrastive learning models were developed for use in computer vision. However, they have been adopted for use with time series ECG signals in subsequent works (Kiyasseh et al., 2021; Diamant et al., 2022; Gopal et al., 2021; Mehari & Strodthoff, 2022; Oh et al., 2022). The principle difference between CLOCS (Kiyasseh et al., 2021), PCLR (Diamant et al., 2022) and the 3KG (Gopal et al., 2021) models is in the way the positive pairs are created. CLOCS treats consecutive non-overlapping segments and/or leads of the same ECG as positive pairs. PCLR treats two ECGs of the same patient as positive pairs. 3KG constructs positive pairs by applying spatial augmentations such as rotation and scaling in vectorcardiogram (VCG) space after converting ECG to VCG, followed by temporal augmentations such as time masking in ECG space after converting VCG back to ECG. Cheng et al. (2020) introduced adversarial training to address intersubject variability while learning ECG representations using contrastive learning. Mehari & Strodthoff (2022) adapted SimCLR (Chen et al., 2020a), CPC (Oord et al., 2018), and SwAV (Caron et al., 2020) to ECG domain to learn ECG representations. Recently, Oh et al. (2022) proposed a pre-training method that combines CMSC from Kiyasseh et al. (2021) and Wave2Vec 2.0 (Baevski et al., 2020) from speech domain to learn local and global contextual ECG representations.

To the best of our knowledge, there is only one work that combines ECGs with other modalities. Raghu et al. (2022) developed a SimCLR-like contrastive learning model that was pre-trained using multi-modal clinical time series data such as ECG signals and structured time series data (labs and vitals). The model utilizes 18-dimensional structured time series data from metabolic panel, blood pressures, heart rate, and SpO2. The model is shown to have achieved improved or comparable performance over training from scratch on two downstream tasks: (i) Elevated mPAP; and (ii) 24-hour mortality rate. To the best of our knowledge, we are the first to fully utilize the large landscape of electronic health records to learn ECG representations.

3 Methods

3.1 sEHR-BERT: Structured EHR Model Pre-training

Several methods have been proposed to model structured EHRs based on BERT (Devlin et al., 2018): BEHRT (Li et al., 2020), Med-BERT (Rasmy et al., 2021), CEHR-BERT (Pang et al., 2021), and CEHR-GAN-BERT (Poulain et al., 2022). However, none of these pre-trained models are publicly available to use in our work. Moreover, the vocabulary in our dataset may not be aligned with the vocabulary of the mentioned models. So we developed sEHR-BERT, a model pre-trained to encode sEHR data and produce sEHR representations based on the BERT architecture (Devlin et al., 2018). We used a vocabulary of size 28593, constructed from ICD diagnosis codes, ICD procedure codes, and medication prescriptions. These are collectively referred to as medical codes in this work. The input to the model is a sequence of medical codes sorted in ascending order based on medical codes' timestamps. Each code is processed by adding its corresponding medical code embedding, time embedding, and medical code type embedding and sent to the transformer encoder. Time embeddings are constructed in a weekly manner based on the medical code's timestamp, which means that all codes that were generated in the same week have the same time

embedding. Medical code type embeddings are divided into different categories (i.e., diseases, symptoms, procedures, special tokens, etc.). We used a custom BERT model with the number of layers, hidden size, and number of self-attention heads set to 5, 320, and 5 respectively. This model has 15M parameters. We initialize the model weights randomly and follow the BERT (Devlin et al., 2018) pre-training strategy, i.e., Masked Language Modeling (MLM) to learn the representations of the structured EHR sequences. We minimize the MLM loss given by $\mathcal{L} = -\frac{1}{K}\sum_{i=1}^{K} \log p\left(D_{m_i}|D_{\tilde{M}};\Theta\right)$, where Θ are parameters of the model, $D = \{D_0, D_1, ..., D_N\}$ is the sequence of medical codes of length N, $M = \{m_0, m_1, ..., m_K\}$ are indices of masked medical codes, and $D_{\tilde{M}}$ denotes the set of unmasked medical codes. During training, the medical codes are masked with a probability of 15%, and the model is trained with AdamW (Loshchilov & Hutter, 2019) optimizer and batch size of 512 for 100 epochs. We set an initial learning rate of 5e-4 and the learning rate is reduced by a factor of 2 if the validation loss stops decreasing continuously for 2 epochs.

3.2 sEHR-ECG-Text: Joint sEHR, ECG and Text Pre-training

In the sEHR-ECG-Text model, we pair the ECG modality with sEHR and text modalities to jointly learn multi-modal representations. MultiModal Versatile Networks (MMV) (Alayrac et al., 2020), applied contrastive learning to video, audio, and text multi-modal data under the assumption that the video and audio modalities are more granular than the text modality. MMV assumes that applying contrastive loss in shared embedding space does not maintain the specificities, so two embedding spaces are learned i.e. a fine-grained embedding space where video and audio are matched and a coarse-grained embedding space where text is matched with video and audio domains. We hypothesize that sEHR, ECG, and text modalities do not exhibit the same level of granularity. Moreover, the ECGs are paired with sEHR and text data in a given time window surrounding the ECG acquisition timestamp, and tokens are trimmed based on the input length accepted by the corresponding encoders as we describe in Sections 4.2.2 and 4.2.3. This implies that the same level of information might not be maintained between sEHR and text, so we follow the framework from MMV in our sEHR-ECG-Text model and compare ECG with sEHR in fine-grained joint ECG-sEHR embedding space and ECG with sEHR and text in coarse-grained joint sEHR-ECG-Text embedding space.

We consider a dataset $S = \mathcal{X}_s \times \mathcal{X}_e \times \mathcal{X}_t$ consisting of triplets $\{(x_s^i, x_e^i, x_t^i)\}_{i=1}^{M}$ where x_s^i is the sEHR sequence of the i-th example, x_e^i is the ECG waveform of the i-th example, and x_t^i is the text sequence of the i-th example, M is the total number of examples in the training set and $\mathcal{X}_s$, $\mathcal{X}_e$, and $\mathcal{X}_t$ denote the domain of sEHR, ECG, and text respectively.

Let $E_m : \mathcal{X}_m \rightarrow \mathbb{R}^{d_m}$ be a parameterized model mapping from modality m to a modality-specific embedding of dimension d_m, where m can be s, e, t for sEHR, ECG and text respectively. Let Ω_z be a shared embedding space of different modalities where modality-specific representations are projected into to maximize or minimize the alignment between different modalities using the contrastive loss objective. For example, Ω_{es}, Ω_{et}, and Ω_{set} denotes ECG-sEHR, ECG-Text and sEHR-ECG-Text shared embedding spaces respectively. Let $P_{m\rightarrow z} : \mathbb{R}^{d_m} \rightarrow \mathbb{R}^{d_z}$ be a projection network, mapping from the representation of modality m of dimension d_m to the representation of dimension d_z in shared embedding space Ω_z. In order to obtain the modality-specific representations, we use a convolutional neural network (CNN) customized to one dimension (E_e) for the ECG modality, the pre-trained sEHR-BERT (E_s) as described in Section 3.1 for the sEHR modality, and pre-trained GatorTron (Yang et al., 2022) (E_t) for the text modality. Global average pooling is applied at the final layer for all three encoders to obtain the representations. We use multi-layer perceptron (MLP) for the projection network to embed modality-specific representations into shared space. We apply the contrastive loss between ECG and sEHR in ECG-sEHR joint embedding space Ω_{es}, and contrastive loss between ECG and text in sEHR-ECG-Text joint embedding space Ω_{set} so that specificities are maintained.

Let v_m^i be the representation of x_m^i obtained by passing x_m^i into modality specific encoder E_m, i.e., $v_m^i = E_m(x_m^i)$, $v_{m,z}^i$ be the representation of x_m^i in the shared embedding space Ω_z obtained by passing v_m^i into projection network $P_{m\rightarrow z}$, i.e., $v_{m,z}^i = P_{m\rightarrow z}(v_m^i)$. To project a representation from one shared space to another shared space, a different projection network is used. For example, to embed ECG representation in ECG-sEHR shared space ($v_{e,es}^i$) into sEHR-ECG-Text shared space ($v_{e,set}^i$) to compare ECG modality with text modality, we utilize a projection network $P_{es\rightarrow set} : \mathbb{R}^{d_z} \rightarrow \mathbb{R}^{d_z}$, i.e., $v_{e,set}^i = P_{es\rightarrow set}(v_{e,es}^i)$. We assume that all the representations that are generated at different levels are L_2 normalized, and we define

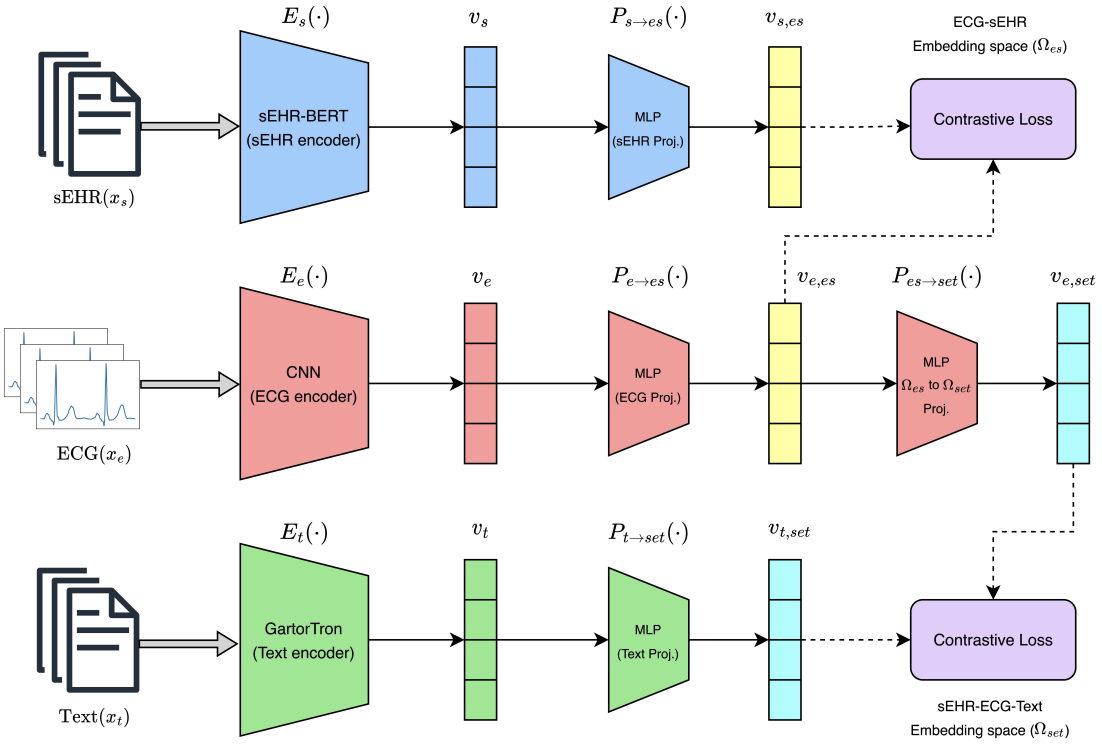

Figure 1: Overview of the pre-training process for the sEHR-ECG-Text model. The model uses a contrastive loss method that applies between the ECG and sEHR in the ECG-sEHR joint embedding space, as well as between the ECG and text in the ECG-sEHR-Text joint embedding space. During ECG-sEHR pre-training, the text branch is not used, while during ECG-Text pre-training, the sEHR branch is not used. In both ECG-sEHR and ECG-Text pre-training, the $P_{es\to set}(\cdot)$ projection is not used.

the cosine similarity between two L_2-normalized vectors $x, y \in \mathbb{R}^{d_z}$ as, $sim(x, y) = x^T \cdot y$. Following Zhang et al. (2022), we define the contrastive objective in a bidirectional manner, i.e., in the case of contrastive loss between ECG and sEHR domains, the loss is directed from ECG to sEHR as well as from sEHR to ECG and similarly between ECG and text domains. In a given minibatch of size N, N ECG-sEHR (x_e, x_s) pairs are considered positive, while the remaining $N^2 - N$ pairs are treated as negative. This same approach is applied to ECG-Text (x_e, x_t) pairs. Let $\mathcal{L}_{es}$ be the contrastive loss between ECG and sEHR, $\mathcal{L}_{et}$ be the contrastive loss between ECG and text, $\mathcal{L}$ be the overall contrastive loss, λ_{es} and λ_{et} be scalar weights $\in [0, 1]$ and $\tau \in \mathbb{R}^+$ be the temperature parameter, then

$$\mathcal{L}_{es} = -\frac{1}{N}\sum_{i=1}^{N}\left(\lambda_{es}\log\frac{\exp\left(sim(v^i_{e,es}, v^i_{s,es})/\tau\right)}{\sum_{k=1}^{N}\exp\left(sim(v^i_{e,es}, v^k_{s,es})/\tau\right)} + (1-\lambda_{es})\log\frac{\exp\left(sim(v^i_{s,es}, v^i_{e,es})/\tau\right)}{\sum_{k=1}^{N}\exp\left(sim(v^i_{s,es}, v^k_{e,es})/\tau\right)}\right) \quad (1)$$

$$\mathcal{L}_{et} = -\frac{1}{N}\sum_{i=1}^{N}\left(\lambda_{et}\log\frac{\exp\left(sim(v^i_{e,set}, v^i_{t,set})/\tau\right)}{\sum_{k=1}^{N}\exp\left(sim(v^i_{e,set}, v^k_{t,set})/\tau\right)} + (1-\lambda_{et})\log\frac{\exp\left(sim(v^i_{t,set}, v^i_{e,set})/\tau\right)}{\sum_{k=1}^{N}\exp\left(sim(v^i_{t,set}, v^k_{e,set})/\tau\right)}\right) \quad (2)$$

the combination of which gives the overall loss and we aim to minimize this loss.

$$\mathcal{L} = \mathcal{L}_{es} + \mathcal{L}_{et} \quad (3)$$

3.3 ECG-sEHR: Joint ECG and sEHR Pre-training

In the ECG-sEHR model, we pair ECG signals with structured EHRs as we describe in more detail in Sections 4.2.1 and 4.2.2. We apply contrastive objective between ECG and sEHR modalities in joint ECG-sEHR embedding space (Ω_{es}), where we minimize the contrastive loss given in Equation 1.

3.4 ECG-Text: Joint ECG and Text Pre-training

In the ECG-Text model, we pair ECG signals with clinical text data from unstructured EHRs as we describe in more detail in Section 4.2.3. ECG and text embeddings are jointly learned by applying the contrastive objective between ECG and text modalities in joint ECG-Text embedding space (Ω_{et}). Following the notation introduced in Section 3.2, we minimize the contrastive loss given by,

$$\mathcal{L}_{et} = -\frac{1}{N}\sum_{i=1}^{N}\left(\lambda_{et}\log\frac{\exp\left(sim(v_{e,et}^{i}, v_{t,et}^{i})/\tau\right)}{\sum_{k=1}^{N}\exp\left(sim(v_{e,et}^{i}, v_{t,et}^{k})/\tau\right)} + (1-\lambda_{et})\log\frac{\exp\left(sim(v_{t,et}^{i}, v_{e,et}^{i})/\tau\right)}{\sum_{k=1}^{N}\exp\left(sim(v_{t,et}^{i}, v_{e,et}^{k})/\tau\right)}\right) \quad (4)$$

3.5 ECG-MTL: Large Scale Multi-Task Learning Model

ECG-MTL model is a Multi-Task Learning (MTL) model trained in a supervised manner to simultaneously predict 81 classification and 37 regression tasks with labels sourced from ECG reports, ECHO reports, echocardiogram measurements, lab tests, age, and gender. In particular, for classification tasks, for each ECG, we obtain the patient's gender (male/female), the presence or absence of 75 heart conditions from ECG reports commonly diagnosed by physicians using ECGs (e.g., left atrial enlargement, atrial fibrillation, left bundle branch block, etc.), and the presence or absence of 5 valvular heart diseases from ECHO reports (i.e., mitral regurgitation, tricuspid regurgitation, aortic regurgitation, aortic stenosis, and mitral stenosis). For regression tasks, we obtain the patient's age at ECG acquisition timestamp, 14 heart specifications from ECHO measurements (e.g., left ventricular ejection fraction, aortic valve area, tricuspid regurgitation velocity, etc.), and 22 blood, urine, other bodily fluid or tissue values, etc. from lab test records (e.g., serum-potassium, hemoglobin measurements, etc.). The motivation behind the MTL model is that common ECG patterns across tasks will be learned better when compared to the individual models. This is achieved by having a shared backbone network and a set of task-specific MLP heads on top of the backbone network for each classification and regression task. We apply binary cross entropy loss for classification and mean squared error loss for regression tasks and we minimize the loss obtained by adding the losses across tasks using the *random loss weighting* strategy (Lin et al., 2022). We provide more details about the tasks used and the performance of each task in Appendix E.

4 Experiments and Results

4.1 Dataset Splits Setup

We used De-Identified EHR data of around 2.4 *million* patients from a medical center consisting of around 9 *million* ECGs to create datasets for pre-training and downstream tasks. We initially split all the patients into the global train, validation, and test sets in a 60%, 5%, and 35% ratio, which are then used to create pre-training datasets and disease cohorts for downstream classification tasks. In particular train, validation, and test sets for pre-training and classification tasks are created by drawing the EHRs from the global train, validation, and test patients respectively. In this way, the representation quality can be evaluated more effectively on downstream tasks as the validation and test patients' data was not seen in the pre-training stage. This means that all datasets across tasks will have train, validation, and test split percentages roughly close to 60%, 5%, and 35% respectively.

4.2 Pre-training Datasets

In this section, we describe the creation of i) the sEHR sequences for sEHR-BERT pre-training; ii) ECG-sEHR (x_e, x_s) pairs for joint pre-training of ECG and sEHR modalities; and iii) ECG-Text (x_e, x_t) pairs for joint ECG-Text pre-training, where x_e represents the ECG signal, x_s represents the sEHR sequence, and x_t represents text sequence. For joint sEHR-ECG-Text pre-training, we construct the triplets (x_s, x_e, x_t) by considering (x_e, x_s) and (x_e, x_t) pairs that have ECG paired with both sEHR and text data. We provide the number of ECGs, number of patients in global splits and pre-training dataset splits in Table 8.

4.2.1 Dataset for sEHR-BERT Pre-training

We utilize ICD-9 (International Classification of Diseases, Ninth Revision), ICD-10 (International Classification of Diseases, Tenth Revision), CPT (Current Procedural Terminology), HCPS (Healthcare Common Procedures Coding System) codes, and medication prescriptions to create the dataset for sEHR-BERT pre-training. Since ICD-9 codes are different from ICD-10 codes, but the corresponding text descriptions are similar, we do a mapping from ICD-9 to ICD-10 to maintain the same phenotypic information. ICD-10 diagnosis codes are shortened to the first three characters as keeping four or more characters provides little to no extra information for large-scale pre-training. For example, the corresponding text descriptions of ICD-10 diagnosis codes I26.0 and I26.9 are pulmonary embolism with acute cor pulmonale and pulmonary embolism without acute cor pulmonale respectively, but these come under a common disease category, i.e., pulmonary embolism (I26). Shortened ICD-10 diagnosis codes, ICD-10 procedure codes, CPT codes, HCPS codes, and medication prescriptions that have an association with at least 50 patients are considered in the vocabulary. We present short ICD-10 vs full ICD-10 diagnosis codes ablation by keeping codes from other sources constant in Section 4.6. A vocabulary of size 28593 is constructed based on shortened ICD-10 diagnosis codes, ICD-10 procedure codes, CPT codes, HCPS codes, and medication prescriptions for the main results. To create the sEHR sequence for sEHR-BERT model pre-training we randomly select one sequence of up to 512 consecutive medical codes from a given patient's timeline. On average, the sequence length of the resulting dataset is 168.

4.2.2 ECG-sEHR Pairs Generation

To create the ECG-sEHR (x_e, x_s) pairs, we first select an ECG of a given patient, x_e, and consider all the shortened ICD-10 diagnosis codes, ICD-10 procedure codes, CPT codes, HCPS codes, and medication prescriptions associated with that patient within a period of one year prior, and one year subsequent, to the acquisition timestamp of that ECG. The medical codes restricted to this time range are arranged sequentially to form the sEHR input sequence x_s. The average length of the constructed sequences is 121.

4.2.3 ECG-Text Pairs Generation

ECGs are paired with unstructured EHR text data from multiple sources. These include ECG reports, ECHO reports, pathology reports, radiology reports, microbiology reports, clinical notes, and surgical notes. These are collectively referred to as patient notes in this work. Although GatorTron-base (Yang et al., 2022) model accepts sequences of length up to 512, we were only able to use a maximum of 400 tokens after the tokenization of the text because of computing resource constraints. At first, for a selected ECG x_e of a given patient, we take all the reports from patient notes that are available in a specific time window before and after the ECG acquisition timestamp. To incorporate patient information from a longer timeline, we select only the reports that contain an entity belonging to a pre-determined list of bio-medical entity types, (i.e., diseases, symptoms, procedures, medications, biomarkers, and gene mutations) using an in-house NLP model. This step is referred to as *entity detection*. Through this, we eliminate irrelevant data, capture long-term dependencies, and potentially improve the quality of the representations. We then use two methods for pairing ECGs with selected patient notes: i) report concatenation; and ii) entity concatenation. These methods differ in how patient notes are processed after *entity detection*. In report concatenation, we use a time window of 1 month and concatenate all the reports containing bio-medical entities. This resulted in an average sequence length of 354 after tokenization. In entity concatenation, we use a time window of 1 year and concatenate not the reports containing entities, but only the entities themselves to further extend the

Table 1: ECG count, unique patient count, and disease prevalence for each downstream task cohort.

	Train			Validation			Test		
Disease	#ECGs	#Patients	Prev.(%)	#ECGs	#Patients	Prev.(%)	#ECGs	#Patients	Prev.(%)
Coronary atherosclerosis	19,281	10,589	38.78	1,604	870	39.66	11,290	6,088	39.13
Myocarditis	53,432	52,299	0.97	4,366	4,260	1.17	31,715	30,984	1.02
Cardiac amyloidosis	34,465	20,011	8.54	2,795	2,462	8.04	20,071	17,508	8.51
Pulmonary hypertension	200,777	73,908	14.71	16,132	6,091	14.10	115,602	42,893	14.34
Low LVEF	166,702	166,702	7.58	13,814	13,814	7.69	97,109	97,109	7.48
AFib in NSR	1,455,626	514,871	7.28	42,627	42,627	7.45	301,022	301,022	7.32

timeline. This resulted in an average sequence length of 266 after tokenization. We use entity concatenation for the main results as it yielded better results when compared to report concatenation. We also present the report concatenation vs entity concatenation ablation in Section 4.6.

4.3 Classification Datasets

In this section, we provide the details of the classification datasets that are used in linear classification and fine-tuning tasks. We evaluate the pre-trained models on both internal and external datasets.

4.3.1 Internal Datasets

We target six cardiac diseases whose diagnostic criteria using ECGs haven't been established in clinical practice, i.e., either the patterns to identify these diseases from ECG are not known or ECG is not the gold standard for definitive diagnosis. These include coronary atherosclerosis, myocarditis, cardiac amyloidosis, pulmonary hypertension (PH), low left ventricular ejection fraction (low LVEF i.e., LVEF$\leq$40), and atrial fibrillation in normal sinus rhythm (AFib in NSR). These diseases are diagnosed by other means and the diagnostic information is available in EHRs. For example, to diagnose PH, an invasive, right heart catheterization (RHC) procedure is performed and to identify low LVEF, an echocardiogram test is performed. we utilize EHRs to associate ECGs with these diseases and generate labels. In particular, we used the following works to create the disease cohorts: cardiac amyloidosis (Grogan et al., 2021), low LVEF (Attia et al., 2019a), PH (Wagner et al.), and AFib in NSR (Attia et al., 2019c). In order to create the coronary atherosclerosis disease cohort, we utilized the coronary artery calcium score (CAC score) which is obtained through computed tomography (CT) test in clinical settings. The positive group was defined as having a CAC score greater than 300. Controls were defined as patients who have undergone a CT with no observed coronary artery calcification (i.e., CAC score of 0) The cohort details describing patient count, ECG count and prevalence for each disease are given in Table 1.

4.3.2 External Datasets

We also test all pre-trained models on two publicly available datasets: i) **PhysioNet2020** (Alday et al., 2020), which is itself a collection of six 12-lead ECG datasets sourced across the world with varying signal lengths and sampling rates; ii) **Chapman** (Zheng et al., 2020), which contains 10-second long, 12-lead ECGs of 10,646 patients. The details about each dataset are given in Table 10. Following Gopal et al. (2021) we merge some of the conditions from the list of 27 conditions in PhysioNet2020 due to their similarity, i.e., complete right bundle branch block and right bundle branch block, premature atrial contraction and supraventricular premature beats, premature ventricular contractions and ventricular premature beats, 1st-degree atrioventricular block and prolonged PR interval, and evaluate on 23 distinct classes. Following Zheng et al. (2020), we also merge 11 cardiac arrhythmia conditions of the Chapman dataset into 4 major classes. The conditions in these datasets are commonly diagnosed by physicians directly using ECG, unlike the diseases in our proprietary internal classification datasets. During evaluation, we resample the datasets whose sampling rate is not 500Hz to 500Hz and we take a 5-second long random crop from each ECG record. We split both datasets into 60%, 10%, and 30% for training, validation, and testing respectively.

4.4 Baseline Models

Random initialization models We train binary classification models on all individual diseases from scratch using the same ECG encoder used while pre-training by randomly initializing the weights.

ECG-only contrastive learning models We compare our models with the current state-of-the-art ECG-only self-supervised learning models. In particular, we compare against the 3KG (Gopal et al., 2021), CLOCS (CMSC) (Kiyasseh et al., 2021) and PCLR (Diamant et al., 2022) models. For identical comparison, we pre-train all three models using the same global splits that we used for the multi-modal contrastive pre-training but with only ECG signal as input and we also use the same ECG encoder that we used while pre-training multi-modal contrastive learning models.

4.5 Downstream Tasks and Results

In this section, we evaluate the pre-trained models' transfer learning capabilities and compare them with different baseline methods on various downstream tasks such as classification, zero-shot retrieval, and out-of-distribution (OOD) detection.

4.5.1 Classification Tasks

We evaluate the pre-trained models on linear classification and fine-tuning tasks. After pre-training, we assess the quality of the representations by extracting embeddings from the pre-trained ECG encoder by passing ECG waveform as input and then employ logistic regression models to train on various cardiovascular diseases using both internally created and publicly available datasets as outlined in Section 4.3. In fine-tuning tasks, we add a classification head (MLP) on top of the pre-trained ECG encoder and fine-tune the entire network. We compare our pre-trained models with various baseline models as described in Section 4.4. One of the most useful applications of pre-trained models is in providing downstream tasks with **data efficiency**, which refers to a model's performance remaining roughly constant despite a reduction in the amount of data used to train it. This is very valuable when a large amount of labeled data is not available due to the low prevalence of the diseases or is too expensive to procure. To demonstrate the data efficiency of different pre-training methods, we create different fractions (1%, 10%, 100%) of the training set by maintaining roughly the same prevalence as the original prevalence of the full training set. For low data environment diseases such as coronary atherosclerosis, myocarditis, and cardiac amyloidosis, we drop 1% split. This is because even the complete cohort sizes are relatively small. We use the area under the ROC curve (AUROC) as our evaluation metric. We execute five separate runs with different seeds for each split of each disease and report the average AUROC score. In the case of internal datasets, we report the AUROC score using a single run on the 100% split to ensure that train, validation, and test sets remain separate, as specified in section 4.1. On external datasets, we report the AUROC score averaged over all disease classifiers.

Results Table 3 shows the performance (AUROC) of various disease linear classifiers and Table 4 shows the performance (AUROC) of various disease fine-tuned classification models. We also present linear classification results on external datasets in Table 2. i) We observe that linear classifiers trained using representations obtained from our pre-trained models consistently outperform all baseline models by large margins across different fractions for all diseases. ii) We find that classification models fine-tuned by initializing with our pre-trained models' weights outperform all baseline models by a large margin in low-data environments and by a small margin in high-data environments. ii) We have noticed that our pre-trained models, when trained on just 10% of the training data for various diseases,

Table 2: Linear classification results (AUROC, averaged over five independent runs) on external datasets, i.e., PhysioNet2020 (23 classes) and Chapman (4 classes).

	PhysioNet2020			Chapman		
Method	1%	10%	100%	1%	10%	100%
ECG-only SSL						
3KG	67.51	76.37	85.71	90.45	96.31	98.50
CLOCS(CMSC)	66.26	74.49	85.79	82.34	89.10	94.37
PCLR	65.26	75.43	83.24	72.34	82.21	90.98
Our models						
sEHR-ECG-Text	70.71	79.27	89.20	94.02	97.38	98.92
ECG-sEHR	68.60	78.20	88.02	89.23	95.54	98.07
ECG-Text	**73.00**	**81.40**	**89.49**	**97.09**	**98.33**	**99.14**
ECG-MTL	69.18	75.38	88.86	93.44	97.03	98.84

Table 3: Linear classification performance (AUROC) on coronary atherosclerosis, myocarditis, cardiac amyloidosis, pulmonary hypertension, low LVEF, AFib in NSR diseases across different fractions of the training set, i.e. 1%, 10%, 100%. Results are compared with random initialization and ECG-only SSL methods.

	Coronary atherosclerosis		Myocarditis		Cardiac amyloidosis		Pulmonary hypertension			Low LVEF			AFib in NSR		
Method	10%	100%	10%	100%	10%	100%	1%	10%	100%	1%	10%	100%	1%	10%	100%
Supervised Baseline															
Random Init.	77.06	82.55	74.63	86.32	91.42	94.47	81.16	88.79	92.85	85.35	92.06	94.38	83.84	89.62	92.19
ECG-only SSL															
3KG	72.77	79.14	72.28	82.90	87.21	90.82	79.63	85.98	87.22	85.06	90.23	91.46	82.61	86.56	86.78
CLOCS (CMSC)	74.70	80.41	66.45	78.21	88.22	91.68	78.95	85.69	87.08	86.42	90.40	91.62	81.99	86.09	86.55
PCLR	76.46	82.58	71.91	82.10	89.85	92.71	84.57	89.51	90.70	89.33	92.66	93.59	86.45	89.62	90.08
Our models															
sEHR-ECG-Text	**84.16**	88.98	86.05	89.85	**93.66**	**95.84**	90.08	93.09	93.85	**91.00**	94.05	95.13	90.03	92.49	92.31
ECG-sEHR	83.68	**89.05**	**86.48**	**90.25**	93.37	**95.84**	**90.40**	93.21	93.99	90.98	94.09	95.08	**90.06**	**92.67**	**92.56**
ECG-Text	82.27	87.52	79.91	88.84	92.36	94.77	89.30	92.41	93.06	90.25	93.67	94.68	89.47	91.89	91.53
ECG-MTL	80.55	86.04	76.29	88.06	92.49	95.72	89.39	**93.30**	**94.08**	90.17	**94.52**	**95.56**	88.35	91.45	91.56

Table 4: Fine-tuned performance (AUROC) on coronary atherosclerosis, myocarditis, cardiac amyloidosis, pulmonary hypertension, low LVEF, AFib in NSR diseases across different fractions of the training set, i.e. 1%, 10%, 100%. Results are compared with random initialization and ECG-only SSL methods.

	Coronary atherosclerosis		Myocarditis		Cardiac amyloidosis		Pulmonary hypertension			Low LVEF			AFib in NSR		
Method	10%	100%	10%	100%	10%	100%	1%	10%	100%	1%	10%	100%	1%	10%	100%
Supervised Baseline															
Random Init.	77.06	82.55	74.63	86.32	91.42	94.47	81.16	88.79	92.85	85.35	92.06	94.38	83.84	89.62	92.19
ECG-only SSL															
3KG	74.31	82.70	72.58	85.30	90.37	94.56	79.46	89.02	92.58	86.83	92.12	94.38	83.94	89.93	92.10
CLOCS (CMSC)	78.16	80.23	72.92	86.69	90.86	94.89	83.36	89.22	92.55	89.54	92.57	94.36	84.95	89.75	91.88
PCLR	80.65	83.51	71.50	85.57	92.67	95.20	86.89	91.15	93.23	90.55	93.76	94.87	87.48	90.54	92.08
Our models															
sEHR-ECG-Text	87.92	89.12	86.04	**90.77**	94.80	96.32	91.61	93.34	93.80	93.37	94.16	95.10	91.28	91.76	92.70
ECG-sEHR	**88.01**	**89.35**	**86.08**	90.59	**94.90**	**96.37**	91.85	93.53	94.16	93.19	94.11	94.99	**91.74**	91.89	92.74
ECG-Text	86.86	88.08	85.39	89.35	94.09	95.98	90.75	92.49	93.40	92.85	93.62	94.92	90.67	91.45	92.35
ECG-MTL	85.16	86.20	83.31	87.23	94.51	96.14	**92.19**	**93.65**	**94.31**	**93.89**	**94.79**	**95.47**	89.85	**91.98**	**92.93**

produce classification results that are just as good or even better (highlighted in boldface) than those obtained by training on all the training data using the baseline models, showing the effectiveness of our pre-trained models for data efficiency. iv) We also show that our models consistently outperform all baseline models on external datasets on linear classification tasks, showing the generalization ability to other datasets. v) We observe that the ECG-MTL model performs the best on low LVEF and PH diseases. We speculate the reason for this is, LVEF which is used to diagnose low LVEF, i.e., LVEF$\leq$40, and tricuspid regurgitation velocity (TRV) which is a suggestive factor to diagnose PH, although not accurate, are part of the regression tasks in the ECG-MTL model.

4.5.2 Retrieval Tasks

Following (Zhang et al., 2022), we also evaluate the pre-trained models on two zero-shot retrieval tasks: i) Zero-shot ECG-ECG Retrieval, ii) Zero-shot Text-ECG Retrieval. We used data only from the global test split to create the queries and candidates for the retrieval tasks. For a given query, we rank the candidates by computing the cosine similarity between the representations of the query and the candidates obtained from pre-trained encoders. For the Text-ECG retrieval task, we obtain ECG and text embeddings from shared ECG-Text embedding space Ω_{et}. We report the precision@k metric for k=100, 500, and 1000, which represents the percentage of top k ranked candidates that are relevant to the query.

Zero-shot ECG-ECG Retrieval We take 1000 different ECGs as search queries for each of the 41 cardiovascular conditions that are based on ECG reports. For every condition, we select 100,000 candidate ECGs, of which 10,000 are classified as positive for the condition and 90,000 are classified as negative for the condition. The query ECGs and positive candidate ECGs are completely exclusive.

Zero-shot Text-ECG Retrieval We choose 1000 distinct ECG reports as search queries for each of the 41 cardiovascular conditions. For each of the conditions, we choose 100,000 ECGs, out of which 10,000 ECGs show the condition and 90,000 ECGs have no connection to the condition. We also make sure that no ECG corresponding to the 1000 distinct ECG report queries is chosen as a candidate for the positive set.

Table 5: Zero-shot ECG-ECG retrieval and Text-ECG retrieval results. The *random* category results are from random guesses. P@k denotes Precision@k.

	ECG-ECG Retrieval			Text-ECG Retrieval		
Method	P@100	P@500	P@1000	P@100	P@500	P@1000
Random	10.00	10.00	10.00	10.00	10.00	10.00
PCLR	38.41	35.34	33.74	-	-	-
3KG	40.35	37.34	35.74	-	-	-
CLOCS	41.13	37.63	35.82	-	-	-
ECG-Text	**55.13**	**49.47**	**46.33**	**73.02**	**66.92**	**63.58**

Table 5 shows the zero-shot ECG-ECG retrieval and ECG-Text retrieval results. Our multi-modal ECG-Text model outperforms random (random guess) and ECG-only contrastive learning models by a large margin on both tasks.

4.5.3 Out-of-Distribution Detection

It is observed that the representations learned via self-supervised learning techniques help to better distinguish between in-distribution (IND) and out-of-distribution (OOD) datasets. We demonstrate this using representations obtained from our ECG-sEHR model to differentiate between two disparate ECG datasets. We take the proprietary ECG pulmonary hypertension (PH) cohort as the IND dataset and Holter ECGs (ECG recorded continuously over 24 hours or longer) from the open-source St Petersburg INCART 12-lead Arrhythmia Database (Tihonenko et al., 2008) as the OOD dataset. Non-overlapping 10-second long segments are taken from Holter ECGs and resampled to 500Hz to be consistent with ECGs from the IND PH dataset. We use the *relative Mahalanobis distance* (RMD) (Ren et al., 2021) method which is based on the Mahalanobis distance of embeddings from the distribution of the nearest predicted class, to determine whether the data is in-distribution or out-of-distribution. We compare the results obtained using representations extracted from the ECG-encoder of the pre-trained ECG-sEHR model with that of the representations obtained from the penultimate layer of the PH binary classifier trained from scratch on the PH cohort and show that the rejection rate at different significance levels is much higher in the case of ECG-sEHR model. The results are given in Table 6, which clearly shows that generic ECG representations are better at detecting out-of-distribution data compared to disease-specific representations.

Table 6: Out-of-distribution detection results using ECG representations obtained from the PH disease model and generic ECG representations obtained from the ECG-sEHR model. *sig.* denotes significance level.

Metric	ECG-sEHR	PH
Rejection at 1% sig.	13.94	10.35
Rejection at 5% sig.	49.67	29.87
IND vs OOD AUC	75.70	62.00

4.6 Ablation Study

We perform two ablations: i) short ICD-10 vs full ICD-10 codes in sEHR modality as described in 4.2.2; and ii) report concatenation vs entity concatenation in text modality as described in 4.2.3. We performed these two ablations using bi-modal contrastive learning models to understand the affect of each component. We performed short ICD-10 vs full ICD-10 codes ablation using ECG-sEHR model and report concatenation vs entity concatenation ablation using ECG-Text model. A vocabulary of size 28593 and 42355 is used for short ICD-10 and full ICD-10 codes respectively. we show the ablation study results on the linear classification task in Table 7. The difference in performance between short ICD-10 and full ICD-10 codes is very minimal which can be attributed to the point that full ICD codes provide little to no extra information, whereas in report concatenation vs entity concatenation ablation, entity concatenation yielded better results in the

Table 7: Ablation study results. Linear classification performance (AUROC) comparison between short ICD codes vs full ICD codes and report concatenation vs entity concatenation.

Method	Coronary atherosclerosis	Myocarditis	Cardiac amyloidosis	Pulmonary hypertension	Low LVEF	AFib in NSR
ECG-sEHR						
Short ICD-10 codes	**89.05**	**90.25**	**95.84**	93.99	95.08	92.56
Full ICD-10 codes	88.87	90.10	95.66	**94.00**	**95.14**	**92.72**
ECG-Text						
Report concatenation	**88.18**	**89.72**	93.41	91.85	93.44	91.19
Entity concatenation	87.52	88.84	**94.77**	**93.06**	**94.68**	**91.53**

majority of the diseases. We speculate that the reason for this is, entity concatenation captures better long-term dependencies than report concatenation as we incorporate information from 1 year around ECGs.

5 Conclusion

The EHRs utilized in this study have undergone a rigorous de-identification process, guaranteeing the utmost privacy and data security. These records have received approval from the Institutional Review Board (IRB) of the medical center, ensuring compliance with ethical guidelines and regulations. Consequently, there are no privacy or data security issues associated with the use of these de-identified EHRs. Our work introduces a series of three multi-modal contrastive learning models and a multi-task learning model. These models leverage both structured and unstructured EHRs to produce high-quality ECG representations. We have demonstrated that our pre-trained models outperform randomly initialized models and other ECG-only contrastive learning models by a wide margin on classification and retrieval tasks. Specifically, we perform the classification tasks using ECGs on cardiovascular diseases whose definitive diagnoses are obtained from more expensive and/or invasive tests in clinical settings. This is a significant breakthrough as ECG tests are widely available, non-invasive, and less expensive. Furthermore, our ECG representations have been shown to excel in detecting out-of-distribution data when compared to disease-specific representations.

6 Future Work

In this work, we make use of both structured EHRs and unstructured EHR text data to learn ECG representations. However, there are additional modalities present in unstructured EHRs such as images (MRI scans, CT scans, X-rays, and histopathology images related to cardiology), videos (echocardiograms/heart ultrasounds), and time-series signals (heart and lung sounds), which can provide even more meaningful information through multi-modal contrastive learning. Another interesting future work would be the integration of federated learning frameworks to leverage multi-institutional medical data. This approach aims to capture a more diverse range of patient information, leading to enhanced ECG representation learning. While the disease models presented in this study undergo training and testing using real-world datasets, it is of utmost importance to conduct clinical validation across a diverse set of health systems before deploying them. This ensures that the models are equitable, unbiased, and trustworthy. We hope our work will serve as an inspiration for future endeavors in harnessing multi-modal EHR data to learn robust ECG representations.

References

Demilade Adedinsewo, Rickey E Carter, Zachi Attia, Patrick Johnson, Anthony H Kashou, Jennifer L Dugan, Michael Albus, Johnathan M Sheele, Fernanda Bellolio, Paul A Friedman, et al. Artificial intelligence-enabled ecg algorithm to identify patients with left ventricular systolic dysfunction presenting to the emergency department with dyspnea. *Circulation: Arrhythmia and Electrophysiology*, 13(8):e008437, 2020.

Joseph C Ahn, Zachi I Attia, Puru Rattan, Aidan F Mullan, Seth Buryska, Alina M Allen, Patrick S Kamath, Paul A Friedman, Vijay H Shah, Peter A Noseworthy, et al. Development of the ai-cirrhosis-ecg score: an electrocardiogram-based deep learning model in cirrhosis. *The American Journal of Gastroenterology*, 117 (3):424–432, 2022.

Jean-Baptiste Alayrac, Adrià Recasens, Rosalia Schneider, Relja Arandjelović, Jason Ramapuram, Jeffrey De Fauw, Lucas Smaira, Sander Dieleman, and Andrew Zisserman. Self-Supervised MultiModal Versatile Networks. In *NeurIPS*, 2020.

Erick A Perez Alday, Annie Gu, Amit J Shah, Chad Robichaux, An-Kwok Ian Wong, Chengyu Liu, Feifei Liu, Ali Bahrami Rad, Andoni Elola, Salman Seyedi, et al. Classification of 12-lead ecgs: the physionet/computing in cardiology challenge 2020. *Physiological measurement*, 41(12):124003, 2020.

Zachi I Attia, Suraj Kapa, Francisco Lopez-Jimenez, Paul M McKie, Dorothy J Ladewig, Gaurav Satam, Patricia A Pellikka, Maurice Enriquez-Sarano, Peter A Noseworthy, Thomas M Munger, et al. Screening for cardiac contractile dysfunction using an artificial intelligence–enabled electrocardiogram. *Nature medicine*, 25(1):70–74, 2019a.

Zachi I Attia, Suraj Kapa, Xiaoxi Yao, Francisco Lopez-Jimenez, Tarun L Mohan, Patricia A Pellikka, Rickey E Carter, Nilay D Shah, Paul A Friedman, and Peter A Noseworthy. Prospective validation of a deep learning electrocardiogram algorithm for the detection of left ventricular systolic dysfunction. *Journal of cardiovascular electrophysiology*, 30(5):668–674, 2019b.

Zachi I Attia, Peter A Noseworthy, Francisco Lopez-Jimenez, Samuel J Asirvatham, Abhishek J Deshmukh, Bernard J Gersh, Rickey E Carter, Xiaoxi Yao, Alejandro A Rabinstein, Brad J Erickson, 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):861–867, 2019c.

Shekoofeh Azizi, Basil Mustafa, Fiona Ryan, Zachary Beaver, Jan Freyberg, Jonathan Deaton, Aaron Loh, Alan Karthikesalingam, Simon Kornblith, Ting Chen, et al. Big self-supervised models advance medical image classification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 3478–3488, 2021.

Shekoofeh Azizi, Laura Culp, Jan Freyberg, Basil Mustafa, Sebastien Baur, Simon Kornblith, Ting Chen, Patricia MacWilliams, S Sara Mahdavi, Ellery Wulczyn, et al. Robust and efficient medical imaging with self-supervision. *arXiv preprint arXiv:2205.09723*, 2022.

Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in neural information processing systems*, 33: 12449–12460, 2020.

Shruthi Bannur, Stephanie Hyland, Flora Liu, Fernando Pérez-García, Maximilian Ilse, Daniel Coelho de Castro, Benedikt Boecking, Harshita Sharma, Kenza Bouzid, Anja Thieme, Anton Schwaighofer, Maria Teodora Wetscherek, Matthew Lungren, Aditya Nori, Javier Alvarez-Valle, and Ozan Oktay. Learning to exploit temporal structure for biomedical vision-language processing. In *The IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2023.

Benedikt Boecking, Naoto Usuyama, Shruthi Bannur, Daniel Coelho de Castro, Anton Schwaighofer, Stephanie Hyland, Maria Teodora Wetscherek, Tristan Naumann, Aditya Nori, Javier Alvarez-Valle, Hoifung Poon, and Ozan Oktay. Making the most of text semantics to improve biomedical vision-language processing. In *The European Conference on Computer Vision (ECCV)*, October 2022.

J Martijn Bos, Zachi I Attia, David E Albert, Peter A Noseworthy, Paul A Friedman, and Michael J Ackerman. Use of artificial intelligence and deep neural networks in evaluation of patients with electrocardiographically concealed long qt syndrome from the surface 12-lead electrocardiogram. *JAMA cardiology*, 6(5):532–538, 2021.

Mathilde Caron, Ishan Misra, Julien Mairal, Priya Goyal, Piotr Bojanowski, and Armand Joulin. Unsupervised learning of visual features by contrasting cluster assignments. *Advances in neural information processing systems*, 33:9912–9924, 2020.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pp. 1597–1607. PMLR, 2020a.

Ting Chen, Simon Kornblith, Kevin Swersky, Mohammad Norouzi, and Geoffrey E Hinton. Big self-supervised models are strong semi-supervised learners. *Advances in neural information processing systems*, 33:22243–22255, 2020b.

Joseph Y Cheng, Hanlin Goh, Kaan Dogrusoz, Oncel Tuzel, and Erdrin Azemi. Subject-aware contrastive learning for biosignals. *arXiv preprint arXiv:2007.04871*, 2020.

Georgios Christopoulos, Jonathan Graff-Radford, Camden L Lopez, Xiaoxi Yao, Zachi I Attia, Alejandro A Rabinstein, Ronald C Petersen, David S Knopman, Michelle M Mielke, Walter Kremers, et al. Artificial intelligence–electrocardiography to predict incident atrial fibrillation: A population-based study. *Circulation: Arrhythmia and Electrophysiology*, 13(12):e009355, 2020.

Ozan Ciga, Tony Xu, and Anne Louise Martel. Self supervised contrastive learning for digital histopathology. *Machine Learning with Applications*, 7:100198, 2022.

Michal Cohen-Shelly, Zachi I Attia, Paul A Friedman, Saki Ito, Benjamin A Essayagh, Wei-Yin Ko, Dennis H Murphree, Hector I Michelena, Maurice Enriquez-Sarano, Rickey E Carter, et al. Electrocardiogram screening for aortic valve stenosis using artificial intelligence. *European heart journal*, 42(30):2885–2896, 2021.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

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):e1009862, 2022.

Conner D Galloway, Alexander V Valys, Jacqueline B Shreibati, Daniel L Treiman, Frank L Petterson, Vivek P Gundotra, David E Albert, Zachi I Attia, Rickey E Carter, Samuel J Asirvatham, et al. Development and validation of a deep-learning model to screen for hyperkalemia from the electrocardiogram. *JAMA cardiology*, 4(5):428–436, 2019.

Shashank Goel, Hritik Bansal, Sumit Bhatia, Ryan A. Rossi, Vishwa Vinay, and Aditya Grover. CyCLIP: Cyclic contrastive language-image pretraining. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho (eds.), *Advances in Neural Information Processing Systems*, 2022.

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 *Machine Learning for Health*, pp. 156–167. PMLR, 2021.

Priya Goyal, Mathilde Caron, Benjamin Lefaudeux, Min Xu, Pengchao Wang, Vivek Pai, Mannat Singh, Vitaliy Liptchinsky, Ishan Misra, Armand Joulin, et al. Self-supervised pretraining of visual features in the wild. *arXiv preprint arXiv:2103.01988*, 2021.

Martha Grogan, Francisco Lopez-Jimenez, Michal Cohen-Shelly, Angela Dispenzieri, Zachi I Attia, Omar F Abou Ezzedine, Grace Lin, Suraj Kapa, Daniel D Borgeson, Paul A Friedman, et al. Artificial intelligence–enhanced electrocardiogram for the early detection of cardiac amyloidosis. In *Mayo Clinic Proceedings*, volume 96, pp. 2768–2778. Elsevier, 2021.

Awni Y Hannun, Pranav Rajpurkar, Masoumeh Haghpanahi, Geoffrey H Tison, Codie Bourn, Mintu P Turakhia, and Andrew Y Ng. Cardiologist-level arrhythmia detection and classification in ambulatory electrocardiograms using a deep neural network. *Nature medicine*, 25(1):65–69, 2019.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 9729–9738, 2020.

Shih-Cheng Huang, Liyue Shen, Matthew P Lungren, and Serena Yeung. Gloria: A multimodal global-local representation learning framework for label-efficient medical image recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 3942–3951, 2021.

Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *International Conference on Machine Learning*, pp. 4904–4916. PMLR, 2021.

Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. Supervised contrastive learning. *Advances in neural information processing systems*, 33:18661–18673, 2020.

Dani Kiyasseh, Tingting Zhu, and David A Clifton. Clocs: Contrastive learning of cardiac signals across space, time, and patients. In *International Conference on Machine Learning*, pp. 5606–5615. PMLR, 2021.

Wei-Yin Ko, Konstantinos C Siontis, Zachi I Attia, Rickey E Carter, Suraj Kapa, Steve R Ommen, Steven J Demuth, Michael J Ackerman, Bernard J Gersh, Adelaide M Arruda-Olson, et al. Detection of hypertrophic cardiomyopathy using a convolutional neural network-enabled electrocardiogram. *Journal of the American College of Cardiology*, 75(7):722–733, 2020.

Junnan Li, Pan Zhou, Caiming Xiong, and Steven Hoi. Prototypical contrastive learning of unsupervised representations. In *International Conference on Learning Representations*, 2021.

Yikuan Li, Shishir Rao, José Roberto Ayala Solares, Abdelaali Hassaine, Rema Ramakrishnan, Dexter Canoy, Yajie Zhu, Kazem Rahimi, and Gholamreza Salimi-Khorshidi. Behrt: transformer for electronic health records. *Scientific reports*, 10(1):7155, 2020.

Baijiong Lin, Feiyang YE, Yu Zhang, and Ivor Tsang. Reasonable effectiveness of random weighting: A litmus test for multi-task learning. *Transactions on Machine Learning Research*, 2022. ISSN 2835-8856.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019.

Ming Y. Lu, Bowen Chen, Andrew Zhang, Drew F. K. Williamson, Richard J. Chen, Tong Ding, Long Phi Le, Yung-Sung Chuang, and Faisal Mahmood. Visual language pretrained multiple instance zero-shot transfer for histopathology images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 19764–19775, June 2023.

Temesgen Mehari and Nils Strodthoff. Self-supervised representation learning from 12-lead ecg data. *Computers in biology and medicine*, 141:105114, 2022.

Ishan Misra and Laurens van der Maaten. Self-supervised learning of pretext-invariant representations. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 6707–6717, 2020.

Jungwoo Oh, Hyunseung Chung, Joon-myoung Kwon, Dong-gyun Hong, and Edward Choi. Lead-agnostic self-supervised learning for local and global representations of electrocardiogram. In *Conference on Health, Inference, and Learning*, pp. 338–353. PMLR, 2022.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.

Chao Pang, Xinzhuo Jiang, Krishna S Kalluri, Matthew Spotnitz, RuiJun Chen, Adler Perotte, and Karthik Natarajan. Cehr-bert: Incorporating temporal information from structured ehr data to improve prediction tasks. In *Machine Learning for Health*, pp. 239–260. PMLR, 2021.

Raphael Poulain, Mehak Gupta, and Rahmatollah Beheshti. Few-shot learning with semi-supervised transformers for electronic health records. In Zachary Lipton, Rajesh Ranganath, Mark Sendak, Michael Sjoding, and Serena Yeung (eds.), *Proceedings of the 7th Machine Learning for Healthcare Conference*, volume 182 of *Proceedings of Machine Learning Research*, pp. 853–873. PMLR, 05–06 Aug 2022.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PMLR, 2021.

Aniruddh Raghu, Payal Chandak, Ridwan Alam, John Guttag, and Collin Stultz. Contrastive pre-training for multimodal medical time series. In *NeurIPS 2022 Workshop on Learning from Time Series for Health*, 2022.

Laila Rasmy, Yang Xiang, Ziqian Xie, Cui Tao, and Degui Zhi. Med-bert: pretrained contextualized embeddings on large-scale structured electronic health records for disease prediction. *NPJ digital medicine*, 4(1):86, 2021.

Jie Ren, Stanislav Fort, Jeremiah Liu, Abhijit Guha Roy, Shreyas Padhy, and Balaji Lakshminarayanan. A simple fix to mahalanobis distance for improving near-ood detection. *arXiv preprint arXiv:2106.09022*, 2021.

Konstantinos C Siontis, Peter A Noseworthy, Zachi I Attia, and Paul A Friedman. Artificial intelligence-enhanced electrocardiography in cardiovascular disease management. *Nature Reviews Cardiology*, 18(7): 465–478, 2021.

Hari Sowrirajan, Jingbo Yang, Andrew Y Ng, and Pranav Rajpurkar. Moco pretraining improves representation and transferability of chest x-ray models. In *Medical Imaging with Deep Learning*, pp. 728–744. PMLR, 2021.

Chetan L Srinidhi and Anne L Martel. Improving self-supervised learning with hardness-aware dynamic curriculum learning: an application to digital pathology. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 562–571, 2021.

Vikto Tihonenko, A Khaustov, S Ivanov, A Rivin, and E Yakushenko. St petersburg incart 12-lead arrhythmia database. *PhysioBank PhysioToolkit and PhysioNet*, 2008.

Geoffrey H Tison, Jeffrey Zhang, Francesca N Delling, and Rahul C Deo. Automated and interpretable patient ecg profiles for disease detection, tracking, and discovery. *Circulation: Cardiovascular Quality and Outcomes*, 12(9):e005289, 2019.

Rachael A. Venn, Xin Wang, Sam Freesun Friedman, Nate Diamant, Shaan Khurshid, Paolo Di Achille, Lu-Chen Weng, Seung Hoan Choi, Christopher Reeder, James P. Pirruccello, Pulkit Singh, Emily S. Lau, Anthony Philippakis, Christopher D. Anderson, Patrick T. Ellinor, Jennifer E. Ho, Puneet Batra, and Steven A. Lubitz. Deep learning of electrocardiograms enables scalable human disease profiling. *medRxiv*, 2022. doi: 10.1101/2022.12.21.22283757.

T. Wagner, Z.I. Attia, S.J. Asirvatham, S. Awasthi, M. Babu, R. Barve, K. Carlson, C.L. Carpenter, R.P. Frantz, P.A. Friedman, A. Prasad, C. Chehoud, E. Kogan, A. Nnewihe, D. Quinn, C. Bridges, S. Kapa, and V. Soundararajan. *An Automated Screening Algorithm Using Electrocardiograms for Pulmonary Hypertension*, pp. A1179–A1179. doi: 10.1164/ajrccm-conference.2021.203.1_MeetingAbstracts.A1179.

Xiyue Wang, Sen Yang, Jun Zhang, Minghui Wang, Jing Zhang, Wei Yang, Junzhou Huang, and Xiao Han. Transformer-based unsupervised contrastive learning for histopathological image classification. *Medical Image Analysis*, 2022.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pp. 38–45, Online, October 2020. Association for Computational Linguistics.

Xi Yang, Nima PourNejatian, Hoo Chang Shin, Kaleb E Smith, Christopher Parisien, Colin Compas, Cheryl Martin, Mona G Flores, Ying Zhang, Tanja Magoc, et al. Gatortron: A large clinical language model to unlock patient information from unstructured electronic health records. *arXiv preprint arXiv:2203.03540*, 2022.

Xiaoxi Yao, David R Rushlow, Jonathan W Inselman, Rozalina G McCoy, Thomas D Thacher, Emma M Behnken, Matthew E Bernard, Steven L Rosas, Abdulla Akfaly, Artika Misra, et al. Artificial intelligence–enabled electrocardiograms for identification of patients with low ejection fraction: a pragmatic, randomized clinical trial. *Nature Medicine*, 27(5):815–819, 2021.

Yuhao Zhang, Hang Jiang, Yasuhide Miura, Christopher D Manning, and Curtis P Langlotz. Contrastive learning of medical visual representations from paired images and text. In *Machine Learning for Healthcare Conference*, pp. 2–25. PMLR, 2022.

Jianwei Zheng, Jianming Zhang, Sidy Danioko, Hai Yao, Hangyuan Guo, and Cyril Rakovski. A 12-lead electrocardiogram database for arrhythmia research covering more than 10,000 patients. *Scientific data*, 7(1):48, 2020.

Appendix

A Training Details

We execute all pre-training and classification tasks using 2 Nvidia V100 (16G) GPUs. However, for the pretraining tasks involving the text domain, we utilize 2 Nvidia A100 (40G) GPUs. All the original ECGs consist of 12 leads and are 10 seconds long with a sampling rate of 500Hz. During training, we use a random crop of 5 seconds in length, i.e., 2500 samples. We optimize all pre-training and classification models with AdamW optimizer (Loshchilov & Hutter, 2019) with (β_1, β_2) set to (0.9, 0.999).

Pre-training Details We initially pre-train sEHR-BERT as described in Section 3.1. For multi-modal contrastive pre-training, we initialize the sEHR encoder with sEHR-BERT model weights and the text encoder with GatorTron-base (Yang et al., 2022) model weights. GatorTron-base (Yang et al., 2022) is a 345M-parameter language model pre-trained on large amounts of de-identified clinical notes (80B words) from the University of Florida Health System, having a vocabulary of size 50176. We use BERT and Megatron-BERT implementation offered by the Huggingface transformers library (Wolf et al., 2020) for sEHR and text encoders respectively. For the ECG encoder, we use ResNet-like architecture (He et al., 2016) customized to 1 dimension which consists of around 1M parameters. The ECG encoder is initialized randomly. More details about the architecture of the ECG encoder used are provided in Appendix C. In joint pre-training, we freeze the first 3 and 18 layers of sEHR and text encoders respectively, and fine-tune the remaining layers. Following Zhang et al. (2022), we set λ_{es}, λ_{et}, and τ to 0.5, 0.5, and 0.1 respectively. We used a batch size of 256 and an initial learning rate of 1e-4. The learning rate is reduced by a factor of 2 if the validation loss stops decreasing continuously for 2 epochs and we early stop the training based on validation loss with an early stopping patience of 10 epochs.

Classification Details For classification tasks, we add a two-layered MLP head on top of the ECG encoder. We also add dropout layers after each hidden layer with a dropout probability of 0.2 for regularisation. A batchsize of 128 is used for all classification models. We used an initial learning rate of 1e-3 for random initialization training for all diseases. For fine-tuning, we used an initial learning rate of 1e-3 for coronary atherosclerosis and myocarditis tasks, and 1e-4 for cardiac amyloidosis, pulmonary hypertension, low LVEF, and AFib in NSR tasks. The learning rate is reduced by a factor of 2 if the validation score stops increasing continuously for 2 epochs and we early stop the training based on validation loss with an early stopping patience of 10 epochs. During fine-tuning, we initialize the ECG encoder with the pre-trained ECG encoder weights and warm up the classification head (MLP) for 1024 steps by freezing the backbone network weights and then fine-tuning the entire network. During prediction, we take 6 consecutive 5-second long crops with a stride of 1 second from the original 10-second long ECG. The median of the predictions of these 6 crops is taken as the final prediction for computing the AUROC score.

B Global Splits and Pre-training Dataset Splits Details

Table 8: Number of ECGs and number of patients in global splits and pre-training dataset splits.

	Train		Validation		Test		Total	
Model	#ECGs	#Patients	#ECGs	#Patients	#ECGs	#Patients	#ECGs	#Patients
Global Splits	5,479,435	1,463,009	450,775	121,932	3,210,110	853,477	9,140,320	2,438,418
sEHR-BERT	-	1,167,991	-	97,333	-	-	-	-
ECG-sEHR	4,553,278	1,196,478	373,649	99,608	-	-	-	-
ECG-Text	5,416,467	1,423,999	445,342	118,628	-	-	-	-
sEHR-ECG-Text	4,526,686	1,177,903	371,367	98,013	-	-	-	-
ECG-MTL	5,462,181	1,455,805	449,432	121,387	3,200,116	849,414	9,111,729	2,426,606

C ECG Encoder Details

We use ResNet-like architecture (He et al., 2016) customized to 1 dimension for time series ECG signals. This consisted of eight 1D convolution layers based on the basic block of ResNet. Details of each layer are given in Table 9. All convolutional layers employ Batch normalization and ReLU activation function and a stride of 2. We add two fully connected layers with hidden sizes 128 and 64 on top of the backbone CNN architecture for classification tasks. We use the same architecture for all pre-training methods.

Table 9: ECG encoder details. IC: Input Channels, OC: Output Channels, K: Kernel size.

Layer	Layer type	IC	OC	K
1	Conv	12	32	5
2	Conv	32	32	5
3	Conv	32	64	5
4	Conv	64	64	3
5	Conv	64	128	3
6	Conv	128	128	3
7	Conv	128	256	3
8	Conv	256	256	3

D External Classification Dataset Details

The details of the PhysioNet2020 (Alday et al., 2020) and Chapman (Zheng et al., 2020) datasets are summarized in Table 10.

Table 10: Details of the PhysioNet2020 and Chapman datasets.

Dataset	#ECGs	#Patients	Signal length	Sampling rate
PhysioNet2020				
CPSC2018	6,877	6,877	6-60 secs	500 Hz
CPSC extra	3,453	3,453	6-60 secs	500 Hz
St Petersburg INCART	74	32	30 mins	257 Hz
PTB	516	516	-	1000Hz
PTB-XL	21,837	21,837	10 secs	500 Hz
Georgia	10,344	10,344	10 secs	500 Hz
Chapman				
Chapman	10,646	10,646	10 secs	500 Hz

E Multi-Task Learning Model Details

The performance of the ECG-MTL model on classification tasks can be found in Table 11, while Table 12 presents the model performance on regression tasks.

Table 11: The performance of classification tasks (AUC) and the source of labels for the ECG-MTL model.

Condition	Source	AUC	Condition	Source	AUC
Mitral Stenosis	ECHO	0.9664	Inferior Injury	ECG	0.9826
Aortic Stenosis	ECHO	0.9186	Junctional Escape Beats	ECG	0.9823
Mitral Regurgitation	ECHO	0.9129	Right Axis Deviation	ECG	0.9806
Tricuspid Regurgitation	ECHO	0.9099	Incomplete LBBB	ECG	0.9805
Aortic Regurgitation	ECHO	0.8551	Ectopic Atrial Rhythm	ECG	0.9795
4:1 AV Block	ECG	0.9992	Short PR Interval	ECG	0.9789
Complete RBBB	ECG	0.9989	Premature Atrial Complexes	ECG	0.9787
Atrial Fibrillation	ECG	0.9987	Ventricular Escape Beats	ECG	0.9781
Complete LBBB	ECG	0.9984	Anterior Injury	ECG	0.9769
Sinus Tachycardia	ECG	0.9984	Left Ventricular Hypertrophy	ECG	0.9745
Sinus Bradycardia	ECG	0.9979	Wandering Atrial Pacemaker	ECG	0.9739
Bifascicular Block	ECG	0.9970	Second degree AV Block Type I	ECG	0.9738
Junctional Tachycardia	ECG	0.9966	Left Axis Deviation	ECG	0.9728
Ventricular Pacemaker	ECG	0.9959	Acute Pericarditis	ECG	0.9727
Dualchamber Pacemaker	ECG	0.9958	Gender	ECG	0.9724
Variable Av Block	ECG	0.9956	Wolff Parkinson White	ECG	0.9716
Junctional Bradycardia	ECG	0.9956	Anteroseptal Infarct	ECG	0.9716
Biventricular Hypertrophy	ECG	0.9954	Incomplete RBBB	ECG	0.9696
Third Degree AV Block	ECG	0.9950	Low QRS Voltage	ECG	0.9692
Multifocal Atrial Tachycardia	ECG	0.9950	Second degree AV Block Type II	ECG	0.9692
Atrial Flutter	ECG	0.9948	Dextrocardia	ECG	0.9691
Supraventricular Tachycardia	ECG	0.9946	Lateral Infarct	ECG	0.9672
Ventricular Tachycardia	ECG	0.9941	Early Repolarization	ECG	0.9652
Anterolateral Injury	ECG	0.9935	3:1 AV Block	ECG	0.9649
Ectopic Atrial Tachycardia	ECG	0.9934	Left Atrial Enlargement	ECG	0.9647
Right Superior Axis Deviation	ECG	0.9933	Nonspecific IVCD	ECG	0.9634
Inferolateral Injury	ECG	0.9929	Prolonged QT	ECG	0.9602
Biatrial Enlargement	ECG	0.9926	Inferior Infarct	ECG	0.9584
Premature Ventricular Complexes	ECG	0.9921	Posterior Infarct	ECG	0.9520
Ectopic Atrial Bradycardia	ECG	0.9920	Sinus Arrhythmia	ECG	0.9505
Left Posterior Fascicular Block	ECG	0.9911	Septal Infarct	ECG	0.9434
Trifascicular Block	ECG	0.9904	Anterior Infarct	ECG	0.9423
Idioventricular Rhythm	ECG	0.9904	Premature Junctional Complexes	ECG	0.9420
Lateral Injury	ECG	0.9898	ST Depression	ECG	0.9400
QRS Widening	ECG	0.9895	ST Elevation	ECG	0.9390
Left Anterior Fascicular Block	ECG	0.9892	Non-Specific T Wave Changes	ECG	0.9316
First Degree AV Block	ECG	0.9888	T Wave Inversion	ECG	0.9251
Right Ventricular Hypertrophy	ECG	0.9886	Normal ECG	ECG	0.9248
Right Atrial Enlargement	ECG	0.9865	ST And T Wave Abnormality	ECG	0.9158
Junctional Rhythm	ECG	0.9857	Poor R Wave Progression	ECG	0.9034
Anterolateral Infarct	ECG	0.9850	ST Segment Abnormality	ECG	0.8585
Normal Sinus Rhythm	ECG	0.9848			

Table 12: The performance of regression tasks and the source of labels for the ECG-MTL model.

Measurement	Source	Pearson correlation	Spearman correlation	R^2 score
Age	ECG	0.8699	0.8273	0.7544
Wall Motion Score Index	ECHO	0.7513	0.7706	0.5627
Left Ventricular EF	ECHO	0.7145	0.5812	0.5090
Left Ventricular Mass	ECHO	0.7072	0.7017	0.5001
Left Atrial Volume	ECHO	0.6965	0.7068	0.4839
Left Ventricle End Diastolic Diameter	ECHO	0.6894	0.6386	0.4749
Mitral Valve E to E' Ratio	ECHO	0.6154	0.6294	0.3777
Interventricular Wall Thickness	ECHO	0.6040	0.5486	0.3633
Tricuspid Regurgitation Velocity	ECHO	0.5844	0.5458	0.3395
Tricuspid Annular Plane Systolic Excursion	ECHO	0.5652	0.5788	0.3183
Left Ventricular Posterior Wall Diastolic Thickness	ECHO	0.5623	0.5163	0.3155
Aortic Valve Area	ECHO	0.5139	0.5003	0.2625
Aortic Valve Systolic Peak Velocity	ECHO	0.5048	0.4731	0.2510
Aortic Valve Systolic Mean Gradient	ECHO	0.5007	0.5331	0.2459
MC CV Dimensionless Index	ECHO	0.4979	0.4841	0.2441
Hemoglobin	Lab Test	0.6550	0.6402	0.4279
Natriuretic Peptide B Prohormone N Terminal	Lab Test	0.6037	0.7756	0.3635
Potassium in Serum or Plasma	Lab Test	0.5662	0.5170	0.3170
Cholesterol in HDL in Serum or Plasma	Lab Test	0.5412	0.5433	0.2914
Creatinine	Lab Test	0.5359	0.4949	0.2830
Urea Nitrogen	Lab Test	0.5201	0.5034	0.2701
Urate	Lab Test	0.5028	0.5088	0.2517
C Reactive Protein, in Serum Plasma Blood	Lab Test	0.4837	0.5816	0.2331
Troponin I, Cardiac	Lab Test	0.4713	0.5296	0.2210
Hemoglobin A1C, Blood	Lab Test	0.4593	0.4627	0.2102
Lymphocytes 100 WBCs	Lab Test	0.4430	0.5127	0.1956
Glucose in Serum, Fasting	Lab Test	0.4403	0.4614	0.1918
Total Cholesterol [Massvolume] in Serum or Plasma	Lab Test	0.4356	0.4453	0.1876
Creatinine Kinase (MB)	Lab Test	0.4313	0.3138	0.1680
Total Bilirubin	Lab Test	0.3949	0.3274	0.1484
Sodium in Serum, Plasma or Blood	Lab Test	0.3854	0.3355	0.1470
LDL Cholesterol in Serum or Plasma	Lab Test	0.3808	0.393	0.1434
Cholesterol, VLDL, Serum Plasma	Lab Test	0.3763	0.3935	0.1397
Triglyceride in Serum or Plasma	Lab Test	0.3545	0.4494	0.1250
D-dimer	Lab Test	0.2902	0.3133	0.0828
Lactate Dehydrogenase (LDH), Serum	Lab Test	0.2865	0.4176	0.0785
Estimated Glomerular Filtration Rate (eGFR)	Lab Test	0.2843	0.2382	0.0729