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. 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 following facts: (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); (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. In particular, we utilize diagnosis codes, procedure codes, and medication prescriptions 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 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 (SOTA) 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 so that codes falling within non-overlapping 7-day windows share a common 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 this section, we describe the pre-training of the sEHR-ECG-Text model, where we pair the ECG modality with sEHR and text modalities to jointly learn multi-modal representations.

3.2.1 Preliminaries

MultiModal Versatile Networks (MMV) (Alayrac et al., 2020) apply contrastive learning to multi-modal data including video, audio, and text under the assumption that the video and audio modalities are more granular than the text modality. MMV discusses that applying contrastive loss in *shared space* (where all modalities are embedded into a single shared vector space) may not maintain specificities, as it implicitly assumes that all modalities have equal granularity. To address this, MMV proposes to learn two separate embedding spaces: a fine-grained space where video and audio are matched and a coarse-grained space where text is matched with video and audio domains. This method is referred to as *fine and coarse spaces (FAC).* We hypothesize that sEHR, ECG, and text modalities do not exhibit equal granularity. Moreover, the ECGs are paired with sEHR and text data within a specific 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 adopt the FAC framework from MMV in our sEHR-ECG-Text model. We describe the methodology in detail in the following sections.

3.2.2 Notation

Let $x \in \mathcal{X}$ be an instance defined by an instantiation of different modalities $\mathcal{M} : x = \{x_m\}$, $m \in \mathcal{M}$. In this study, we employ three modalities: ECG $x_e \in \mathcal{X}_e$, sEHR $x_s \in \mathcal{X}_s$, and text $x_t \in \mathcal{X}_t$. Specifically, x_s, x_e, and x_t represent the sequence of sEHR codes, ECG waveform samples, and sequence of text tokens respectively. Let $E_m : \mathcal{X}_m \rightarrow \mathbb{R}^{d_m}$ be a parameterized modality-specific encoder that takes as input an instance x_m from modality m and produces a modality-specific representation of dimension d_m. These modality-specific representations are embedded into a shared space $\Omega_z \subset \mathbb{R}^{d_z}$, where z represents the list of modalities that we embed into this space. For instance, $z = es$ to denote joint ECG-sEHR space Ω_{es}, $z = et$ to denote joint ECG-Text space Ω_{et}, or $z = set$ to denote joint sEHR-ECG-Text space Ω_{set}. In this shared space, we maximize or minimize the alignment between different modalities using the contrastive loss objective. The projection head $P_{m \rightarrow z} : \mathbb{R}^{d_m} \rightarrow \mathbb{R}^{d_z}$ is used to embed modality specific-representations $v_m = E_m(x_m)$ into the shared space Ω_z, and we denote the resulting vector as $v_{m,z} = P_{m \rightarrow z}(E_m(x_m))$, which signifies the representation of the input modality x_m in the shared space Ω_z.

3.2.3 Data Encoding

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 heads.

3.2.4 Multi-Modal Contrastive Objective

As mentioned before, we use FAC framework inspired from MMV (Alayrac et al., 2020) where ECG and sEHR are compared in fine-grained joint ECG-sEHR space Ω_{es}, while ECG is compared with text in coarse-grained joint sEHR-ECG-Text space Ω_{set}. Given a minibatch containing N instances $\{x^i\}_{i=1}^N$, we denote $v^i_{m,z} = P_{m \rightarrow z}(E_m(x^i_m))$ as the representation of the modality m in the shared space Ω_z for the i-th instance. Following Zhang et al. (2022), we define the contrastive objective bidirectionally. For example, in the case

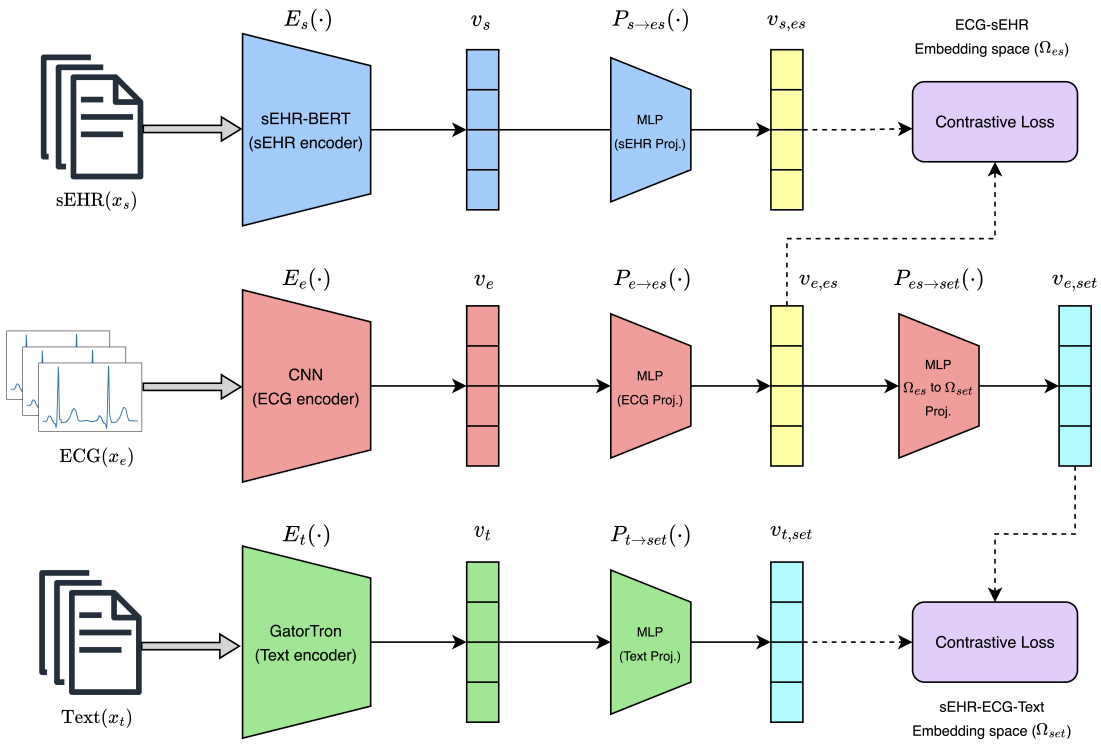


Figure 1: Overview of the sEHR-ECG-Text model pre-training. The model takes as input three modalities $\mathcal{M} : x = \{x_m\}$, $m \in \mathcal{M}$, i.e., ECG (x_e), sEHR (x_s), and text (x_t). $E_m(\cdot)$ and v_m denote the modality-specific encoder and modality-specific representation respectively. Ω_z denotes the shared embedding space, where z represents the list of modalities that we embed into this space. For instance, $z = es$ to denote joint ECG-sEHR space Ω_{es}. $P_{m\to z}(\cdot)$ denotes the projection head used to embed modality specific representation v_m into shared space Ω_z. $v_{m,z}$ denotes the representation of the input modality x_m in the shared space Ω_z. The model is trained by applying contrastive loss between ECG and sEHR in the fine-grained ECG-sEHR space (Ω_{es}) and between ECG and text in the coarse-grained sEHR-ECG-Text space (Ω_{set}).

of contrastive loss between ECG and sEHR domains, the loss is directed from ECG to sEHR and vice-versa. In the context of contrastive objective, we consider N pairs of ECG-sEHR (x_e, x_s) as positive, while the remaining $N^2 - N$ pairs are treated as negative. The same approach is applied to contrastive loss between ECG and text domains. Let $sim(x, y) = x^T y/\|x\|\|y\|$ denote the cosine similarity between two vectors $x, y \in \mathbb{R}^{d_z}$, $\mathcal{L}_{es}$ be the contrastive loss between ECG and sEHR, $\mathcal{L}_{et}$ be the contrastive loss between ECG and text, λ_{es} and λ_{et} be scalar weights $\in [0, 1]$, and $\tau \in \mathbb{R}^+$ be the temperature parameter. The combination of λ_{es} and λ_{et} gives the overall loss, denoted by $\mathcal{L}$ (Equation 3), and we aim to minimize this loss.

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

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

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

Figure 1 illustrates the pre-training of the sEHR-ECG-Text model. See Appendix B.1 for more details about implementation and training. We also present *shared space* versus *FAC spaces* ablation in Appendix E. We provide the architecture of the shared space sEHR-ECG-Text model in Appendix B.2.

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. We provide the ECG-sEHR model architecture in Appendix B.2.

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}). We provide the ECG-Text model architecture in Appendix B.2. Following the notation introduced in Section 3.2, we minimize the contrastive loss given by,

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

In this model, note that ECG and text are compared in joint ECG-Text space (Ω_{et}), which differs from the sEHR-ECG-Text model where ECG and text are compared in the sEHR-ECG-Text space (Ω_{set}).

4 Experiments and Results

4.1 Dataset Splits Setup

We used 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. These EHRs 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. 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. This approach ensures that we can effectively evaluate the quality of representations on downstream tasks, as the data of validation and test patients is not seen during the pre-training phase. Consequently, all datasets across tasks 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 following datasets for pre-training the proposed models: (i) sEHR sequences for pre-training the sEHR-BERT model, (ii) ECG-sEHR (x_e, x_s) pairs for pre-training the ECG-sEHR model, and (iii) ECG-Text (x_e, x_t) pairs for pre-training the ECG-Text model. For sEHR-ECG-Text model pre-training, we construct 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. See Appendix A.1 for details on the number of instances and patients used in different pre-training models.

4.2.1 Dataset for sEHR-BERT Pre-training

We use 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 differ from ICD-10 codes, although their corresponding text descriptions are similar, we map ICD-9 to ICD-10 to maintain consistent 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 associated with at least 50 patients are included in the vocabulary, resulting in a size of 28593. We present short ICD-10 vs. full ICD-10 diagnosis codes ablation while keeping codes from other sources consistent in Appendix E. 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 derived from various sources, including ECG reports, ECHO reports, pathology reports, radiology reports, microbiology reports, clinical notes, and surgical notes, collectively referred to as "patient notes" in this work. Despite the GatorTron-base (Yang et al., 2022) model's capacity to handle sequences of up to 512 tokens, we were limited to work with a maximum of 400 tokens due to computing resource constraints. Given the abundance of text data within patient notes, we implemented a filtering process to extract only the most relevant information. Specifically, we selected patient notes that contained entities from a predefined list of biomedical entity types, such as diseases, symptoms, procedures, medications, biomarkers, and gene mutations, using an in-house NLP model. The next step involves pairing ECGs with the selected patient notes. We employed two distinct methods for this purpose: report concatenation and entity concatenation.

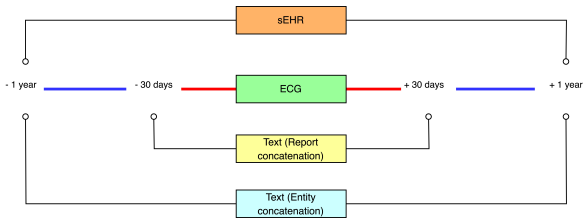


Figure 2: The figure shows the linkage of sEHR and text modalities with ECG modality within specific time windows: one year for the sEHR modality, one month for report concatenation, and one year for entity concatenation in the case of the text modality.

Report concatenation: For each patient's ECG, we concatenate all selected patient notes that are available within one month around the ECG acquisition timestamp. This approach resulted in an average sequence length of 354 tokens after tokenization.

Entity concatenation: It's important to note that aligning ECGs with patient information spanning a more extended timeframe is advantageous for understanding ECG patterns more effectively. To effectively capture a broader spectrum of medical insights while adhering to token constraints, we concatenate only the identified entities from the patient notes within a one-year timeframe around the ECG acquisition timestamp. This resulted in an average sequence length of 266 after tokenization. Figure 2 illustrates how sEHR and text modalities are linked with ECG modality within specific time windows around the ECG timestamp.

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 Appendix E.

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. See Appendix A.2 for the summary of the disease labeling process, the number of ECGs, and the number of patients used.

4.3.2 External Datasets

We also evaluate all pre-trained models on two publicly available datasets: (i) **PhysioNet2020** (Alday et al., 2020), which consists of a collection of six 12-lead ECG datasets with varying signal lengths and sampling rates, (ii) **Chapman** (Zheng et al., 2020), which contains 10-second long 12-lead ECGs (see Appendix A.3 for more details). The diseases in these datasets are commonly diagnosed by physicians directly using ECGs, unlike the diseases in our proprietary internal classification datasets. To replicate the state-of-the-art results on the PhysioNet2020 dataset presented by 3KG (Gopal et al., 2021), we followed their detailed procedure: (i) merge some of the conditions from the list of 27 conditions 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; (ii) resample the ECG signals to 500Hz; (iii) take non-overlapping 5-second long crops from each ECG record exhaustively and associate those with the label of the original record; (iv) split the dataset into 80%, 10%, and 10% for training, validation, and testing respectively; (v) train a 23-class multilabel classification model. For the Chapman dataset, in line with Kiyasseh et al. (2021), we merge 11 cardiac arrhythmia conditions of the dataset into 4 major classes and split the dataset into 60%, 20%, and 20% 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 that was used during 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.

Table 1: Linear classification results (AUROC, mean and standard deviation over 5 runs with different random seeds) for PhysioNet2020 (23 classes) and Chapman (4 classes) datasets. Results within 95% confidence intervals of the best result are shown in **bold**.

	PhysioNet2020			Chapman		
Method	1%	10%	100%	1%	10%	100%
Supervised baseline						
Random Init.	0.684 ± 0.024	0.842 ± 0.006	0.913 ± 0.006	0.552 ± 0.031	0.964 ± 0.003	0.987 ± 0.002
ECG-only SSL						
3KG	0.774 ± 0.001	0.837 ± 0.003	0.872 ± 0.000	0.934 ± 0.008	0.968 ± 0.001	0.984 ± 0.000
CLOCS(CMSC)	0.770 ± 0.005	0.836 ± 0.002	0.864 ± 0.000	0.876 ± 0.019	0.936 ± 0.001	0.946 ± 0.000
PCLR	0.721 ± 0.003	0.798 ± 0.003	0.834 ± 0.001	0.722 ± 0.007	0.821 ± 0.004	0.872 ± 0.000
Our models						
sEHR-ECG-Text	0.813 ± 0.005	0.882 ± 0.003	0.911 ± 0.000	0.957 ± 0.004	0.979 ± 0.001	0.985 ± 0.000
ECG-sEHR	0.788 ± 0.005	0.866 ± 0.002	0.896 ± 0.000	0.937 ± 0.010	0.969 ± 0.000	0.976 ± 0.000
ECG-Text	**0.820 ± 0.003**	**0.887 ± 0.001**	**0.915 ± 0.000**	**0.974 ± 0.002**	**0.987 ± 0.001**	**0.990 ± 0.000**

4.5.1 Classification Tasks

We evaluate the pre-trained models on linear classification and fine-tuning tasks. We evaluate the representation quality by extracting embeddings from the pre-trained ECG encoder by passing ECG waveform as input and training logistic regression models 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 ensures consistent performance with reduced training data. 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 this, we create different fractions (1%, 10%, 100%) of the training set by maintaining the original disease prevalence. For low-data diseases such as coronary atherosclerosis, myocarditis, and cardiac amyloidosis, we drop 1% split due to small dataset sizes. We use AUROC and AUPRC as our evaluation metrics. We conduct each experiment using five random seeds and report the mean and standard deviation. For the 1% and 10% splits, we conducted experiments with five distinct fractional splits derived from the original training split (100%).

Results and Discussion. Table 1 shows the linear classification results on external datasets. Table 2 shows classification results on coronary atherosclerosis, myocarditis, and cardiac amyloidosis diseases (low-data) and Table 3 shows classification results on pulmonary hypertension, low LVEF, and AFib in NSR diseases (high-data). We observed the following findings: (i) 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) Fine-tuned classification models initialized with our pre-trained models' weights outperform all baseline models by a large margin in low-data environments and a small margin in high-data environments. (iii) Classification models trained with only 10% of the training data using our pre-trained models produce results that are as good as or better than those obtained by training on the entire dataset with baseline models. This demonstrates the effectiveness of our pre-trained models in achieving data efficiency. (iv) When utilizing the complete training dataset, on the linear classification task, our ECG-Text model achieves an AUROC score of 0.915 on PhysioNet2020, surpassing the SOTA performance reported in Gopal et al. (2021) (3KG) by 2.5% (0.915 vs. 0.890). Additionally, on the Chapman dataset, it achieves an AUROC score of 0.990, surpassing the SOTA performance reported in Kiyasseh et al. (2021) (CLOCS) by 3.1% (0.990 vs. 0.959). It's worth noting that, on equal grounds, we surpass 3KG by 4.3% (0.915 vs. 0.872) and CLOCS by 4.4% (0.990 vs. 0.946). (v) We hypothesize that the main reason for the poor performance of ECG-only contrastive learning models on internal datasets is their exclusive dependence on ECGs for learning representations, i.e., comparing different instances of ECG data. This may not be sufficient to

learn the complex patterns of various medical conditions. In contrast, our methods align ECGs with EHR data, providing rich contextual information about a patient's health history, including diagnoses, procedures, medications, and more. This approach is beneficial for learning ECG patterns more effectively.

Comparison between ECG-sEHR and ECG-Text models. The ECG-sEHR model demonstrates superior performance on internal datasets, whereas the ECG-Text model excels on external datasets. We hypothesize that this difference arises from the fact that external datasets, such as PhysioNet2020 and Chapman, primarily comprise diseases that are commonly diagnosed from ECGs (i.e., arrhythmias, conduction blocks, etc.). These diagnoses are well-documented in the textual modality, particularly within ECG reports. As a result, the ECG-Text model performs best on these external datasets. In contrast, our internal datasets contain diseases for which labels are derived from EHR data. These labels are better captured by the sEHR modality. Therefore, the ECG-sEHR model outperforms the ECG-Text model when applied to internal datasets. It's worth noting that the sEHR-ECG-Text model demonstrates strong generalization across all internal datasets and achieves comparable performance with the ECG-Text model in the case of external datasets, i.e., PhysioNet2020 and Chapman, as it is trained with both sEHR and text modalities. Furthermore, it offers the advantage of comparing different modalities for retrieval tasks.

Table 2: Results (AUROC, mean and standard deviation over 5 runs with different random seeds) for coronary atherosclerosis, myocarditis, and cardiac amyloidosis classification tasks: (a) linear classification, (b) fine-tuned classification. Results within 95% confidence intervals of the best result are shown in **bold**. See Appendix D for AUPRC metrics.

(a) Linear classification

	Coronary atherosclerosis		Myocarditis		Cardiac amyloidosis	
Method	10%	100%	10%	100%	10%	100%
Supervised baseline						
Random Init.	0.772 ± 0.012	0.831 ± 0.004	0.775 ± 0.019	0.868 ± 0.003	0.912 ± 0.004	0.945 ± 0.001
ECG-only SSL						
3KG	0.722 ± 0.008	0.786 ± 0.000	0.699 ± 0.022	0.808 ± 0.000	0.869 ± 0.009	0.906 ± 0.000
CLOCS(CMSC)	0.743 ± 0.004	0.801 ± 0.000	0.662 ± 0.021	0.783 ± 0.000	0.876 ± 0.006	0.918 ± 0.000
PCLR	0.761 ± 0.004	0.825 ± 0.000	0.718 ± 0.022	0.818 ± 0.000	0.898 ± 0.006	0.928 ± 0.000
Our models						
sEHR-ECG-Text	**0.840 ± 0.003**	0.890 ± 0.000	**0.860 ± 0.019**	**0.901 ± 0.001**	**0.934 ± 0.007**	**0.960 ± 0.000**
ECG-sEHR	**0.836 ± 0.011**	**0.891 ± 0.000**	**0.859 ± 0.011**	0.896 ± 0.000	**0.932 ± 0.006**	0.959 ± 0.000
ECG-Text	0.821 ± 0.006	0.875 ± 0.000	0.785 ± 0.010	0.881 ± 0.000	0.922 ± 0.006	0.949 ± 0.000

(b) Fine-tuned classification

	Coronary atherosclerosis		Myocarditis		Cardiac amyloidosis	
Method	10%	100%	10%	100%	10%	100%
Supervised baseline						
Random Init.	0.772 ± 0.012	0.831 ± 0.004	0.775 ± 0.019	0.868 ± 0.003	0.912 ± 0.004	0.945 ± 0.001
ECG-only SSL						
3KG	0.751 ± 0.007	0.823 ± 0.007	0.725 ± 0.008	0.853 ± 0.015	0.900 ± 0.007	0.945 ± 0.002
CLOCS(CMSC)	0.782 ± 0.005	0.815 ± 0.009	0.716 ± 0.024	0.857 ± 0.013	0.912 ± 0.006	0.948 ± 0.001
PCLR	0.812 ± 0.004	0.828 ± 0.005	0.722 ± 0.032	0.853 ± 0.018	0.924 ± 0.003	0.951 ± 0.001
Our models						
sEHR-ECG-Text	**0.880 ± 0.003**	0.888 ± 0.005	**0.862 ± 0.005**	**0.905 ± 0.003**	**0.948 ± 0.003**	**0.961 ± 0.002**
ECG-sEHR	**0.881 ± 0.003**	**0.892 ± 0.001**	**0.866 ± 0.008**	**0.902 ± 0.003**	**0.949 ± 0.001**	**0.961 ± 0.002**
ECG-Text	0.871 ± 0.002	0.880 ± 0.001	**0.857 ± 0.008**	0.892 ± 0.004	0.940 ± 0.002	**0.959 ± 0.001**

Table 3: Results (AUROC, mean and standard deviation over 5 runs with different random seeds) for pulmonary hypertension, low LVEF, and AFib in NSR classification tasks: (a) linear classification, (b) fine-tuned classification. Results within 95% confidence intervals of the best result are shown in **bold**. See Appendix D for AUPRC metrics.

(a) Linear classification

	Pulmonary hypertension			Low LVEF			AFib in NSR		
Method	1%	10%	100%	1%	10%	100%	1%	10%	100%
Supervised baseline									
Random Init.	0.805 ± 0.008	0.887 ± 0.005	0.927 ± 0.002	0.857 ± 0.006	0.919 ± 0.002	0.944 ± 0.000	0.834 ± 0.013	0.896 ± 0.002	0.922 ± 0.001
ECG-only SSL									
3KG	0.779 ± 0.018	0.862 ± 0.002	0.874 ± 0.000	0.846 ± 0.016	0.902 ± 0.002	0.914 ± 0.000	0.824 ± 0.009	0.866 ± 0.000	0.870 ± 0.000
CLOCS(CMSC)	0.782 ± 0.021	0.857 ± 0.002	0.872 ± 0.000	0.864 ± 0.011	0.905 ± 0.002	0.916 ± 0.000	0.818 ± 0.009	0.862 ± 0.000	0.868 ± 0.000
PCLR	0.845 ± 0.016	0.894 ± 0.001	0.907 ± 0.000	0.894 ± 0.006	0.927 ± 0.000	0.936 ± 0.000	0.865 ± 0.007	0.898 ± 0.000	0.904 ± 0.000
Our models									
sEHR-ECG-Text	**0.900 ± 0.005**	**0.932 ± 0.001**	0.939 ± 0.000	**0.911 ± 0.013**	**0.941 ± 0.001**	**0.951 ± 0.000**	**0.902 ± 0.008**	0.928 ± 0.001	0.931 ± 0.000
ECG-sEHR	**0.899 ± 0.007**	**0.933 ± 0.001**	**0.940 ± 0.000**	**0.910 ± 0.014**	**0.942 ± 0.001**	**0.951 ± 0.000**	**0.902 ± 0.007**	**0.930 ± 0.001**	**0.933 ± 0.000**
ECG-Text	0.889 ± 0.006	0.924 ± 0.001	0.931 ± 0.000	**0.901 ± 0.014**	0.938 ± 0.001	0.947 ± 0.000	0.891 ± 0.008	0.923 ± 0.000	0.925 ± 0.000

(b) Fine-tuned classification

	Pulmonary hypertension			Low LVEF			AFib in NSR		
Method	1%	10%	100%	1%	10%	100%	1%	10%	100%
Supervised baseline									
Random Init.	0.805 ± 0.008	0.887 ± 0.005	0.927 ± 0.002	0.857 ± 0.006	0.919 ± 0.002	0.944 ± 0.000	0.834 ± 0.013	0.896 ± 0.002	0.922 ± 0.001
ECG-only SSL									
3KG	0.800 ± 0.010	0.891 ± 0.002	0.925 ± 0.001	0.868 ± 0.007	0.922 ± 0.002	0.945 ± 0.000	0.842 ± 0.009	0.900 ± 0.001	0.922 ± 0.001
CLOCS(CMSC)	0.833 ± 0.002	0.892 ± 0.003	0.927 ± 0.001	0.893 ± 0.002	0.926 ± 0.002	0.945 ± 0.001	0.849 ± 0.005	0.897 ± 0.002	0.920 ± 0.001
PCLR	0.865 ± 0.010	0.911 ± 0.002	0.933 ± 0.001	0.905 ± 0.002	**0.937 ± 0.002**	**0.949 ± 0.000**	0.874 ± 0.006	0.906 ± 0.002	0.922 ± 0.001
Our models									
sEHR-ECG-Text	**0.916 ± 0.004**	**0.934 ± 0.001**	0.938 ± 0.003	**0.933 ± 0.002**	**0.939 ± 0.003**	**0.949 ± 0.001**	**0.909 ± 0.008**	**0.919 ± 0.002**	**0.926 ± 0.001**
ECG-sEHR	**0.917 ± 0.005**	**0.935 ± 0.001**	**0.940 ± 0.001**	**0.932 ± 0.001**	**0.939 ± 0.002**	**0.949 ± 0.000**	**0.914 ± 0.005**	**0.920 ± 0.003**	**0.928 ± 0.001**
ECG-Text	0.907 ± 0.003	0.926 ± 0.001	0.934 ± 0.001	0.927 ± 0.004	**0.938 ± 0.001**	0.948 ± 0.001	0.901 ± 0.008	0.916 ± 0.003	**0.926 ± 0.002**

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, and (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.

Table 4: 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**

Zero-shot Text-ECG Retrieval. We take 1000 distinct ECG reports as search queries for each of the 41 cardiovascular conditions. For each of the conditions, we take 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 4 shows the zero-shot ECG-ECG retrieval and ECG-Text retrieval results. Our ECG-Text model outperforms the random guess method 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 5, which clearly shows that generic ECG representations are better at detecting out-of-distribution data compared to disease-specific representations.

Table 5: 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	0.757	0.620

5 Conclusion

Our work introduces a series of three multi-modal contrastive learning models. 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.

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.

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.

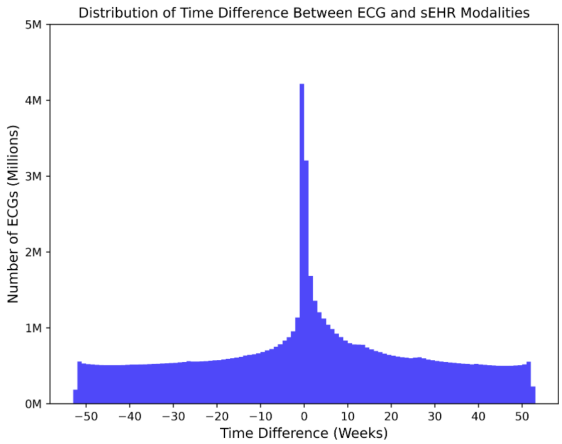
Appendix

A Dataset Details

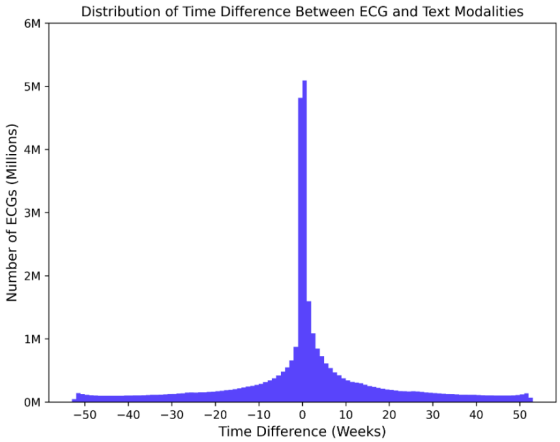
A.1 Pre-training Dataset Details

Table 6: The number of ECGs and the number of patients used during pre-training of the proposed models, i.e., sEHR-BERT, sEHR-ECG-Text, ECG-sEHR, and ECG-Text.

	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	-	-	-	-
sEHR-ECG-Text	4,526,686	1,177,903	371,367	98,013	-	-	-	-
ECG-sEHR	4,553,278	1,196,478	373,649	99,608	-	-	-	-
ECG-Text	5,416,467	1,423,999	445,342	118,628	-	-	-	-



(a) Time difference distribution between ECG and sEHR modalities.



(b) Time difference distribution between ECG and text modalities.

Figure 3: The figures (a) and (b) show the distribution of time difference between ECG and other modalities: (a) the distribution of time difference between ECG and sEHR modalities, (b) the distribution of time difference between ECG and text modalities.

Table 6 outlines the number of ECGs and the number of patients used during pre-training of the proposed models. Figure 3a shows the time difference distribution between ECG and sEHR modalities, while Figure 3b displays the time difference distribution between ECG and text modalities. The Y-axis in Figure 3a represents the number of ECGs associated with at least one medical code for the sEHR modality in one-week intervals and the Y-axis in Figure 3b represents the number of ECGs associated with at least one biomedical entity (mentioned in Section 4.2.3) for the text modality in one-week intervals. The distributions are plotted using one-year time window (i.e., approximately 52 weeks) around ECG acquisition timestamp.

A.2 Internal Classification Dataset Details

Table 7: Summary of the disease labeling process for coronary atherosclerosis, myocarditis, cardiac amyloidosis, pulmonary hypertension, and low LVEF. The *Case* column details the identification of patients with the specific disease, the *Control* column explains how control patients are produced for the corresponding disease, and the *Time window* indicates the time frame for associating ECGs with diagnoses. (d_1, d_2) denotes the time frame extending d_1 days before and d_2 days after the first diagnosis timestamp. All ECGs of the control patients are taken into the control cohort. The abbreviations CAC, TTE, mPAP, TRV, and LVEF represent coronary artery calcium, transthoracic echocardiogram, mean pulmonary arterial pressure, tricuspid regurgitation velocity, and left ventricular ejection fraction, respectively.

Disease	Case	Control	Time window
Coronary atherosclerosis	CAC score ≥ 300	CAC score $= 0$	(365, 365)
Myocarditis	Manual curation by expert physicians using EHR data	No history of myocarditis	(7, 7)
Cardiac amyloidosis	Manual curation by expert physicians using EHR data	Patients who have undergone TTE and no history of amyloidosis	(180, 180)
Pulmonary hypertension	mPAP ≥ 25 mmHg or TRV ≥ 3.4 m/s	mPAP ≤ 20 mmHg or TRV ≤ 2.8 m/s	(30, 30)
Low LVEF	LVEF ≤ 40	LVEF > 40	(14, 14)

Table 8: Number of ECGs, number of patients, and disease prevalence in coronary atherosclerosis, myocarditis, cardiac amyloidosis, pulmonary hypertension, low LVEF, and AFib in NSR classification datasets.

	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

In this section, we summarize the disease labeling process and outline the number of ECGs, the number of patients, and the disease prevalence used during training, validation, and testing of internal disease classification models. Table 7 shows the summary of the disease labeling process for coronary atherosclerosis, myocarditis, cardiac amyloidosis, pulmonary hypertension, and low LVEF. In the case of AFib in NSR, for the case-cohort, we encompass all NSR ECGs: those from 30 days before the first occurrence of AFib, during the occurrences, and extending 30 days after the last occurrence of AFib. The control cohort comprises all NSR ECGs obtained from patients with no evidence of AFib. Table 8 shows the number of ECGs, the number of patients, and the prevalence for each disease.

A.3 External Classification Dataset Details

In this section, we provide an overview of the number of ECGs and the number of patients in PhysioNet2020 and Chapman datasets (see Table 9). The physioNet2020 dataset consists of ECG signals from multiple sources with varying signal lengths and sampling rates.

Table 9: Details of the PhysioNet2020 and Chapman datasets, including the number of ECGs, number of patients, signal lengths, and sampling rates.

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

B Model Architectures and Implementation Details

B.1 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).

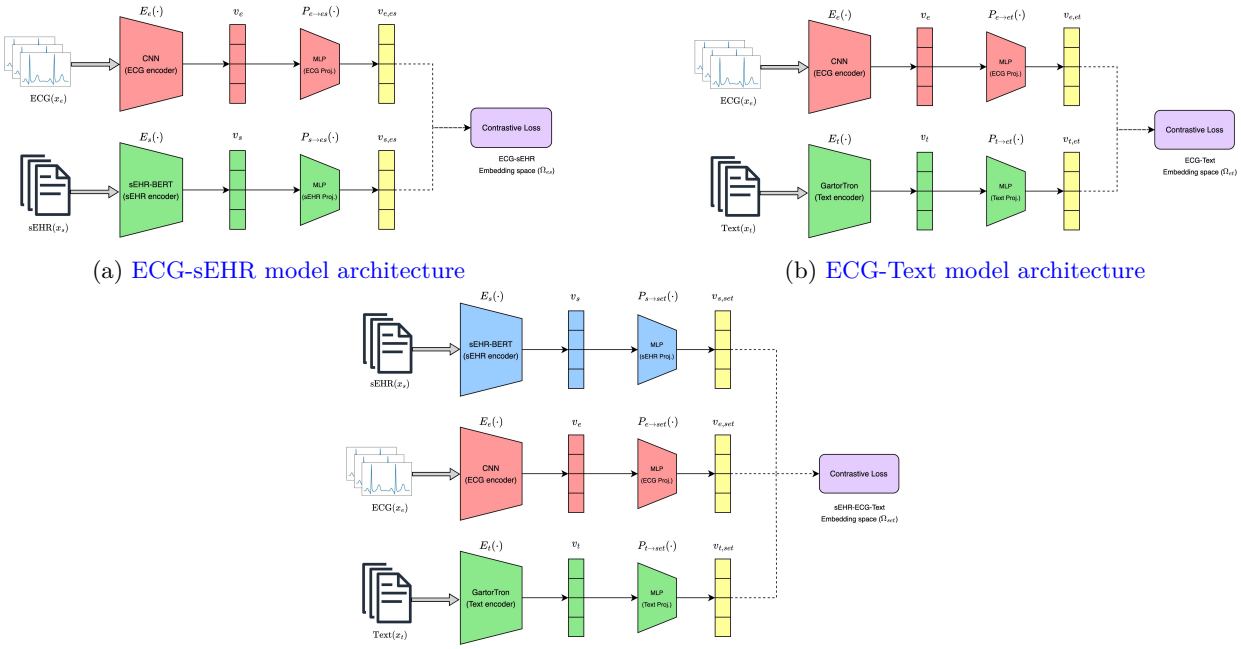
Pre-training Details We initially pre-train sEHR-BERT as described in Section 3.1. For multi-modal contrastive pre-training, we initialize the sEHR encoder with sEHR-BERT model weights and the text encoder with GatorTron-base (Yang et al., 2022) model weights. GatorTron-base (Yang et al., 2022) is a 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. 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 τ to 0.1. We assign equal weighting to both the directions of contrastive learning, i.e., from ECG to sEHR and sEHR to ECG, similarly for ECG and text, i.e., $(\lambda_{es}, \lambda_{et})$ is set to (0.5, 0.5). We used a batch size of 256 and an initial learning rate of 1e-4 for our models. For ECG-only contrastive learning models, we used a batch size of 512 and an initial learning rate of 1e-3. 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 batch size 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.2 Multi-Modal Architectures

In this section, we present the precise architectures of various multi-modal contrastive learning models. Figure 4a and Figure 4b show the architectures of the ECG-sEHR model, described in Section 3.3, and the ECG-Text model described in Section 3.4, respectively. Figure 4c shows the architecture of sEHR-ECG-Text shared space model mentioned in Section 3.2.



(a) ECG-sEHR model architecture

(b) ECG-Text model architecture

(c) sEHR-ECG-Text shared space architecture

Figure 4: The figures (a) and (b) show the architectures of bi-modal contrastive learning methods. (a) illustrates the architecture for the ECG-sEHR model, while (b) presents the architecture for the ECG-Text model. Figure (c) shows the shared space architecture of the sEHR-ECG-Text model, where contrastive loss between ECG and sEHR and, between ECG and text, is applied in sEHR-ECG-Text shared space (Ω_{set}).

B.3 ECG Encoder Architecture

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 10. 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 10: ECG encoder architecture used for all experiments. IC, OC, and K represent the number of input channels, the number of output channels, and kernel size, respectively.

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

C Statistical Testing

We assess the statistical significance of our generalized model, sEHR-ECG-Text, by comparing it to each baseline using a two-sided t-test. This evaluation is conducted on linear classification tasks with 10% training splits. Each experiment is repeated 10 times, with 10 different fractional splits derived from the original training data. Our model significantly outperformed all the baselines with the p-value less than 1e-5. We show the average AUROC score over 10 runs along with 95% confidence intervals in Table 11.

Table 11: The table shows the average AUROC, with 95% confidence intervals in parentheses on linear classification tasks, i.e., coronary atherosclerosis, myocarditis, cardiac amyloidosis, pulmonary hypertension, low LVEF, and AFib in NSR, using 10% training splits. The results were calculated by conducting each experiment 10 times, using 10 different fractional splits obtained from the original training split. The sEHR-ECG-Text model significantly outperforms all the baselines with the p-value less than 1e-5.

Method	Coronary atherosclerosis	Myocarditis	Cardiac amyloidosis	Pulmonary hypertension	Low LVEF	AFib in NSR
Supervised baseline						
Random Init.	0.770 (0.760, 0.779)	0.778 (0.767, 0.789)	0.910 (0.907, 0.913)	0.886 (0.882, 0.890)	0.917 (0.915, 0.920)	0.897 (0.895, 0.898)
ECG-only SSL						
3KG	0.721 (0.715, 0.727)	0.686 (0.669, 0.704)	0.869 (0.862, 0.875)	0.862 (0.860, 0.863)	0.902 (0.900, 0.903)	0.866 (0.865, 0.866)
CLOCS(CMSC)	0.736 (0.729, 0.742)	0.660 (0.649, 0.671)	0.879 (0.875, 0.883)	0.857 (0.856, 0.859)	0.905 (0.904, 0.906)	0.861 (0.861, 0.862)
PCLR	0.759 (0.757, 0.762)	0.716 (0.704, 0.728)	0.898 (0.894, 0.903)	0.895 (0.894, 0.895)	0.927 (0.927, 0.928)	0.898 (0.897, 0.899)
Our model						
sEHR-ECG-Text	0.840 (0.838, 0.843)	0.859 (0.846, 0.872)	0.934 (0.930, 0.937)	0.932 (0.931, 0.933)	0.942 (0.941, 0.943)	0.928 (0.928, 0.929)

D Additional Evaluation Metrics

In this section, we provide the AUPRC scores for classification tasks performed on internal datasets. Table 12 shows AUPRC scores for coronary atherosclerosis, myocarditis, and cardiac amyloidosis classification tasks (an extension of Table 2) and Table 13 shows AUPRC scores for pulmonary hypertension, low LVEF, and AFib in NSR classification tasks (an extension of Table 3). We observe a similar trend in the AUPRC metric when comparing our models to baseline models, as we did with AUROC.

Table 12: An extension of Table 2 to include AUPRC scores (mean and standard deviation over 5 runs with different random seeds) for coronary atherosclerosis, myocarditis, and cardiac amyloidosis classification tasks: (a) linear classification, (b) fine-tuned classification. Results within 95% confidence intervals of the best result are shown in **bold**.

(a) Linear classification

	Coronary atherosclerosis		Myocarditis		Cardiac amyloidosis	
Method	10%	100%	10%	100%	10%	100%
Supervised baseline						
Random Init.	0.686 ± 0.013	0.768 ± 0.006	0.075 ± 0.019	0.239 ± 0.013	0.687 ± 0.009	0.795 ± 0.004
ECG-only SSL						
3KG	0.643 ± 0.014	0.722 ± 0.000	0.047 ± 0.013	0.078 ± 0.000	0.522 ± 0.020	0.628 ± 0.000
CLOCS(CMSC)	0.661 ± 0.011	0.734 ± 0.000	0.033 ± 0.005	0.069 ± 0.000	0.534 ± 0.014	0.670 ± 0.000
PCLR	0.668 ± 0.008	0.761 ± 0.000	0.038 ± 0.010	0.082 ± 0.001	0.618 ± 0.028	0.721 ± 0.001
Our models						
sEHR-ECG-Text	**0.780 ± 0.008**	0.841 ± 0.000	**0.200 ± 0.031**	**0.304 ± 0.001**	**0.775 ± 0.017**	0.842 ± 0.000
ECG-sEHR	**0.773 ± 0.018**	**0.843 ± 0.000**	**0.178 ± 0.028**	0.291 ± 0.001	**0.777 ± 0.019**	**0.845 ± 0.000**
ECG-Text	0.754 ± 0.011	0.819 ± 0.000	0.122 ± 0.029	0.210 ± 0.001	0.692 ± 0.021	0.778 ± 0.001

(b) Fine-tuned classification

	Coronary atherosclerosis		Myocarditis		Cardiac amyloidosis	
Method	10%	100%	10%	100%	10%	100%
Supervised baseline						
Random Init.	0.686 ± 0.013	0.768 ± 0.006	0.075 ± 0.019	0.239 ± 0.013	0.687 ± 0.009	0.795 ± 0.004
ECG-only SSL						
3KG	0.674 ± 0.012	0.756 ± 0.009	0.041 ± 0.006	0.227 ± 0.031	0.667 ± 0.020	0.798 ± 0.004
CLOCS(CMSC)	0.707 ± 0.008	0.747 ± 0.011	0.053 ± 0.018	0.255 ± 0.015	0.693 ± 0.033	0.808 ± 0.002
PCLR	0.735 ± 0.010	0.763 ± 0.005	0.043 ± 0.009	0.257 ± 0.026	0.750 ± 0.010	0.821 ± 0.003
Our models						
sEHR-ECG-Text	**0.822 ± 0.006**	0.838 ± 0.007	0.248 ± 0.012	**0.343 ± 0.019**	**0.805 ± 0.007**	**0.845 ± 0.005**
ECG-sEHR	**0.827 ± 0.004**	**0.845 ± 0.004**	**0.270 ± 0.012**	**0.349 ± 0.013**	**0.812 ± 0.011**	**0.848 ± 0.006**
ECG-Text	0.813 ± 0.006	0.828 ± 0.002	0.196 ± 0.017	0.302 ± 0.008	0.779 ± 0.006	0.831 ± 0.003

E Ablation Study

We perform three ablations: (i) short ICD-10 vs. full ICD-10 codes in the sEHR modality as described in 4.2.2 using the ECG-sEHR model, (ii) report concatenation vs. entity concatenation in text modality as described in 4.2.3 using the ECG-Text model, and (iii) shared space vs. fine and coarse spaces using the sEHR-ECG-Text model as described in Section 3.2. 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 14. 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. In the report concatenation vs. entity concatenation ablation, entity concatenation yielded better results for the majority of the diseases. We speculate that this is because entity concatenation captures better long-term dependencies than report concatenation as we incorporate information from one-year time window around ECGs. Note that the sEHR-ECG-Text model with FAC spaces architecture yielded slightly better results than the shared space architecture.

Table 13: An extension of Table 3 to include AUPRC scores (mean and standard deviation over 5 runs with different random seeds) for pulmonary hypertension, low LVEF, and AFib in NSR classification tasks: (a) linear classification, (b) fine-tuned classification. Results within 95% confidence intervals of the best result are shown in **bold**.

(a) Linear classification

	Pulmonary hypertension			Low LVEF			AFib in NSR		
Method	1%	10%	100%	1%	10%	100%	1%	10%	100%
Supervised baseline									
Random Init.	0.420 ± 0.011	0.585 ± 0.011	0.710 ± 0.005	0.354 ± 0.013	0.527 ± 0.012	0.659 ± 0.003	0.356 ± 0.030	0.545 ± 0.011	0.642 ± 0.004
ECG-only SSL									
3KG	0.385 ± 0.022	0.512 ± 0.005	0.537 ± 0.000	0.322 ± 0.032	0.480 ± 0.006	0.525 ± 0.000	0.375 ± 0.016	0.454 ± 0.003	0.466 ± 0.000
CLOCS(CMSC)	0.394 ± 0.029	0.521 ± 0.003	0.548 ± 0.000	0.369 ± 0.030	0.503 ± 0.008	0.542 ± 0.000	0.373 ± 0.013	0.457 ± 0.002	0.472 ± 0.000
PCLR	0.487 ± 0.043	0.605 ± 0.007	0.633 ± 0.000	0.467 ± 0.036	0.600 ± 0.003	0.630 ± 0.000	0.443 ± 0.022	0.539 ± 0.003	0.556 ± 0.000
Our models									
sEHR-ECG-Text	**0.618 ± 0.012**	0.722 ± 0.003	0.743 ± 0.000	**0.529 ± 0.046**	**0.664 ± 0.003**	**0.695 ± 0.000**	**0.598 ± 0.009**	0.662 ± 0.001	0.670 ± 0.000
ECG-sEHR	**0.627 ± 0.020**	**0.727 ± 0.002**	**0.747 ± 0.000**	**0.542 ± 0.052**	**0.663 ± 0.004**	0.693 ± 0.000	**0.606 ± 0.014**	**0.672 ± 0.002**	**0.679 ± 0.000**
ECG-Text	0.595 ± 0.013	0.689 ± 0.002	0.713 ± 0.000	**0.493 ± 0.030**	0.634 ± 0.003	0.668 ± 0.000	0.561 ± 0.015	0.637 ± 0.002	0.646 ± 0.000

(b) Fine-tuned classification

	Pulmonary hypertension			Low LVEF			AFib in NSR		
Method	1%	10%	100%	1%	10%	100%	1%	10%	100%
Supervised baseline									
Random Init.	0.420 ± 0.011	0.585 ± 0.011	0.710 ± 0.005	0.354 ± 0.013	0.527 ± 0.012	0.659 ± 0.003	0.356 ± 0.030	0.545 ± 0.011	0.642 ± 0.004
ECG-only SSL									
3KG	0.429 ± 0.015	0.600 ± 0.008	0.703 ± 0.004	0.422 ± 0.013	0.582 ± 0.007	0.673 ± 0.003	0.430 ± 0.018	0.569 ± 0.004	0.640 ± 0.004
CLOCS(CMSC)	0.497 ± 0.004	0.605 ± 0.014	0.714 ± 0.004	0.483 ± 0.011	0.602 ± 0.005	0.671 ± 0.006	0.449 ± 0.008	0.563 ± 0.004	0.638 ± 0.003
PCLR	0.545 ± 0.021	0.660 ± 0.010	0.726 ± 0.004	0.546 ± 0.004	0.649 ± 0.002	**0.692 ± 0.002**	0.509 ± 0.013	0.585 ± 0.006	0.643 ± 0.002
Our models									
sEHR-ECG-Text	**0.668 ± 0.012**	0.723 ± 0.006	0.746 ± 0.009	**0.651 ± 0.004**	**0.660 ± 0.008**	**0.692 ± 0.008**	**0.626 ± 0.021**	**0.631 ± 0.006**	**0.658 ± 0.010**
ECG-sEHR	**0.678 ± 0.014**	**0.729 ± 0.004**	**0.752 ± 0.004**	0.643 ± 0.003	**0.664 ± 0.007**	**0.695 ± 0.005**	**0.640 ± 0.014**	**0.634 ± 0.009**	**0.661 ± 0.003**
ECG-Text	0.639 ± 0.009	0.694 ± 0.010	0.731 ± 0.004	0.618 ± 0.008	0.638 ± 0.006	0.678 ± 0.004	0.595 ± 0.016	0.616 ± 0.008	0.655 ± 0.008

Table 14: Results of the ablation study. Comparison of linear classification performance (AUROC, mean and standard deviation over 5 runs with different random seeds) between short ICD codes and full ICD codes (ECG-sEHR), report concatenation and entity concatenation (ECG-Text), and shared space versus fine and coarse spaces (sEHR-ECG-Text). The best results of each ablation are shown in **bold**.

Method	Coronary atherosclerosis	Myocarditis	Cardiac amyloidosis	Pulmonary hypertension	Low LVEF	AFib in NSR
ECG-sEHR						
Short ICD-10 codes	**0.891 ± 0.000**	**0.896 ± 0.000**	**0.959 ± 0.000**	**0.940 ± 0.000**	**0.951 ± 0.000**	0.933 ± 0.000
Full ICD-10 codes	0.888 ± 0.000	0.894 ± 0.000	0.958 ± 0.000	**0.940 ± 0.000**	**0.951 ± 0.000**	**0.934 ± 0.000**
ECG-Text						
Entity concatenation	0.875 ± 0.000	0.881 ± 0.000	**0.949 ± 0.000**	**0.931 ± 0.000**	**0.947 ± 0.000**	**0.925 ± 0.000**
Report concatenation	**0.879 ± 0.000**	**0.895 ± 0.001**	0.939 ± 0.000	0.921 ± 0.000	0.938 ± 0.000	**0.925 ± 0.000**
sEHR-ECG-Text						
Fine and coarse spaces	**0.890 ± 0.000**	**0.901 ± 0.001**	**0.960 ± 0.000**	**0.939 ± 0.000**	**0.951 ± 0.000**	**0.931 ± 0.000**
Shared space	0.883 ± 0.000	0.884 ± 0.001	0.955 ± 0.000	0.936 ± 0.000	0.949 ± 0.000	0.928 ± 0.000