HuBERT-ECG: a self-supervised foundation model for broad and scalable cardiac applications

Edoardo Coppola[1], Mattia Savardi[2], Mauro Massussi[2,3], Marianna Adamo[2,3], Marco Metra[2,3], Alberto Signoroni[2]

[1] Department of Information Engineering – University of Brescia, Brescia, Italy
[2] Dept. of Medical and Surgical specialties, Radiological Sciences and Public Health – Univ. of Brescia, Brescia, Italy
[3] Cardiac Catheterization Laboratory and Cardiology, ASST Spedali Civili di Brescia, Brescia, Italy

Abstract

Deep learning models have shown remarkable performance in electrocardiogram (ECG) analysis, but their success has been constrained by the limited availability and size of ECG datasets, resulting in systems that are more task specialists than versatile generalists. In this work, we introduce HuBERT-ECG, a foundation ECG model pre-trained in a self-supervised manner on a large and diverse dataset of 9.1 million 12-lead ECGs encompassing 164 cardiovascular conditions. By simply adding an output layer, HuBERT-ECG can be fine-tuned for a wide array of downstream tasks, from diagnosing diseases to predicting future cardiovascular events. Across diverse real-world scenarios, HuBERT-ECG achieves AUROCs from 84.3% in low-data settings to 99% in large-scale setups. When trained to detect 164 overlapping conditions simultaneously, our model delivers AUROCs above 90% and 95% for 140 and 94 diseases, respectively. HuBERT-ECG also predicts death events within a 2-year follow-up with an AUROC of 93.4%. We release models and code.

Introduction

The electrocardiogram (ECG) has long been a cornerstone of cardiovascular diagnostics, serving as a non-invasive, cost-effective, and widely available tool for assessing cardiac and noncardiac diseases[1]. Through its recurring waveforms, the ECG captures a unique “language of the heart”, encoding both physiological and pathological vital information through distinctive electrophysiological “fingerprints”[2]. This language, although seemingly simple in its structure, reveals complex and meaningful patterns that allow deviations from expected cardiac function to be identified. By examining these signals, we gain valuable insights into the heart's health, making the study of its language both scientifically essential and clinically transformative.

Although the widespread use of the ECG, integrating deep learning (DL) into its analysis represents a transformational opportunity to significantly improve its clinical utility[3,4]. In fact, DL algorithms can transform the raw ECG traces into powerful digital biomarkers, enabling early detection, risk stratification, and intervention across a wide range of cardiovascular conditions. Previous studies have demonstrated the robust potential of DL in specific areas of ECG interpretation. For example, DL models have consistently shown solid performance in the detection of conditions where the ECG is the gold standard for diagnosis (e.g. atrial fibrillation, tachycardia, and bradycardia), leading to validated algorithms[5]. In contrast, DL applications in the diagnosis of morphological conditions, such as heart failure, pulmonary thromboembolism and aortic stenosis, are less widespread but growing[6]. Although these conditions rely primarily on imaging modalities, there is valuable work demonstrating that DL can successfully use ECG data to predict cardiovascular conditions beyond those traditionally

associated with ECG patterns, providing a valuable complement to conventional imaging approaches[7]. Finally, the use of DL to predict future cardiovascular events (CVE), such as myocardial infarction and stroke, represents the most ambitious application of this technology. While early results are promising, this area presents greater challenges due to the complexity of predicting outcomes from ECG patterns, making it an active and exciting field of research[8].

From a machine learning perspective, existing limitations in terms of size of available ECG datasets, range of clinical conditions, patient numerosity and demographic diversity have led to the development of ECG models that are more task specialists[9] rather than competent and versatile generalists, traits that typically characterise foundation models[10]. In fact, specialised models often struggle to adapt to new domains or even different distributions within the same task[11,12]. For example, if an ECG model is trained on a dataset in which every recording has been labelled as positive or negative for atrial fibrillation, such model can only detect atrial fibrillation and is not able to diagnose other conditions without being retrained on the new dataset and its cardiac abnormalities. Similarly, if the ECGs were all taken from people aged over 65, i.e., where the condition is most common, the model might struggle to detect the same condition in a dataset of younger patients. In stark contrast, by self-supervised pre-training on large and diverse unlabelled datasets, foundation models can learn robust and generalised data representations that are transferable to a wide variety of downstream tasks, requiring minimal fine-tuning for specific domains or distributions[13].

Unimodal foundation models have achieved significant success across traditional domains—including natural language processing (e.g., the GPT series[14–17], BERT[18], RoBERTa[19], T5[20]), computer vision (e.g., DINO[21], MAE[22]), and audio processing (e.g., Wav2Vec2[23], HuBERT[24])—by leveraging large-scale self-supervised pre-training on extensive unannotated datasets. More recently, these advancements have also facilitated multimodal learning, where models integrate multiple data modalities to achieve a more holistic understanding of information within data (e.g., Google Gemini[a], GPT-4o[b]). In the medical domain, progress has been slower due to the limited availability of large medical datasets[13,25]. However, this trend has begun to shift, leading to the emergence of foundation medical models such as CONCH[26], Prov-GigaPath[27] and Virchow[28] for computational pathology; KAD[29] and CheXzero[30] for radiology; MedSAM[31], for medical image segmentation; EchoCLIP[32] for echocardiography; and BiomedGPT[33] for various biomedical tasks. However, despite these advances, there remains a significant gap: the lack of a foundation model for electrocardiography that is pre-trained on a truly large-scale dataset using self-supervision. Such a model would need to be (1) versatile and adaptable to various use cases, especially where data is scarce; (2) capable of achieving high performance with minimal fine-tuning; and (3) designed to support research in underexplored areas.

In this work, inspired by HuBERT[24] architecture, we present HuBERT-ECG, a new ECG foundation model capable of performing a wide range of tasks, from diagnosing diseases to predicting future CVEs. Unlike most ECG models, HuBERT-ECG is pre-trained in a self-supervised manner on an extensive dataset of 9.1 million ECGs measured from a large and diverse population across four countries. The annotated instances in our dataset, which are more than 2.4 million ECGs, are labelled with one or more conditions from a comprehensive set of 164 diagnoses, allowing our model to learn and generalise over a wide spectrum of cardiac pathologies. We evaluate the proposed model on 16 datasets and their aggregation, simulating real-world scenarios and use cases of varying complexity. HuBERT-ECG demonstrates both efficiency and accuracy across all datasets, showing the potential to address three key areas of clinical practice: 1) identification of cardiac conditions where the ECG

[a] https://gemini.google.com/
[b] https://openai.com/index/hello-gpt-4o/

is the primary diagnostic tool; 2) diagnosis of morphological conditions where the ECG plays a supportive diagnostic role; and 3) prediction of future CVEs.

Finally, to support ongoing research on cardiovascular diseases, we release HuBERT-ECG models and code to the community. HuBERT-ECG is available in three model sizes or configurations—SMALL, BASE, and LARGE—allowing researchers to choose the configuration that best meets the unique requirements of their specific use cases. This range in model size is designed to accommodate different levels of complexity across applications, from datasets with limited examples and simple conditions to those requiring nuanced recognition of challenging, multifaceted cardiac issues. By providing insights from our dataset evaluations, aimed at simulating diversified real-world scenarios, we intend to guide the effective adoption and customisation of HuBERT-ECG for a variety of clinical and research applications.

Results

ECG Representation Learning through Self-supervised Pre-training

To learn robust and transferable 12-lead ECG representations, we draw inspiration from HuBERT[24], a powerful self-supervised foundation model for speech representation learning. HuBERT's pre-training approach involves predicting predetermined cluster assignments for masked continuous speech embeddings through multiple pre-training iterations, effectively capturing both acoustic and linguistic information in an increasingly refined manner. After collecting, pre-processing and assembling multiple data sources into a large and diverse dataset of 9.1 million ECG instances covering four countries (Fig. 1a-d, Fig. 2a), as described in "Methods" (*Data and Pre-processing*), we follow a similar approach to pre-train HuBERT-ECG (Fig. 2b), as outlined in "Methods" (*HuBERT-ECG Architecture and Theoretical Framework*). In particular, as detailed in "Methods" (*Unsupervised Label Discovery*), clustering models are fitted offline on feature descriptors of raw ECG fragments to generate cluster assignments for masked ECG embeddings. These feature descriptors, along with the clustering models built upon them, evolve between pre-training iterations. The BASE model configuration undergoes two pre-training iterations: we use Mel Frequency Cepstral Coefficients[34] as feature descriptor during the first iteration, while we employ latent ECG representations from intermediate model layers during the second one. Finally, to assess scalability, we also pre-train a SMALL and a LARGE configuration, both in a single iteration, using the latent representation extracted from the intermediate layers of the BASE model. The specifications and implementation details of the pre-training are provided in "Methods" (*Implementation* and *Self-supervised Pre-training*), while a summary of the architecture of the three model configurations is given in Table 1.

Supervised fine-tuning on downstream data and evaluation

To evaluate the performance of HuBERT-ECG in different real-world settings of varying complexity, we compile 16 downstream datasets that simulate a range of situations and use cases, including large and small ECG collections, labelled with many or few possible conditions, often unevenly distributed. These data sources (Fig. 1a) include: the labelled partition of Ribeiro[35] (also known as CODE); CPSC and CPSC-Extra[36]; PTB[37]; PTB-XL[38] which, with 6 different sets of conditions, gives rise to 6 different datasets; the publicly available partition of Georgia[39]; Chapman-Shaoxing[40] (Chapman); Ningbo First Hospital[41] (Ningbo); part of the dataset used in Tianchi Arrhythmia Competition[c] (Hefei); Shandong Provincial Hospital[42] (SPH); and SaMi-Trop[43]. HuBERT-ECG is evaluated on extracted hold-out test sets after fine-tuning procedures requiring only a small fraction of the pre-

[c] https://tianchi.aliyun.com/competition/entrance/231754/introduction

training time. Additionally, to mirror the real-world clinical settings that cardiologists navigate daily—managing a diverse spectrum of cardiovascular diseases across patients with varying health backgrounds and complexities—we combine all annotated datasets into a new, comprehensive dataset, which we name *Cardio-Learning*. This dataset consists of 2.4 million ECGs from four countries labelled with one or more conditions from a comprehensive, highly imbalanced set of 164 conditions. We present HuBERT-ECG's results on these 16+1 datasets and compare them with those achieved by training the same model from scratch. We also benchmark our performance against the state-of-the-art wherever possible. To maintain evaluation consistency, address label distribution imbalances, and apply a robust, threshold-invariant metric, we report performance using the area under the receiver operating characteristic curve (AUROC). Fine-tuning procedures are detailed in "Methods" (*Supervised Fine-tuning*). We present HuBERT-ECG results, organized across diverse application scenarios, to highlight its potential as a versatile tool for supporting cardiologists in various clinical contexts.

HuBERT-ECG is efficient in low-data diagnostic settings. In diagnostic low-data scenarios, such as those simulated by PTB, CPSC and CPSC-Extra datasets, HuBERT-ECG exhibits distinct macro trends. For PTB (Fig. 3a), featuring 515 instances and 17 possible conditions, even co-occurring, the SMALL and BASE configurations perform comparably well, with fine-tuned models achieving AUROCs of 84.3% (± 2.9%) and 84.8% (± 3.0%), respectively, while the LARGE one lags behind with a score of 82.2% (±2.6%). In CPSC (Fig. 3b), characterised by 6,878 examples and 9 possibly concurrent conditions, the SMALL size model achieves an AUROC of 94.5% (± 0.2%), while the BASE and LARGE models outperform the former with scores of 95% (± 0.2%) and 94.9% (± 0.6%), respectively. On this dataset, our best model is 3 points behind that of Na et al.[44], which is specifically pre-trained and fine-tuned for arrhythmia detection only. On CPSC-Extra (Fig. 3c), a much more difficult context characterised by 3,453 ECGs and 52 possibly co-existing conditions, HuBERT-ECG SMALL delivers the best performance, reaching an AUROC of 89.4% (± 7.6%), whereas the BASE and LARGE variants are likely to overfit and provide inferior results of 75.6% (± 2.0%) and 77.4% (± 2.3%), respectively. Fine-tuned models also provide 3-5.8% improvements in AUROC over their randomly initialised counterparts, with the minimal exception of the BASE configuration in CPSC-Extra.

HuBERT-ECG remains efficient in increasingly difficult contexts. The difficulty of solving diagnostic tasks by learning from examples in a dataset certainly depends on both the number of examples and the number of the diagnostic classes. However, such numbers are not sufficient to provide a good complexity estimate as some classes may be intrinsically harder to detect than others. In general, a small number of instances per class increases complexity, as does a large number of possible conditions, but augmenting the number of instances only for a few classes can skew the label distribution and complicate the learning process. These challenges are evident in the Georgia and Chapman datasets, where the 62 conditions represented in the former and the 54 in the latter are far from being evenly distributed across the 10,345 and 10,247 examples in the respective data sources. In these complex diagnostic scenarios, fine-tuning HuBERT-ECG consistently leads to better performance than training from scratch, yielding improvements of 6.5-7.8% on Georgia and 3.9-6.1% on Chapman. Specifically, for Georgia (Fig. 3d), HuBERT-ECG SMALL achieves an AUROC of 81.9% (±0.5%), which increases to 83.2% (± 0.7%) with the BASE model, but decreases slightly to 82.1% (± 0.6%) with the LARGE configuration. On Chapman (Fig. 3e), the SMALL size model reaches an AUROC of 85.9% (± 0.8%), with a slight dip to 85.5% (± 0.7%) for the BASE version, before rising again to 85.7% (± 0.7%) with the LARGE one.

HuBERT-ECG enables accurate diagnostics in large-scale, complicated scenarios. Since they are pre-trained on an extensive dataset, diagnostic foundation models are expected to guarantee high performance in large-scale, complicated scenarios. In these situations, a model is expected to identify

a disease among many possible conditions even when it has not seen a large number of significant examples to learn that disease, or when it must account for significant variations between patients, which can heavily influence its predictions. To explore these scenarios, we leverage large datasets with a few to tens of possible conditions including Ribeiro, PTB-XL, Hefei, SPH and Ningbo. On Ribeiro (Fig. 3f), which is a large-scale dataset with more than 2.3 million labelled ECGs and 6 possible conditions, every HuBERT-ECG configuration performs excellently, regardless of weights initialization. The BASE model performs best, achieving an AUROC of 99.89% with fine-tuned weights and a score of 99.84% with randomly initialised ones (i.e., trained from scratch). The SMALL and LARGE configurations perform similarly, obtaining AUROCs of 99.72% and 99.74%, in the former case, and 99.82% and 99.83% in the latter. On this dataset, with an abundance of examples to learn from, pre-training does not show much advantage over training from scratch. Nonetheless, our best model, i.e., the fine-tuned HuBERT-ECG BASE, outperforms the model developed by Ribeiro et al.[35] in terms of sensitivity, specificity, AUROC and area under precision and recall curve (AUPRC), while other model configurations are better only according to certain metrics (Table 2). PTB-XL features 21,837 instances annotated with 71 possible labels covering diagnostic, form and rhythm conditions. The authors also state that 44 diagnostic conditions can be aggregated into 23 diagnostic subclasses and 5 more coarse diagnostic superclasses, while there are 12 possible rhythm statements and 19 different form abnormalities. Therefore, one can evaluate a model on 6 different datasets, referred to as PTB-XL *All*, *Form*, *Rhythm*, *Diagnostic* (*Diag.*), *Diag. Subclass*, *Diag. Superclass*, that differ from each other in the reported labels. On these data sources, fine-tuning pre-trained models shows its advantages for almost each model configuration when compared to its randomly initialised counterpart (Fig. 3g-l). Our fine-tuned models show good performance, with HuBERT-ECG BASE achieving an AUROC greater than 90% in 5 out of 6 datasets. Noteworthy, these results are close to some of the best models from the literature (Table 3) that are optimized for the PTB-XL benchmark, i.e., pre-trained on datasets where PTB-XL is the predominant component and specifically fine-tuned on it. In our case, PTB-XL is a very small fraction of the pre-training set and represents one of many different scenarios HuBERT-ECG can handle. On Hefei, which is characterised by 20,036 samples and 29 conditions, all fine-tuned models surpass the respective randomly initialized versions. When fine-tuned, HuBERT-ECG SMALL achieves an AUROC of 96.05% (± 0.36%), while the BASE and LARGE fine-tuned variants perform slightly better and marginally worse, respectively, with AUROCs of 96.61% (± 0.25%) and 95.36% (± 0.63%) (Fig. 3m). In SPH, a novel dataset of 25,770 ECGs annotated with 44 primary diagnostic statements, all our models achieve similar AUROCs that increase with the model size. In particular, the fine-tuned HuBERT-ECG LARGE obtains the highest AUROC with a score of 94.3%, followed by the BASE and SMALL configurations with 93.8% and 93.5%, respectively (Fig. 3n). To our knowledge, we are the first to address this dataset in its full complexity. Finally, on Ningbo (Fig. 3o), a dataset with 34,905 examples and 76 possibly co-existing conditions, the fine-tuned SMALL model achieves an AUROC of 93.25% (± 0.52%), while the randomly initialised variant achieves 92.95% (± 0.68%). When fine-tuned, the BASE model performs slightly better, achieving an AUROC of 94.39% (± 0.53%), compared to 91.76% (± 1.26%) when trained from scratch. The LARGE model performs marginally worse, reaching 93.96% (± 0.55%) with fine-tuned weights and 91.80% (± 0.78%) with random initialisation.

HuBERT-ECG can be a good predictor of future cardiovascular events. Although ECG-based works on the prediction of future CVEs have shown promising results, this field is still underexplored and the number of publicly available datasets is limited. The SaMi-Trop[43] dataset provides an opportunity to evaluate HuBERT-ECG in this domain. It includes 1632 ECGs from patients with Chagas disease monitored during a 2-year follow-up period, 268 of which are normal recordings and 104 are marked as “death events”. Interestingly, there are two trends in the prediction of mortality events within the follow-up period that seem to be at odds with the results presented so far: 1) randomly initialised models outperform fine-tuned counterparts; 2) their performance improves as

their size increases (Fig. 3p). In fact, HuBERT-ECG SMALL with random initialisation achieves an AUROC of 82.9% (± 6.6%), compared to 74.5% (± 4.3%) with fine-tuned weights, while the LARGE size obtains AUROCs of 83.6% (± 7.0%) and 73.6% (± 15.1%) with and without a random initialisation, respectively. In contrast, the BASE configuration achieves AUROCs of 82.5% (± 6.8%) with randomly initialised weights and 78.7% (± 5.6%) with fine-tuned parameters. The seemingly paradoxical finding of larger models performing better with limited data can be attributed to the intrinsic complexity of predicting *future* CVEs—a task that evidently benefits from models with more trainable parameters. In parallel, the inferior adaptability of the pre-trained HuBERT-ECG to this context is only apparent, as we show in the next section that a different fine-tuning strategy can provide enormous improvements in this task. For comparison purposes, we benchmark our performance against that by Ferreira et al.[45], representing the state-of-the-art on this dataset. Unlike us, they used a Random Forest estimator fitted on a combination of handcrafted ECG features, sociodemographic variables and self-reported symptoms. Rather than using a more robust, threshold-independent metric, they evaluated their model in terms of G-mean score, where the G-mean=$\sqrt{sensitivity \cdot specificity}$. In terms of G-mean, our best model surpasses theirs, reporting a score of 77.8% against one of 77%, while, in terms of AUROC, no comparison is possible due to the lack of predictions and code.

HuBERT-ECG navigates extremely complex scenarios. In real-world clinical practice, cardiologists are expected to diagnose a wide spectrum of heart-related conditions, adapting to the nuances of each patient's presentation and health profile, regardless of how frequently a condition appears. To emulate this complexity, we employ newly assembled, comprehensive *Cardio-Learning* dataset, which contains more than 2.4 million examples and 164 potentially overlapping conditions, where ECG is the primary diagnostic tool, a non-primary supportive diagnostic tool, or used to estimate the risk of future CVEs. The fine-tuned HuBERT-ECG BASE provides a macro-averaged AUROC of 86.64%, while HuBERT-ECG SMALL and LARGE follow approximately at the same distance with scores of 84.43% and 84.52%, respectively (Fig. 4a-c). When the conditions are grouped according to the ECG diagnostic role—primary diagnostic tool or supportive—the BASE model configuration achieves AUROCs of 86.21% and 88.67%, respectively. In contrast, the SMALL and LARGE models lag behind, with AUROCs of 84.33% and 85.02% for the SMALL model, and 84.24% and 85.98% for the LARGE model. In particular, as mentioned in the previous paragraph, we observe a huge improvement in the prediction of death events in SaMi-Trop cases, where the SMALL, BASE and LARGE configurations reach new AUROC values of 83.4%, 93.4%, and 87.5%, respectively.

Discussion

In this paper, we present HuBERT-ECG, a new foundation model for ECG analysis available in three scalable configurations (SMALL, BASE, and LARGE) to meet different deployment needs. Pre-trained in a self-supervised manner on a massive dataset of 9.1 million ECGs from diverse populations in four countries, HuBERT-ECG demonstrates strong performance across 164 cardiovascular conditions, as validated on 16+1 datasets. When fine-tuned, the model configurations achieve AUROCs above 90% for 135, 140, and 134 conditions, respectively, and exceed 95% AUROC for 95, 94, and 94 conditions, respectively (Fig. 4a-c), highlighting the generalisability of its self-supervised representations across highly different contexts. These contexts are reflected in the 16+1 data sources used, ranging from small to very large ECG collections, each annotated with either small or large sets of possibly overlapping conditions. By analysing the model performance, concerning the future use of HuBERT-ECG, we aim to (1) shed light on the data requirements necessary for effective fine-tuning, and (2) guide performance optimisation on new datasets.

In low-data diagnostic settings, which are common and require reduced data collection effort, HuBERT-ECG shows both efficiency and accuracy, even when the number of conditions grows faster than the number of examples, resulting in highly imbalanced label distributions. In these scenarios, the SMALL and BASE configurations are the most suitable options, demonstrating superior generalisation capabilities, while the LARGE configuration is less recommended. As the size of the dataset increases in terms of patients, examples and conditions, the diagnostic scenarios we simulate become more reflective of the real world. In these contexts, although requiring appropriate collection efforts for fine-tuning, HuBERT-ECG delivers remarkable results regardless of label quality and granularity. Scaling up the model size is also beneficial, although smaller configurations maintain competitive performance, thus providing flexibility for complicated diagnostic settings. Notably, on Ribeiro, the largest dataset we use, HuBERT-ECG demonstrates the ability to achieve excellent precision-sensitivity and specificity-sensitivity trade-offs (Table 2).

Furthermore, HuBERT-ECG demonstrates robust performance in highly complex scenarios, such as those well represented by Cardio-Learning, which aggregates ECGs from multiple sources and countries without task-specific separation. In this setup, HuBERT-ECG effectively identifies conditions where the ECG serves as a primary or supportive diagnostic tool. Interestingly, fine-tuning on Cardio-Learning allows HuBERT-ECG to leverage inter-relationships between conditions, resulting in marked improvements in AUROC for conditions that appear only in single subsets of Cardio-Learning. For example, when the BASE model is fine-tuned on the PTB dataset, the AUROCs for coronary heart disease (CHD) and heart failure (HF), conditions that appear only in this dataset, are 76.56% and 75.14%, respectively. At the same time, fine-tuning on the larger Cardio-Learning dataset raises these values significantly to 97.33% for CHD and 99.62% for HF. Similarly, in classifying death events among SaMi-Trop patients, the model achieves AUROCs of 83.4%, 93.4%, and 87.5% in the SMALL, BASE, and LARGE configurations, respectively. These findings suggest that learning to recognise a wide range of cardiovascular conditions simultaneously may improve the model's predictive accuracy for future CVEs or when the ECG is not the primary diagnostic tool.

While HuBERT-ECG is currently focused on predicting a variety of cardiovascular events, there is considerable potential to expand its application toward more personalised clinical insights, including predicting patient response to therapies. For example, in the management of heart failure, HuBERT-ECG could be fine-tuned to predict which patients are likely to benefit from resynchronisation therapy, thereby refining patient selection criteria[46]. This capability would support a more personalised approach to therapy, aimed not only at predicting adverse events but also at improving outcomes, reducing hospitalisations, and improving quality of life. Such advancements could pave the way for a new era of personalised cardiovascular care, where ECG-based foundation models help clinicians tailor therapies with greater precision[47].

Despite its highly promising results, our approach has certain limitations. First, the countries from which the data is collected, although more than in previous studies, do not include large and populated regions (e.g. Africa and India), which we aim to cover in future versions. Second, the scarcity of accessible ECG datasets for predicting future CVEs partially limits our ability to fully assess the model's potential in this still underexplored area. A wider availability of suitable datasets to test the model in this challenging field is desirable to advance research. Third, making direct comparisons with previous studies is challenging, as many focus on limited subsets of available datasets—often reducing the actual number of conditions—, do not open-source their implementations, or do not follow standard practices in metric computation, thereby hampering reproducibility and fair benchmarking. Finally, we note that while some conditions are better classified in Cardio-Learning than in their original datasets, others show lower performance. Further investigation into the reasons for this decrease is crucial to determine whether the limitations stem from confounding factors that may arise when predicting numerous co-occurring conditions with possibly overlapping patterns

simultaneously, or from potential inaccuracies in the ECG ground-truth labels, as is the case, for instance, with PTB-XL[38].

Methods

Data and Preprocessing

Most ECG-related studies not focused on specific clinical questions rely on datasets from the Physionet Challenges[48–50] and overlook other large, valid sources. In contrast, for HuBERT-ECG pre-training and fine-tuning, we use both public and access-on-demand 12-lead ECG datasets, including the labelled and unlabelled partitions from Ribeiro[35], CPSC and CPSC-Extra[36], PTB[37] and PTB-XL[38], the publicly available partition from Georgia[39], Chapman-Shaoxing[40] and Ningbo First Hospital[41], the partition from the Tianchi Arrhythmia Competition[d], Shandong Provincial Hospital[42], MIMIC-IV ECG[51], and SaMi-Trop[43]. For each collected labelled dataset, we homogenise the names of all the conditions to avoid having the same labels under different names in different sources. For self-supervised pre-training, we remove all labels and clinical annotations from the aforementioned datasets, excluding SaMi-Trop, creating an unlabelled dataset of 9.1 million ECGs (Fig. 2a – Pretrain). The effectiveness of the data selection in capturing ECG signal diversity is illustrated through a UMAP[52] projection of the ECG embeddings (Fig. 1e). To assess HuBERT-ECG downstream utility, we fine-tune and test it on every collected dataset, with PTB-XL generating 6 different datasets that differ from each other in the presented conditions. In addition, we create a challenging new dataset, which we name *Cardio-Learning*, by merging all the above sources into one, comprising over 2.4 million ECGs with 164 potentially co-occurring conditions that can be grouped into three categories based on the diagnostic role of the ECG, as shown in Fig. 1f. Remarkably, as shown in Figs. 1b and 1d, the ECGs used in this study were measured from patients with a broad age distribution and diverse geographical backgrounds spanning four countries. To our knowledge, this is the largest and possibly one of the most diverse dataset ever assembled in terms of the number of conditions, demographics, and geographic origin of the individuals.

As pre-processing, similar to Natarajan et al.[53], we first apply a finite impulse response bandpass filter to exclude frequencies outside the range [0.05, 47] Hz, which is reported to contain the dominant components of P waves, T waves and QRS complexes[54]. Secondly, we investigate how sampling rate, which has no standard value and regulates the degree of dilution of the information content, affects both the upstream and downstream performance. We experimentally find that resampling the ECGs at 100 Hz provides the optimal trade-off between downstream performance and training time, while preserving all the meaningful physiological information content according to the Nyquist-Shannon theorem (Supplementary Information – Sec. 1.2). We then rescale our signal to the [-1, 1] range, analogously to what was done by Natarajan et al.[53]. Finally, unlike other works, we use 5-second 12-lead ECGs instead of 10-second recordings, hence halving memory consumption and speeding up both training and inference. In addition, the selected temporal and spectral parameters are compatible with those derived in Mehari & Strodthoff[55].

HuBERT-ECG Architecture and Theoretical Framework

A schematic illustration of the HuBERT-ECG architecture, its pre-training phases and its fine-tuning are shown in Fig. 2b.

Discovering fine labels for ECG fragments. We consider HuBERT-ECG to operate as a masked prediction model, henceforth denoted by the letter *h*. First, standard 12-lead ECGs are flattened into 1D signals and then split into non-overlapping fragments. This fragmentation is necessary to frame

[d] https://tianchi.aliyun.com/competition/entrance/231754/introduction

short portions of the ECG signal from which we can extract feature descriptors to fit a clustering model. Its purpose is to discover and provide the fragments with cluster assignments (i.e., labels) that are finer than the coarse wave-based ones used by Choi et al.[56]. Conceptually, let $X = [x_1, x_2, \ldots, x_N]$ be a flattened 12-lead ECG composed of N non-overlapping fragments x_i, $F = [[f_1, f_2, \ldots f_d]_1, [f_1, f_2, \ldots f_d]_2, \ldots, [f_1, f_2, \ldots, f_d]_N]$ the set of d-dimensional features f_i extracted from each fragment, and K a clustering model fitted to F that finds C different clusters. Then, for each index $i = 1, 2, \ldots, N$, under the hypothesis that f_i is a good descriptor of x_i, we can claim that $K(f_i) = z_i \in \{1, 2, \ldots, C\}$ is the label assignment of x_i and construct $Z = [z_1, z_2, \ldots, z_N]$, the sequence of such assignments for each ECG fragment. For the sake of both terminological flexibility and clarity, we will henceforth use the terms labels and cluster assignments interchangeably.

Representation learning by predicting labels for masked embeddings. Subsequently, the reshaped ECG is fed into a convolutional waveform embedder that captures local contextual information and generates $E = [e_1, e_2, \ldots, e_N]$, a sequence of continuous ECG embeddings. A set of random indices $M = \{j \mid j \in \{1, 2, \ldots, N\}\}$, such that $|M| < N$, is then generated to determine which embeddings are to be masked, i.e., replaced by a special learnable embedding e_{MASK}. After masking, a Transformer encoder[57], which, instead, learns global contextual information, consumes the new masked sequence of embeddings E' and produces, for all the N indices, a probability distribution over the C possible labels: $p_h(E', i)\ \forall i \in \{1, 2, \ldots, N\}$. Such a distribution is eventually used in a standard cross-entropy loss that, however, considers only the indices of M: $L(h; E, M, Z) = \sum_{j \in M} log(p_h(E', j))$. By training the model h to predict cluster assignments of masked embeddings E'_j, $j \in M$, which correspond to fragments of the input ECG not seen by the encoder, we force it to learn the most from the visible ones.

Multi-task learning via cluster ensembles. To increase the granularity of the representations being learned during the pre-training, one can use labels generated by an ensemble of clustering models characterised by an increasing number of clusters. This is equivalent to a multi-task learning framework where tasks are generated as clusters are being discovered. The rationale for this design choice is that, while a single clustering model may introduce imprecise or coarse cluster assignments, an ensemble of models with an increasing number of clusters may mitigate the introduction of noisy targets and provide useful complementary information to the model. Denoting by Γ the number of clustering models composing the ensemble, which is equal to the number of tasks being solved, we can rewrite the loss function as $L(h; E, M, \{Z^\gamma\}) = \sum_{\gamma \in \Gamma} \sum_{j \in M} log(p_h^\gamma(E', j))$, where Z^γ is the sequence of cluster assignments generated by the γ-th clustering model and $\{Z^\gamma\}$ is the set of cardinality Γ comprising all such sequences.

Refining cluster assignments. As in the work of Hsu et al.[24], it is possible to improve the quality of the learned representations by generating an even more refined cluster assignment of the ECG fragments for subsequent pre-training iterations. However, this "finer" mapping does not refer to a temporal refinement, but rather to a more nuanced clustering of the fragments. To generate it, we can cluster latent representations extracted from intermediate layers of the partially pre-trained model h. Therefore, even though the cluster assignments remain aligned with the original fragments in input X, their higher quality after this refinement impacts positively on downstream performance.

Implementation

While the design of HuBERT-ECG follows that proposed by Hsu et al.[24], we make two modifications: 1) the initial convolutional embedder and 2) the masking strategy. First, since ECG signals are sampled rather sparsely compared to audio signals, we do not need high-stride convolutions with large filters and observe that a shallower convolutional block with narrower filters works equally well. In our model, the convolutional embedder generates embeddings at a 0.64 second framerate

from a flattened 5-second 12-lead ECG sampled at 100 Hz (the downsampling factor is 64x). Second, instead of randomly selecting p% of the embeddings as starting points for masked spans, we choose to mask only the selected embeddings without spanning over adjacent ones. This approach avoids the imprecision caused by overlapping spans, which makes it difficult to determine the exact number of masked embeddings. Our method is equivalent to constructing singleton spans and provides more accurate control over the masking process, allowing us to tune the optimal value of p to use during the self-supervised pre-training (Supplementary Information – Sec. 2.1). Additionally, masking single embeddings rather than spans has also been shown to be effective in the work of Hu et al.[58]

After masking, the Transformer encoder consumes the masked sequence of embeddings and produces an output sequence $O = [o_1, o_2, \ldots, o_N]$. The probability distribution over the cluster assignments from a generic clustering model of the ensemble is parameterised by a look-up embedding matrix A^{γ} with shape $C^{\gamma} \times W$ according to the following formula:

$$p_h^{\gamma}(E', i) = \frac{exp(cos\_sim(B^{\gamma} \cdot o_i,\ A_k^{\gamma})/\tau)}{\sum_{c\prime=1}^{C} exp(cos\_sim(B^{\gamma} \cdot o_{c\prime}, A_{c\prime}^{\gamma})/\tau)}, \forall k \in \{1, \ldots C^{\gamma}\}$$

where C^{γ} is the number of clusters found by the γ-th clustering model, B is a projection matrix to make O match the embedding dimension W, A_k is the look-up embedding for the k-th cluster assignment, $cos\_sim(\cdot,\cdot)$ is the cosine similarity between two vectors, and τ is the temperature that scales the logits. The superscript γ denotes the γ-th task being solved when using an ensemble of Γ clustering models. In particular, there are as many projection and look-up embedding matrices as there are tasks in the ensemble.

Since we have a large and diverse dataset, we propose HuBERT-ECG in the SMALL, BASE and LARGE model sizes. As we scale the size, we keep the same convolutional embedder and increase the encoder depth and width, as well as the label embedding dimension W. Table 1 summarises the architecture of these three versions of our model. After pre-training, to fine-tune HuBERT-ECG for specific downstream tasks, we delete the look-up embedding and projection matrices and attach a randomly initialised linear layer atop the encoder to map the pooled output sequence into logits. Specifically, we exploit the Pytorch[59] implementation provided by Hugging Face and modify its source code to suit our needs. All models are trained on a node equipped with NVIDIA A100 GPUs.

Unsupervised label discovery

We pre-train HuBERT-ECG BASE for two consecutive iterations. To generate target labels for the first one, we perform a k-means clustering[60] with 100 clusters on feature descriptors consisting of 39-dimensional vectors of Mel Frequency Cepstral Coefficients[61] (MFCCs) (13 coefficients with first and second-order derivatives). Such features have already been successfully used in ECG analysis[62–64], as we confirm in Supplementary Information – Sec. 1.1 where we compare various feature descriptors. Additionally, although we do not see any significant benefit from using a cluster ensemble (Supplementary Information – Sec. 2.2), when we situate HuBERT-ECG in a multitask framework we add the labels generated by two additional MFCC-based k-means models with 200 and 300 clusters, respectively.

For the second iteration, to produce better and finer labels, we run the k-means algorithm again, increasing the number of clusters. We use 500 clusters of latent representations extracted from the 8th encoding layer of HuBERT-ECG BASE after the first iteration. At the end of the second iteration, we pre-train the SMALL and LARGE model configurations for one iteration using labels generated by clustering representations extracted from the 9th encoding layer into 500 and 1000 clusters, respectively (Supplementary Information – Sec. 1.4). Due to memory constraints, we cannot load the entire dataset into memory and, therefore, we opt for a batched version of k-means provided by scikit-learn[65], in which we yield batches of MFCC descriptors, or latent representations, and perform

incremental updates of centroid positions. We set a batch size of 9300, obtained by yielding 93 descriptors from 100 ECGs, and use k-means++[66] with 20 random restarts for a better initialisation.

Self-supervised pre-training

For each pre-training iteration we reserve 90% of the unlabelled dataset as the training set and the remaining 10% as the internal validation set. We set a batch size of 448 instances and an optimal masking percentage *p* of 33%. The first iteration consists of 80k steps, while the second iteration counts 770k steps. The third iteration, instead, consists of 362.5k and 422.5k steps for the SMALL and LARGE model sizes, respectively. We use Adam[67] optimiser with $\beta = (0.9,0.98)$, an initial weight decay[68] of 0.01 and a dropout probability[69] of 0.1, and a learning rate scheduler that ramps up for the first 8% of the training steps and then decays linearly to zero. Peak learning rates are 5e-5/5e-5/2.5e-5 for the BASE, SMALL and LARGE model configurations, respectively. In addition, we find benefits from exploiting a *dynamic regularisation* during pre-training (Supplementary Information – Sec. 1.3). This technique "penalises" the model by increasing its dropout probability and weight decay if it does not improve its performance on the internal validation set for *penalty-count* consecutive times. Otherwise, if it improves its performance on the internal validation set, with respect to the best validation loss or the best validation accuracy, the model is "rewarded" by reversing the effects of the last penalty, i.e., decreasing its dropout probability and weight decay. It is important to note that the initial weight decay and dropout probability are the minimum achievable values, while *penalty-count* emerges as a significant hyperparameter to be tuned according to the frequency with which validations are performed. During the first and third iterations, with randomly initialised models, we set the *penalty-count* to 4 and perform an internal validation every 2500 steps. For the second iteration, the internal validation is performed every 5000 steps, while keeping the same value of *penalty-count*.

Supervised fine-tuning

In order to assess the capabilities of HuBERT-ECG on clinically relevant datasets simulating real-world scenarios, we fine-tune it on every labelled dataset we consider, each one being characterised by a possibly very different number of instances (Fig. 1a), patients' age distribution (Fig. 1d), and possible conditions, each of which belonging to one of the three classes mentioned in Fig. 1f. When training-validation-test splits are predefined, or when at least the test set is known and fixed, as in the case of PTB-XL, SPH or Ribeiro, we perform the same split in order to allow fair comparisons with previous works. However, for datasets where no such split is known or applicable (i.e., Hefei, Ningbo, Chapman, CPSC, Georgia), we first extract a fixed hold-out test set in a stratified fashion containing 10% of the dataset instances, ensuring that all the dataset classes are represented. We then perform a 4-fold cross-validation to tune the hyperparameters of the four models on the remaining instances, selecting the best candidate from each fold for the evaluation on the test set. Finally, we average the four sets of results obtained from inference. When the cardinality of the dataset is extremely low compared to the model size, as in the case of PTB, CPSC-Extra and SaMi-Trop, we split the dataset four times in a stratified manner into four <*training*, *test*> folds. Then, for each fold, we skip any hyperparameter tuning, train four models until we reach near-zero training error and run inference directly on the corresponding test set with the last model checkpoint. During each fine-tuning, we optimise both the loss function and a macro-averaged AUROC on the validation set, selecting for inference at test time the model candidate that achieves the highest AUROC. Lastly, to generate Cardio-Learning, we merge the training, validation, and test sets from each dataset to form the corresponding overall training, validation, and test sets, taking care to ensure no data leakage between them. Notably, before any fine-tuning, we analyse the label distribution of each dataset and drop the instances labelled with conditions that occur only once. We do so because these conditions are either unlearnable or untestable, as their single instances cannot be included in both the training and test sets. However, if a condition occurs twice in the dataset, we assign one instance to the training set

and one to the test set, allowing the model to attempt to learn the condition while providing a way to assess its generalisability within the limits of this setup. We consider the performance on the validation set as a lower bound on the true test performance. Eventually, when a condition occurs three times, we place one example per split.

To fine-tune HuBERT-ECG we follow a simple protocol: we attach atop the pre-trained model a randomly initialized linear layer and fine-tune all the weights of the resulting model, except those of the convolutional embedder. In contrast to Chen et al.[70] and Hsu et al.[24], we do not use the *freezing-steps* hyperparameter, as we did not see any efficacy in keeping the encoder's parameters fixed while training only the last linear layer. Inspired by Devlin et al.18, we reduce the batch size to 64 instances and decrease the learning rate to 1e-5. Also, to gain more control over the search for a good candidate, we validate our models every 50 or 500 steps, depending on the dataset size, hence more frequently than what we do during the pre-training. Due to the high number of trainable parameters and the limited number of instances of most datasets, HuBERT-ECG overfits easily and we find no effectiveness in using either a strong dropout (up to 0.5) or a high weight decay (up to 0.1), nor in some freezing encoding layers. However, experiments with LayerDrop[72] ([0.0, 0.1, 0.15, 0.2]) show that it can help contain the validation loss divergence and metrics degradation. In addition, we see improvements when using a time-aligned random crop as data augmentation, a strategy that we also replicate at test time when we take the most confident prediction among those made on multiple crops of the same instance. In summary, we observe the best performance when fine-tuning the entire Transformer encoder, zeroing the dropout probability, keeping the weight decay at 0.01, and sweeping over the LayerDrop probability. We track experiments using Weights and Biases[e].

Data availability

All datasets supporting the findings described in this manuscript are public, except for Ribeiro. This dataset, the test set of which is publicly available, is accessible for scientific research upon request to the respective owner.

Code availability

The full pipeline utilised in this study is available at https://github.com/Edoar-do/HuBERT-ECG. This includes: (1) code for data preprocessing, starting from raw data to creating the train-validation-test splits used in our research; (2) scripts for replicating the training and inference of every model developed across all datasets; and (3) code for reproducing our performance validation. Eventually, to both facilitate reproducibility and enable rapid implementation, we make available on Hugging Face[f] (4) the pre-trained weights for all model configurations, as well as (5) the fine-tuned weights on the Cardio-Learning dataset.

References

1. Smilowitz, N. R. & Berger, J. S. Perioperative Cardiovascular Risk Assessment and Management for Noncardiac Surgery. *JAMA* **324**, 279 (2020).
2. van de Vegte, Y. J. *et al.* Genetic insights into resting heart rate and its role in cardiovascular disease. *Nat Commun* **14**, 4646 (2023).
3. Muzammil, M. A. *et al.* Artificial intelligence-enhanced electrocardiography for accurate diagnosis and management of cardiovascular diseases. *J Electrocardiol* **83**, 30–40 (2024).

[e] https://wandb.ai/site
[f] https://huggingface.co/Edoardo-BS/HuBERT-ECG/

4. Friedman, P. A. The Electrocardiogram at 100 Years: History and Future. *Circulation* **149**, 411–413 (2024).
5. Ahsan, M. M. & Siddique, Z. Machine learning-based heart disease diagnosis: A systematic literature review. *Artif Intell Med* **128**, 102289 (2022).
6. Kalmady, S. V. *et al.* Development and validation of machine learning algorithms based on electrocardiograms for cardiovascular diagnoses at the population level. *NPJ Digit Med* **7**, 133 (2024).
7. Goto, S. *et al.* Multinational Federated Learning Approach to Train ECG and Echocardiogram Models for Hypertrophic Cardiomyopathy Detection. *Circulation* **146**, 755–769 (2022).
8. Kim, M. *et al.* Deep learning for predicting rehospitalization in acute heart failure: Model foundation and external validation. *ESC Heart Fail* (2024) doi:10.1002/ehf2.14918.
9. Musa, N. *et al.* A systematic review and Meta-data analysis on the applications of Deep Learning in Electrocardiogram. *J Ambient Intell Humaniz Comput* **14**, 9677–9750 (2023).
10. Moor, M. *et al.* Foundation models for generalist medical artificial intelligence. *Nature* **616**, 259–265 (2023).
11. Finlayson, S. G. *et al.* The Clinician and Dataset Shift in Artificial Intelligence. *New England Journal of Medicine* **385**, 283–286 (2021).
12. Howell, M. D., Corrado, G. S. & DeSalvo, K. B. Three Epochs of Artificial Intelligence in Health Care. *JAMA* **331**, 242 (2024).
13. Bommasani, R. *et al.* On the Opportunities and Risks of Foundation Models. (2021).
14. Radford, A., Narasimhan, K., Salimans, T. & Sutskever, I. Improving Language Understanding by Generative Pre-Training. (2018).
15. Radford, A. *et al. Language Models Are Unsupervised Multitask Learners*. https://www.semanticscholar.org/paper/Language-Models-are-Unsupervised-Multitask-Learners-Radford-Wu/9405cc0d6169988371b2755e573cc28650d14dfe (2018).
16. Brown, T. B. *et al.* Language Models are Few-Shot Learners. (2020).
17. OpenAI *et al.* GPT-4 Technical Report. (2023).
18. Devlin, J., Chang, M.-W., Lee, K. & Toutanova, K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. (2018).
19. Liu, Y. *et al.* RoBERTa: A Robustly Optimized BERT Pretraining Approach. (2019).
20. Raffel, C. *et al.* Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. (2019).
21. Caron, M. *et al.* Emerging Properties in Self-Supervised Vision Transformers. (2021).
22. He, K. *et al.* Masked Autoencoders Are Scalable Vision Learners. (2021).
23. Baevski, A., Zhou, H., Mohamed, A. & Auli, M. wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations. (2020).
24. Hsu, W.-N. *et al.* HuBERT: Self-Supervised Speech Representation Learning by Masked Prediction of Hidden Units. (2021).
25. Schäfer, R. *et al.* Overcoming data scarcity in biomedical imaging with a foundational multi-task model. *Nat Comput Sci* **4**, 495–509 (2024).
26. Lu, M. Y. *et al.* A visual-language foundation model for computational pathology. *Nat Med* **30**, 863–874 (2024).
27. Xu, H. *et al.* A whole-slide foundation model for digital pathology from real-world data. *Nature* **630**, 181–188 (2024).
28. Vorontsov, E. *et al.* A foundation model for clinical-grade computational pathology and rare cancers detection. *Nat Med* **30**, 2924–2935 (2024).
29. Zhang, X., Wu, C., Zhang, Y., Xie, W. & Wang, Y. Knowledge-enhanced visual-language pre-training on chest radiology images. *Nat Commun* **14**, 4542 (2023).
30. Tiu, E. *et al.* Expert-level detection of pathologies from unannotated chest X-ray images via self-supervised learning. *Nat Biomed Eng* **6**, 1399–1406 (2022).

31. Ma, J. *et al.* Segment anything in medical images. *Nat Commun* **15**, 654 (2024).
32. Christensen, M., Vukadinovic, M., Yuan, N. & Ouyang, D. Vision–language foundation model for echocardiogram interpretation. *Nat Med* **30**, 1481–1488 (2024).
33. Zhang, K. *et al.* A generalist vision–language foundation model for diverse biomedical tasks. *Nat Med* (2024) doi:10.1038/s41591-024-03185-2.
34. Xu, M. *et al.* HMM-Based Audio Keyword Generation. in 566–574 (2004). doi:10.1007/978-3-540-30543-9_71.
35. Ribeiro, A. H. *et al.* Automatic diagnosis of the 12-lead ECG using a deep neural network. *Nat Commun* **11**, 1760 (2020).
36. Liu, F. *et al.* An Open Access Database for Evaluating the Algorithms of Electrocardiogram Rhythm and Morphology Abnormality Detection. *J Med Imaging Health Inform* **8**, 1368–1373 (2018).
37. Bousseljot, R., Kreiseler, D. & Schnabel, A. Nutzung der EKG-Signaldatenbank CARDIODAT der PTB über das Internet. *Biomedizinische Technik/Biomedical Engineering* 317–318 (2009) doi:10.1515/bmte.1995.40.s1.317.
38. Wagner, P. *et al.* PTB-XL, a large publicly available electrocardiography dataset. *Sci Data* **7**, 154 (2020).
39. Goldberger, A. L. *et al.* PhysioBank, PhysioToolkit, and PhysioNet: Components of a New Research Resource for Complex Physiologic Signals. *Circulation* **101**, (2000).
40. Zheng, J. *et al.* A 12-lead electrocardiogram database for arrhythmia research covering more than 10,000 patients. *Sci Data* **7**, 48 (2020).
41. Zheng, J. *et al.* Optimal Multi-Stage Arrhythmia Classification Approach. *Sci Rep* **10**, 2898 (2020).
42. Liu, H. *et al.* A large-scale multi-label 12-lead electrocardiogram database with standardized diagnostic statements. *Sci Data* **9**, 272 (2022).
43. Cardoso, C. S. *et al.* Longitudinal study of patients with chronic Chagas cardiomyopathy in Brazil (SaMi-Trop project): a cohort profile. *BMJ Open* **6**, e011181 (2016).
44. Na, Y., Park, M., Tae, Y. & Joo, S. Guiding Masked Representation Learning to Capture Spatio-Temporal Relationship of Electrocardiogram. (2024).
45. Ferreira, A. M. *et al.* Two-year death prediction models among patients with Chagas Disease using machine learning-based methods. *PLoS Negl Trop Dis* **16**, e0010356 (2022).
46. Wouters, P. C. *et al.* Electrocardiogram-based deep learning improves outcome prediction following cardiac resynchronization therapy. *Eur Heart J* **44**, 680–692 (2023).
47. Leclercq, C. *et al.* Wearables, telemedicine, and artificial intelligence in arrhythmias and heart failure: Proceedings of the European Society of Cardiology Cardiovascular Round Table. *EP Europace* **24**, 1372–1383 (2022).
48. Clifford, G. *et al.* AF Classification from a Short Single Lead ECG Recording: the Physionet Computing in Cardiology Challenge 2017. in (2017). doi:10.22489/CinC.2017.065-469.
49. Perez Alday, E. A. *et al.* Classification of 12-lead ECGs: the PhysioNet/Computing in Cardiology Challenge 2020. *Physiol Meas* **41**, 124003 (2020).
50. Reyna, M. A. *et al.* Will Two Do? Varying Dimensions in Electrocardiography: The PhysioNet/Computing in Cardiology Challenge 2021. in *2021 Computing in Cardiology (CinC)* 1–4 (IEEE, 2021). doi:10.23919/CinC53138.2021.9662687.
51. Gow, B. *et al.* MIMIC-IV-ECG: Diagnostic Electrocardiogram Matched Subset. (2023).
52. McInnes, L., Healy, J. & Melville, J. UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. (2018).
53. Natarajan, A. *et al.* A Wide and Deep Transformer Neural Network for 12-Lead ECG Classification. in (2020). doi:10.22489/CinC.2020.107.

54. Hong, S., Zhang, W., Sun, C., Zhou, Y. & Li, H. Practical Lessons on 12-Lead ECG Classification: Meta-Analysis of Methods From PhysioNet/Computing in Cardiology Challenge 2020. *Front Physiol* **12**, (2022).
55. Mehari, T. & Strodthoff, N. Towards Quantitative Precision for ECG Analysis: Leveraging State Space Models, Self-Supervision and Patient Metadata. *IEEE J Biomed Health Inform* **27**, 5326–5334 (2023).
56. Choi, S. *et al.* ECGBERT: Understanding Hidden Language of ECGs with Self-Supervised Representation Learning. (2023).
57. Vaswani, A. *et al.* Attention Is All You Need. (2017).
58. Hu, R., Chen, J. & Zhou, L. Spatiotemporal self-supervised representation learning from multi-lead ECG signals. *Biomed Signal Process Control* **84**, 104772 (2023).
59. Paszke, A. *et al.* PyTorch: An Imperative Style, High-Performance Deep Learning Library. (2019).
60. LLoyd, S. P. Least Squares Quantization in PCM. *IEEE Trans Inf Theory* **28**, (1982).
61. Xu, M. *et al.* HMM-based audio keyword generation. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* **3333**, (2004).
62. Boussaa, M., Atouf, I., Atibi, M. & Bennis, A. ECG signals classification using MFCC coefficients and ANN classifier. in *2016 International Conference on Electrical and Information Technologies (ICEIT)* 480–484 (IEEE, 2016). doi:10.1109/EITech.2016.7519646.
63. Arpitha, Y., Madhumathi, G. L. & Balaji, N. Spectrogram analysis of ECG signal and classification efficiency using MFCC feature extraction technique. *J Ambient Intell Humaniz Comput* **13**, 757–767 (2022).
64. Singh, A. K. & Krishnan, S. ECG signal feature extraction trends in methods and applications. *Biomed Eng Online* **22**, 22 (2023).
65. Pedregosa, F. *et al.* Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research* **12**, 2825–2830 (2011).
66. Arthur, D. & Vassilvitskii, S. k-means++: the advantages of careful seeding. in *Proceedings of the eighteenth annual ACM-SIAM symposium on Discrete algorithms* 1027–1035 (ACM-SIAM Symposium on DIscrete Algorithms, 2007).
67. Kingma, D. P. & Ba, J. Adam: A Method for Stochastic Optimization. (2014).
68. Krogh, A. & Hertz, J. A. A Simple Weight Decay Can Improve Generalization. *Adv Neural Inf Process Syst* **4**, (1991).
69. Hinton, G. E., Srivastava, N., Krizhevsky, A., Sutskever, I. & Salakhutdinov, R. R. Improving neural networks by preventing co-adaptation of feature detectors. (2012).
70. Chen, W. *et al.* Reducing Barriers to Self-Supervised Learning: HuBERT Pre-training with Academic Compute. (2023).
71. Devlin, J., Chang, M.-W., Lee, K., Google, K. T. & Language, A. I. *BERT: Pre-Training of Deep Bidirectional Transformers for Language Understanding*. https://github.com/tensorflow/tensor2tensor.
72. Fan, A., Grave, E. & Joulin, A. Reducing Transformer Depth on Demand with Structured Dropout. (2019).
73. Mehari, T. & Strodthoff, N. Self-supervised representation learning from 12-lead ECG data. *Comput Biol Med* **141**, 105114 (2022).
74. Strodthoff, N., Wagner, P., Schaeffter, T. & Samek, W. Deep Learning for ECG Analysis: Benchmarks and Insights from PTB-XL. *IEEE J Biomed Health Inform* **25**, 1519–1528 (2021).

Acknowledgements

This work has been partly funded by 1) Regione Lombardia, Italy, through the initiative "Programme of measures for economic recovery: development of new cooperation agreements with universities for research, innovation and technology transfer" - DGR n. XI/4445/2021; 2) European Union - Next Generation EU, through the Italian Ministry of Research PRIN 2022, project n. 2022A49KR3 "QT-SEED Quality-of-life Technological and Societal Exploitation of ECG Diagnostics".

Competing Interest

The authors declare no competing non-financial interests.

		SMALL	BASE	LARGE
Convolutional embedder	strides	4, 2, 2, 2, 2		
	kernels width	10, 3, 3, 2, 2		
	channels	512		
Transformer encoder	layers	8	12	16
	internal dimension	512	768	960
	feed-forward dimension	2048	3072	3840
	attention heads	8	8	12
Embedding dimension W		256	256	512
Number of parameters		30M	93M	188M

Table 1 | HuBERT-ECG architecture summary.

Condition	Models	Sensitivity	Specificity	AUROC	AUPRC
Atrial Fibrillation	Ribeiro et al.	0.769	1.000	0.885	0.773
	HuBERT-ECG SMALL	1.000	0.977	1.000	0.9774
	HuBERT-ECG BASE	1.000	1.000	**1.000**	**1.000**
	HuBERT-ECG LARGE	1.000	0.975	0.998	0.924
Atrio-ventricular Block type I	Ribeiro et al.	0.929	0.995	0.962	0.807
	HuBERT-ECG SMALL	0.929	0.952	0.989	0.8309
	HuBERT-ECG BASE	0.750	0.999	**0.998**	**0.956**
	HuBERT-ECG LARGE	1.000	0.949	0.996	0.875
Left bundle branch block	Ribeiro et al.	1.000	1.000	1.000	1.000
	HuBERT-ECG SMALL	1.000	0.994	1.000	0.995
	HuBERT-ECG BASE	0.900	1.000	**1.000**	**1.000**
	HuBERT-ECG LARGE	0.933	0.995	0.999	0.976
Right bundle branch block	Ribeiro et al.	1.000	0.995	0.997	0.895
	HuBERT-ECG SMALL	1.000	0.985	0.999	0.980
	HuBERT-ECG BASE	0.940	0.995	**0.999**	**0.974**
	HuBERT-ECG LARGE	0.971	0.987	0.998	0.969
Sinus bradycardia	Ribeiro et al.	0.938	0.996	0.967	0.782
	HuBERT-ECG SMALL	1.000	0.981	0.998	0.880
	HuBERT-ECG BASE	1.000	0.993	**0.998**	**0.886**
	HuBERT-ECG LARGE	1.000	0.980	0.998	0.895
Sinus Tachycardia	Ribeiro et al.	0.937	0.997	0.985	0.923
	HuBERT-ECG SMALL	0.973	0.982	0.997	0.935
	HuBERT-ECG BASE	0.973	0.995	**0.999**	**0.964**
	HuBERT-ECG LARGE	0.973	0.984	0.996	0.948
MACRO-AVERAGED	Ribeiro et al.	0.935	0.997	0.966	0.863
	HuBERT-ECG SMALL	0.984	0.979	0.997	0.933
	HuBERT-ECG BASE	0.927	0.999	**0.999**	**0.963**
	HuBERT-ECG LARGE	0.979	0.978	0.997	0.931

Table 2 | Fine-tuned HuBERT-ECG performance on Ribeiro benchmarked against that from Ribeiro et al.[35] according to multiple metrics.

Models	PTB-XL All	PTB-XL Form	PTB-XL Rhythm	PTB-XL Diag.	PTB-XL Diag. Subclass	PTB-XL Diag. Superclass
	Macro-averaged AUROC					
HuBERT-ECG SMALL	0.900	0.838	0.941	0.919	0.913	0.907
HuBERT-ECG BASE	0.902	0.855	0.953	0.917	0.917	0.911
HuBERT-ECG LARGE	0.896	0.828	0.935	0.905	0.919	0.903
Hu et al.[58(*)]	**0.947**	~ 0.895	**~ 0.980**	**~ 0.950**	**~ 0.940**	**~ 0.938**
Mehari & Strodthoff [73(**)]	0.942	N.A.	N.A.	N.A.	N.A.	N.A.
Strodthoff et al.[74(***)]	0.925	**0.896**	0.957	0.937	0.929	0.928
Na et al.[44(***)]	N.A.	N.A.	N.A.	N.A.	N.A.	0.933

Table 3 | Fine-tuned HuBERT-ECG performance on PTB-XL datasets against the state-of-the-art. "N.A." stands for "Not Available". (*) Code not available. Results preceded by '~' symbol are estimated by looking at paper graphs. (**) Based on the available code, the AUROC is calculated on individual batches and then averaged—introducing batch-size dependency—with custom handling for any NaN values that arise due to class imbalances within batches. (***) Based on the available code, the AUROC is computed after aggregating predictions and corresponding ground-truth labels across all batches—a standard approach we also adopt.

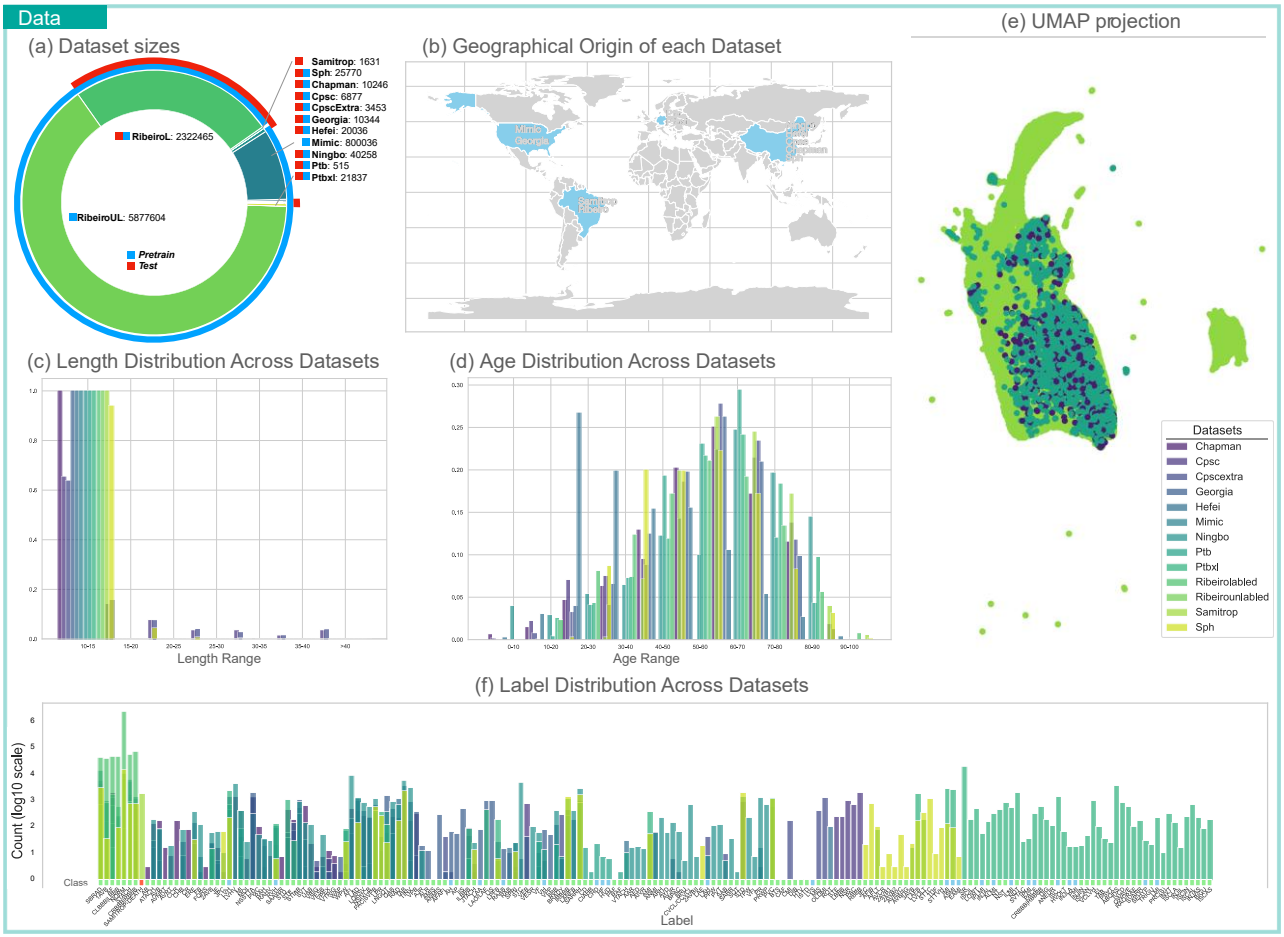


Fig 1. (a) Used dataset and their sizes. The blue bar highlights the dataset used for the pre-training while the red one is the testing one. (b) Geographical origin of the considered datasets. (c) ECG length in seconds. (d) Age distribution. (e) UMAP projection (uniform sample of 30% of all the dataset for the sake of visualisation). (f) Label distribution and grouping according to 3 classes (see the coloured bar below the distribution): Class 0: Green colour for "The ECG is the primary diagnostic tool"; Class 1: Blue colour for "The ECG is a supportive, not primary, diagnostic tool"; Class 2: Red colour for "ECG is used to predict future CVEs". Label abbreviations and corresponding diagnosis are reported in Supplementary Table 1.

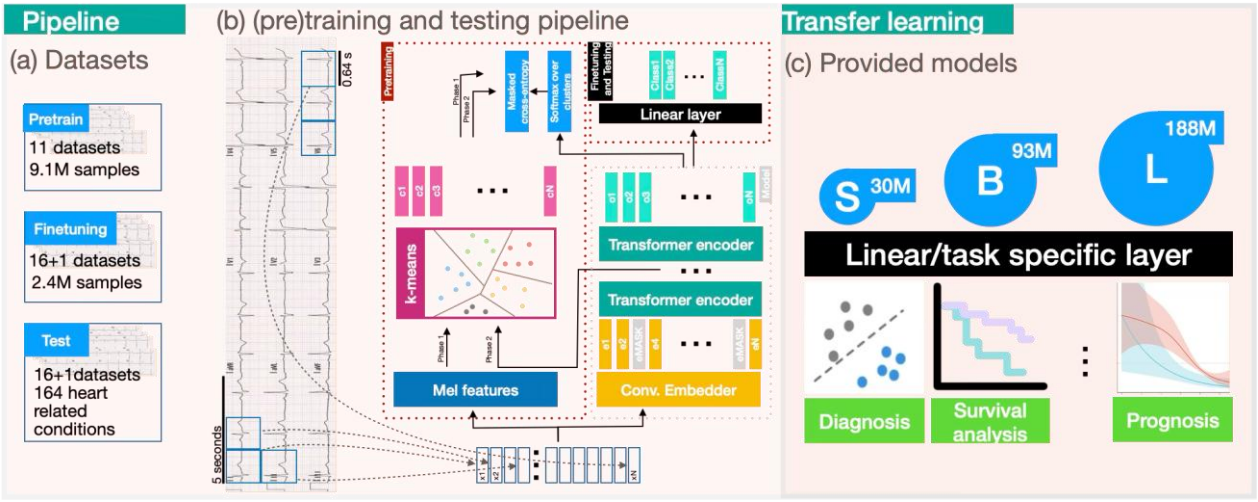


Fig 2. An overview of the proposed deep learning pipeline. (a) **Datasets**: The model is pre-trained on 11 datasets comprising over 9.1 million samples. Then, the model is fine-tuned and tested on 16 datasets and their aggregation covering 164 heart-related conditions. (b) **Pre-training and Testing Pipeline**: The pipeline begins with Mel feature extraction, followed by k-means clustering, and includes two training phases with masked cross-entropy loss and softmax classification. A convolutional embedder and a Transformer encoder process the data, culminating in a linear output layer for classification. (c) **Transfer Learning and Provided Models**: Three model variants (**S**MALL, **B**ASE, **L**ARGE) with increasing parameter counts (30M, 93M, and 188M) are pre-trained, tested and, eventually, provided to enable transfer learning for many downstream tasks such as diagnosis, survival analysis, and prognosis. Fine-tuning can be performed by simply adding a task-specific linear layer.

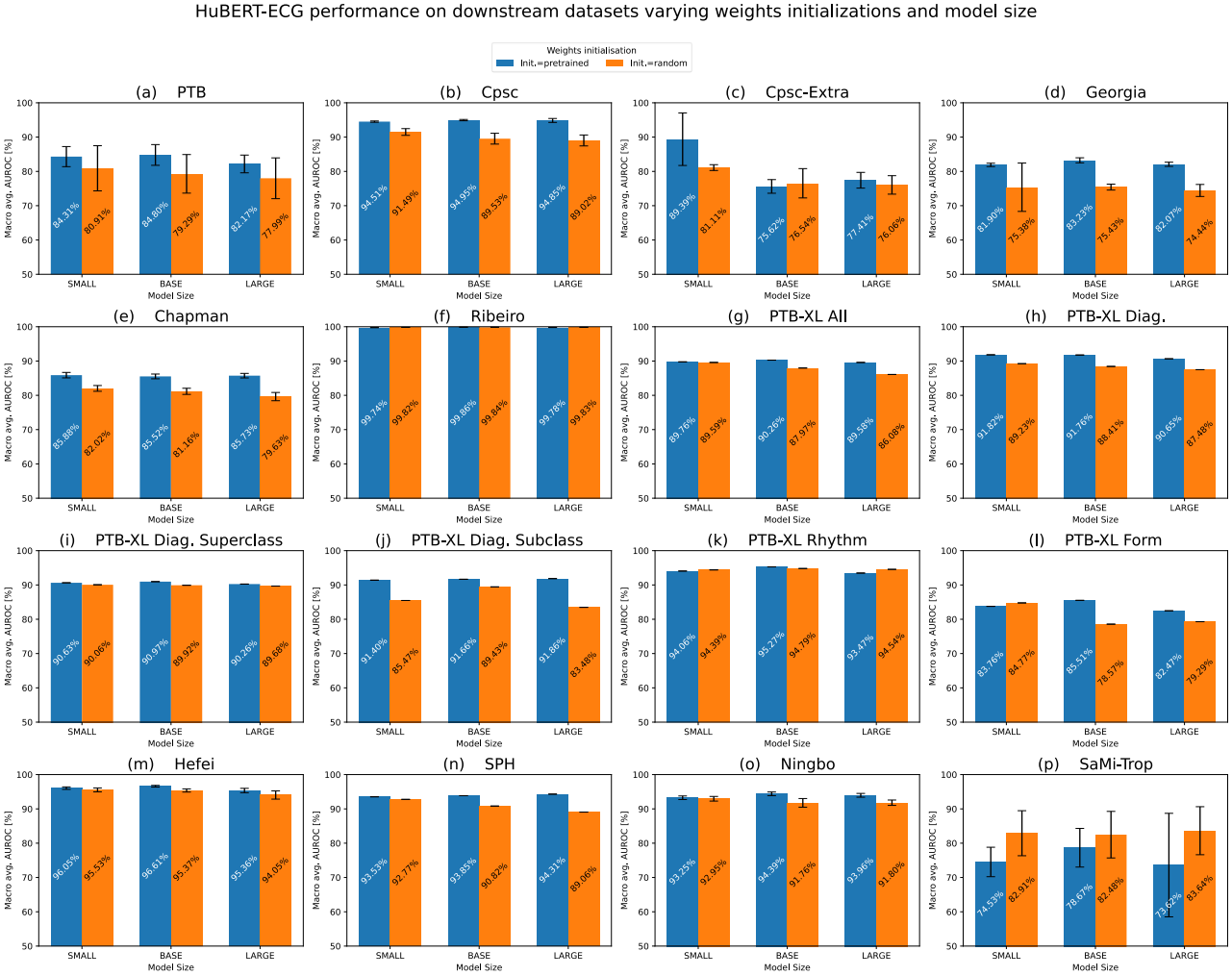


Fig. 3. HuBERT-ECG performance across 16 downstream datasets varying weights initialization and model size. Performance assessed through 4-fold cross-validation are reported with standard deviations.

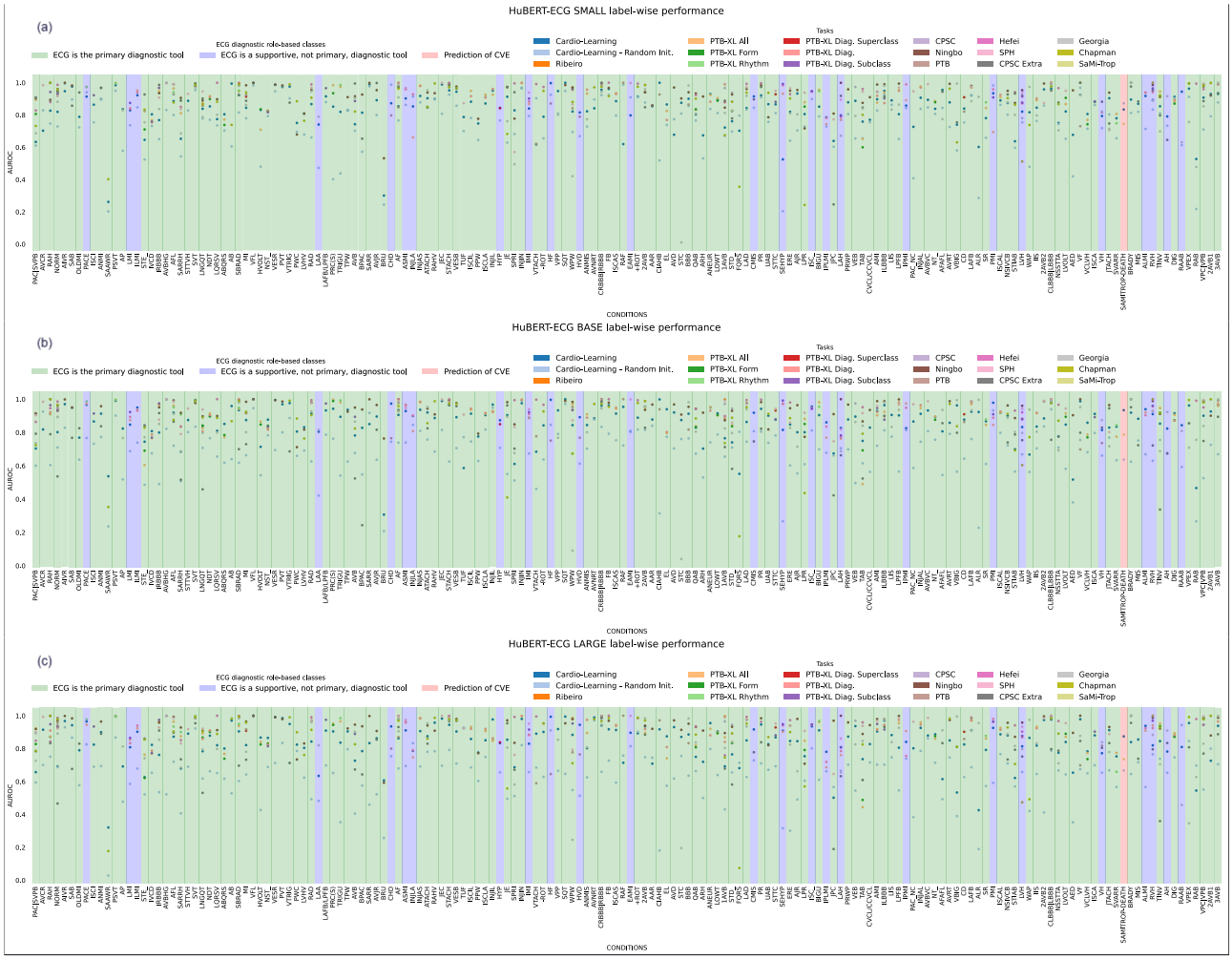


Fig. 4. Hubert-ECG (a) SMALL, (b) BASE, (c) LARGE label-wise performance on every dataset, including Cardio-Learning. Label backgrounds are coloured based on the diagnostic role of the ECG. Supplementary Table 1 reports label abbreviations and corresponding diagnosis.

Supplementary Information

1. Pre-training analyses

This section presents an analysis of the impact of pre-training with different ECG feature descriptors and sampling rates on downstream performance. We then investigate the interesting effects that follow the application of the dynamic regularisation during pre-training. Finally, we examine the clustering quality across both HuBERT-ECG encoding layers and pre-training iterations to measure the downstream impact of these factors. For these purposes, we pre-train HuBERT-ECG BASE for one iteration with fixed hyperparameters and running configurations described in each of the following sections. Then, to evaluate the impact of different choices, we simply perform linear evaluations, i.e., training of a randomly initialised linear layer atop a frozen backbone, on a development set that we extract from the dataset used by Ribeiro et al.[1]. We build such a set, hereafter referred to as *Ribeiro-dev*, by deliberately excluding all normal ECGs, as we are more interested in assessing the ability to detect cardiac abnormalities rather than normal ECGs. We avoid a full fine-tuning due to limited computational resources, but we are confident that the linear evaluation results serve as lower bound since they are obtained considering fewer trainable parameters and using a development set that is three times larger than all public ECG sources combined.

1.1 Exploring feature descriptors for ECG fragments

In order to identify the most appropriate ECG fragment descriptor to use in the initial pre-training iteration, we perform k-means clustering on the entire training set with 3 different feature descriptors extracted from the ECG fragments. The first descriptor takes into account 16 simple time-frequency features, the second one, is based on 39 Mel Frequency Cepstral coefficients[3] (MFCCs) following Hsu et al.[2], and the third one combines the first 13 MFCCs and all the time-frequency features. For each descriptor, we run k-means with 10, 30, 50, 100, 150, 200 and 300 clusters and compute the corresponding sum of squared errors, or inertia (Supplementary Fig. 1). We observe that MFCC- and time-frequency based descriptors provide comparable inertia and a global mean squared error in the order of 1e-5, and thus appear as promising candidates. Following the elbow method[4], and driven by the intuition that more clusters capture too fine and specific ECG patterns for the first pre-training iteration, we set C = 100 as the best number of clusters to use irrespective of the descriptor. Given this number of clusters, we compute the Davies-Bouldin[5] index to have an additional selection criterion and obtain extremely low scores. Interestingly, while clustering on MFCCs yields higher inertia than clustering on time-frequency features, clusters on the former are slightly more compact, as suggested by a marginally lower Davies-Bouldin score. Therefore, we experiment with the use of both MFCC- and time-frequency based k-means models during pre-training and, then, linearly evaluate the resulting models on *Ribeiro-dev* until performance plateau or a maximum number of 80k training steps is reached. The linear evaluation results, presented in Supplementary Table 1, show that pre-training with labels generated by an MFCC-based k-means model yields better downstream performance.

1.2 How the sampling rate affects downstream performance

The sampling rate of an ECG depends on the specific settings of the machine recording the electrical activity of the heart and there is no standard practice in this regard. For example, ECGs from the Ribeiro et al.[1] dataset are sampled at 400 Hz, while ECGs from PTB-XL[6] and SPH[7] are sampled at 400 and 500 Hz, respectively. We investigate the effect that the ECGs sampling rate has on downstream performance. This parameter is important as it regulates the dimensionality of the data, therefore the computation speed, and the degree of dilution of the information content. While it is true that training with ECGs sampled at lower frequencies is computationally less expensive, it is also true that sampling ECGs at lower frequencies may not provide enough samples to capture significant features and nuances in the input signal. Increasing the sampling rate may remedy this issue, at the cost of slower training and higher memory occupation. However, an oversampled ECG may contain redundant samples that overly dilute the information content and may be perceived as noise by a deep learning model. For these reasons, after band-pass filtering our ECGs to exclude frequencies outside [0.05, 47] Hz, which is reported to contain the dominant components of P waves, T waves and QRS complexes[8], we investigate the effects that sampling a 12-lead ECG at 50 and 100 Hz on linear evaluation performance over *Ribeiro-dev*. For these experiments, we linearly evaluate pre-trained models until the performance plateau or 80k training steps are performed in order to determine the optimal sampling rate to work with. To always have the same number of embeddings being processed by the Transformer encoder, we design waveform convolutional embedders that become increasingly shallow as the sampling rate decreases, while also maintaining the same number of ECG fragments. All the other design choices and running configurations are fixed across the experiments, as shown in Supplementary Table 2. Supplementary Fig. 2a illustrates the validation loss curves observed during pre-training at 50 and 100 Hz while Supplementary Fig. 2b shows the corresponding linear evaluation performance on *Ribeiro-dev*. Two discernible trends

are evident from the former: as the sampling rate decreases, the validation loss curves start at lower values than those corresponding to higher sampling rates; furthermore, as the sampling rate decreases, the models tend to show earlier overfitting to the training data. We suspect that at lower sampling rates, the information needed to solve the upstream task becomes more readily accessible, thereby facilitating the learning process. However, when clinical downstream labels are introduced, there is a noticeable performance discrepancy between models operating at 50 Hz and those operating at twice this sampling rate (Supplementary Fig. 2b and Supplementary Table 3). We attribute this discrepancy to the insufficient number of samples available to capture label-related patterns. In summary, sampling at 100 Hz captures all the meaningful physiological information, satisfies the Nyquist-Shannon theorem, and represents a desirable trade-off between computational cost and accurate downstream performance. All results are shown in Supplementary Table 3.

1.3 The effects of the dynamic regularisation

In order to gain a deeper insight into the impact of dynamic regularisation on training time and the time required to tune regularisation terms, including weight decay and dropout probabilities, we consider four different pre-training setups, each followed by a linear evaluation to assess the actual benefit on downstream performance. The first and the second one consist in performing the entire first pre-training iteration (260k steps with no early stopping) with a weight decay of 0.01 and a dropout probability of 0.1 (referred to as default), with and without dynamic regularisation, respectively, The last two setups, instead, consist in performing the same pre-training iteration, with and without dynamic regularisation, but setting the initial weight decay and dropout probability to the maximum values found by the dynamic regularisation. All experiments consider the BASE architecture, ECGs sampled at 100 Hz, a MFCC-based k-means as label generator (C = 100), a batch size of 448 instances and a masking percentage $p = 33\%$. The results of our experiments, presented in Supplementary Table 4 and labelled with letters for clarity, reveal interesting insights. When using dynamic regularisation with default weight decay and dropout probability (setup A), HuBERT-ECG converges in 80k steps, resulting in the best macro-averaged AUROC of 0.933 during linear evaluation. In contrast, disabling the dynamic regularisation while keeping the default regularisation terms (setup B), significantly slows down the convergence and leads to inferior downstream results. Instead, when comparing setups C and D, we observe two opposite behaviours: setup C shows faster convergence but worse upstream performance, while setup D shows slower convergence but better upstream performance. Despite these discordant trends, both setups produce close downstream results, both surpassing those from setup B, but still falling short of those of setup A. This suggests that pre-training with the dynamic regularisation, or the maximum regularisation terms it finds, speeds up pre-training and allows the model to adapt itself to avoid overfitting, resulting in improved linear evaluation when compared to scenarios where the dynamic regularisation is not used at all. Nevertheless, we hypothesise that initiating pre-training with already high regularising terms, irrespective of whether dynamic regularisation is employed or not (setups C and D), may impair the model's learning capability at the most crucial stage, resulting in inferior, albeit marginal, downstream performance. To investigate into the source of these benefits, it seems plausible to suggest that a dynamic dropout rate encourages the model to assign greater importance to alternative hidden paths when necessary, without permanently excluding the dropped ones. In such cases, when also the weight decay increases, this complementary form of regularisation prevents significant changes to the weights of those paths, thereby maintaining training stability.

1.4 The downstream impact of clustering quality across encoding layers and iterations

In the work of Hsu et al.[2], prior to the second and third pre-training iterations, the most suitable number of clusters was determined, as well as the layer from which latent features had to be extracted. This was achieved through the analysis of the quality of clusters of hidden representations extracted from each model layer. In particular, an automatic-speech-recognition model was used to produce frame-level phonetic labels that served as targets to measure the correctness of forced-aligned cluster assignments. The same approach is not applicable to our case as, to the best of our knowledge, there is no powerful open-source ECG model that can produce frame-level forced-aligned cluster assignments. Consequently, to determine which layer's latent features should be clustered and how many clusters are to be found, we make use once again of traditional compactness and separability metrics. In addition, in order to limit the set of layers to explore, we combine such metrics with findings from the NLP domain regarding layers transferability[9]. After sampling 10% of pre-training ECGs, we cluster their fragments' latent representations from the 5^{th} – 10^{th} Transformer layer and measure the clustering quality in terms of inertia, Davies-Bouldin index and Calinsky-Harabasz[10] index. We exclude from consideration shallower layers on the grounds that they would produce too coarse representations to be of use in clustering. Similarly, deeper layers are excluded on the grounds that they would generate representations that would lead to too task-specific labels for a generic pre-training based on pseudo-labels. Upon completion of the first iteration, we identify 500 and 1000 clusters of latent features from each of these layers, since we consider that setting a higher number of clusters than that used in the first iteration is necessary for two reasons: 1) to try to generate much finer cluster assignments; 2) to avoid measuring, through clustering, the degree of separability of the classes learnt during the previous iteration. The

results of this final analysis are presented in Supplementary Fig. 3. As can be seen, each metric shows a monotonic trend, regardless of the representations being clustered. Therefore, they do not provide an unequivocal indication of the best encoding layer to extract features from. To continue the pre-training of the BASE configuration, we then choose to cluster latent representations from the 8^{th} layer into 500 clusters, as this point seems to mark a non-negligible change in the metrics we are considering. After completing the second iteration, we perform a linear evaluation with the same running configurations reported in previous paragraphs to quantify the relative improvement with respect to previous evaluations. As shown in Supplementary Fig. 4, when the first-iteration model saturates, the second-iteration one achieves results that are approximately 5 AUROC points better and still has room for improvement.
Once HuBERT-ECG BASE completes the second iteration, we repeat the clustering step described above and report the results in Supplementary Fig. 3. Once again, the metrics we consider show a monotonic trend that does not facilitate the choice of an acceptable number of clusters (500 or 1000) nor which layer the latent representations should be extracted from. For these reasons, to start pre-training the SMALL and LARGE model configurations, we decide to extract latent features from the 9^{th} layer and to cluster them into 500 and 1000 clusters, respectively. We believe that pre-training these third-iteration models requires, for any configuration, selecting a deeper layer than those selected for the previous iterations. Furthermore, we believe that increasing the quantity and fineness of cluster assignments is beneficial for a more complex model such as HuBERT-ECG LARGE. Conversely, for the SMALL model size, maintaining the same number and granularity of cluster assignments can facilitate the task of mimicking the BASE model without significant loss of performance. The benefits of these choices are also displayed in Supplementary Fig. 4. Upon completion of 362.5k- and 422.5k-step pre-trainings, the SMALL and LARGE configurations perform similarly when linearly evaluated on *Ribeiro-dev* for 80k steps: both achieve a macro-averaged AUROC slightly lower than that of the BASE configuration, but show signs of saturation within this linear evaluation training time.

2 Ablation Study

In this section, we present a short sequence of ablation studies to investigate the effects of specific architectural choices and the impact of important hyperparameters. In particular, we study (1) the impact of our masking strategy, and (2) the effects of multi-task learning experimenting with multiple cluster ensembles. To do so we perform multiple pre-trainings of HuBER-ECG BASE (first iteration only) followed by linear evaluations on *Ribeiro-dev* with fixed hyperparameters. If not otherwise mentioned, the experimental configurations are the same of the previous section.

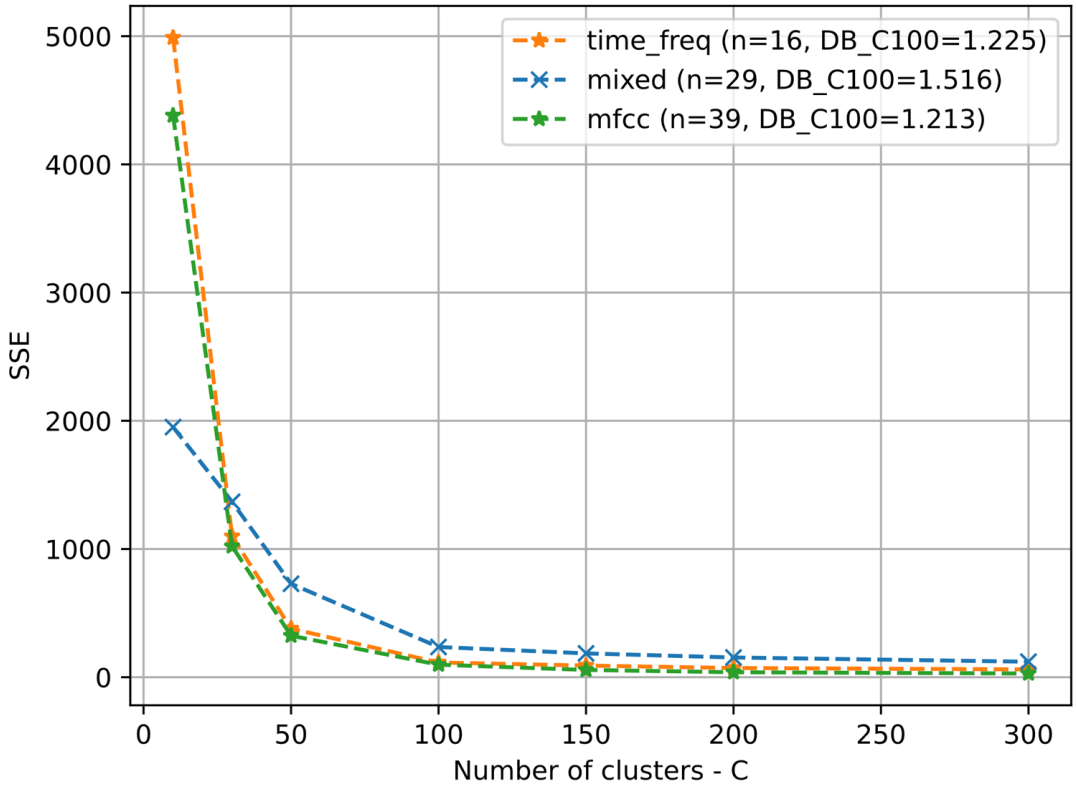
2.1 Impacts of the Masking Strategy

We consider setting the value of the masking percentage p of crucial importance to make HuBERT-ECG learn high quality representations of 12-lead ECGs. An excessively low value would generate a trivial upstream task, while an exaggeratedly high value would result into a nearly impossible one. For this reason, we experiment with multiple values of this hyperparameter in order to see how it impacts on downstream performance and plot the results of these experiments in Figure 4. For the sake of comparison, we also plot the exact percentage of embeddings that are masked when we use the masking strategy proposed by Hsu et al.[2]. Beyond some statistical and label noise, what emerges clearly is that setting $p = 33\%$ guarantees the best results, while following the masking strategy used to pre-train HuBERT leads to suboptimal performance. We believe this finding is due to the more regular patterns and high information redundancy of ECGs compared to audio signals.

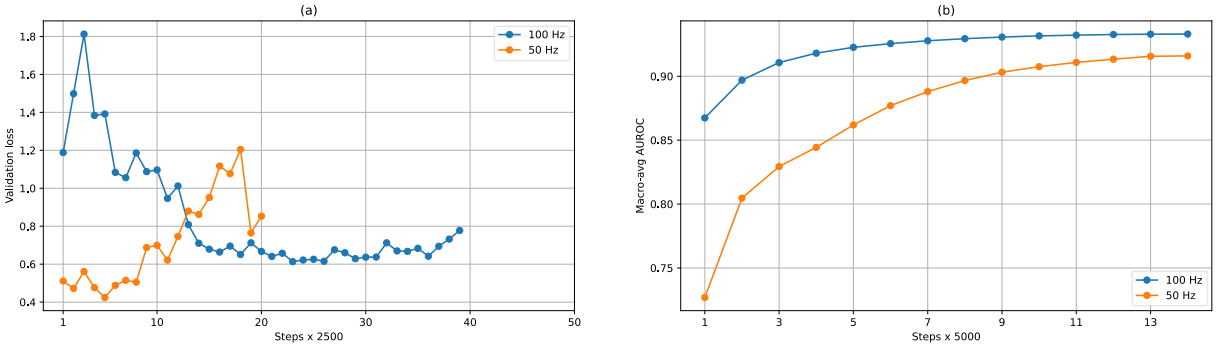
2.2 The Effects Of Multi-Task Learning

To observe the downstream effects of pre-training HuBERT-ECG in a multi-task learning framework, we perform three pre-trainings with an increasing number of tasks to solve. Initially, we pre-train HuBERT-ECG using labels generated by a single MFCC-based k-means model with 100 clusters. Then, we pre-train again with an additional clustering model with 200 clusters and, eventually, we consider an ensemble of three k-means models with 100, 200 and 300 clusters. Supplementary Table 5 reports the attained results. Interestingly, adding just a new clustering model does not change pre-training nor affects downstream performance. In contrast, although marginal, we see an improvement when considering an ensemble of three k-means models providing much more granular labels. Since solving three tasks at once is harder than solving just one of them, we are not surprised in seeing that such performance gains follow a much longer pre-training. We also believe that the cardinality of the ensemble is not as relevant as the maximum number of clusters, which we think is more useful to the model to capture label-specific patterns in downstream data. This is analogous to the refinement of the cluster assignments that is performed prior to the second iteration, except for how such refinement is achieved, since both aim to generate finer and more granular labels for the ECG fragments. Considering both the longer pre-training time and the marginal improvement obtained on the results reached after training with labels generated by a single clustering model, we do not proceed in pre-training with cluster ensembles.

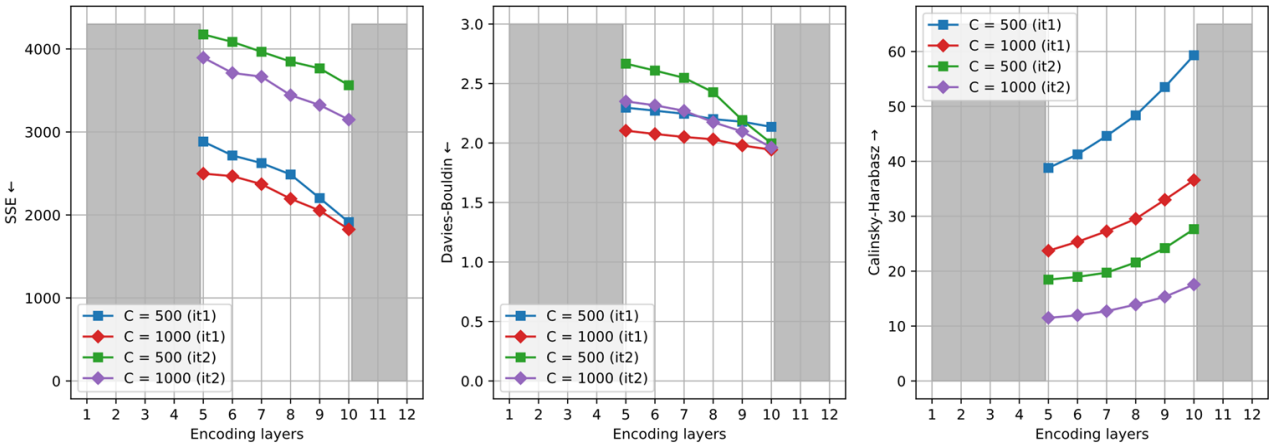
Supplementary Figures



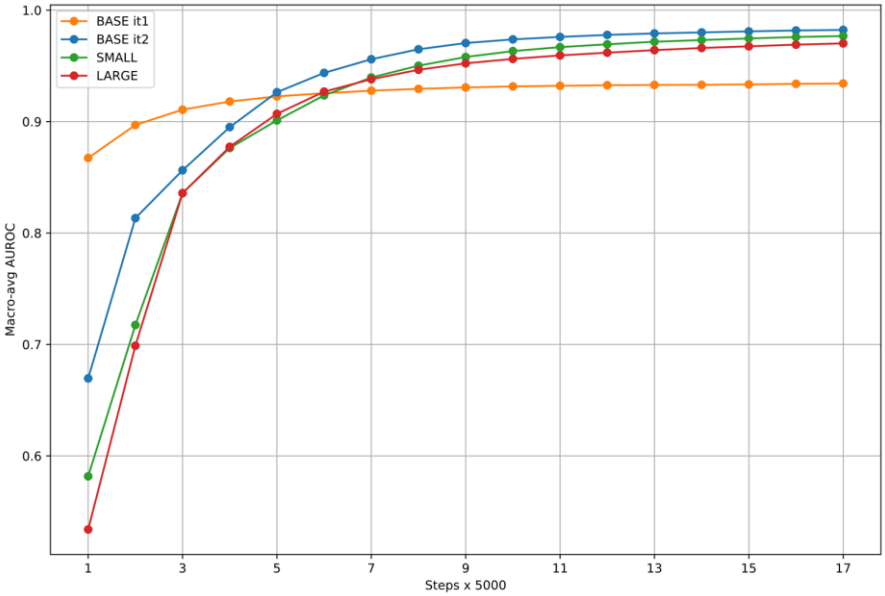
Supplementary Fig. 1 | Performance comparison of multiple k-means clustering runs in terms of Sum of Squared Errors (SSE). Once fixed the optimal number of clusters (C = 100), the Davies-Bouldin index (DB) is computed and reported as "DB_C100". Lower DB values indicate better clustering.



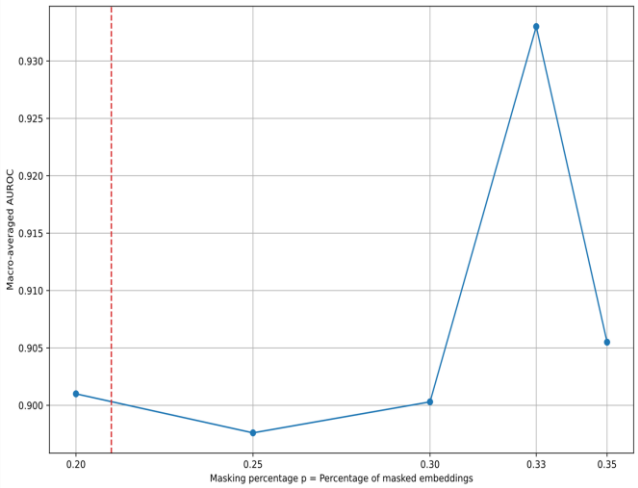
Supplementary Fig. 2 | (a) Validation loss curves obtained when pre-training HuBERT-ECG with ECGs sampled at 50 and 100 Hz. **(b)** Downstream performance obtained when HuBERT-ECG is linearly evaluated on *Ribeiro-dev* after being pre-trained with ECGs sampled at 50 and 100 Hz.



Supplementary Fig. 3 | Clustering quality metrics (inertia, Davies-Bouldin index, Calisky-Harabasz index) across HuBERT-ECG BASE encoding layers after the first (it1) and second (it2) pre-training iterations. Symbols ↑and ↓ indicate whether a metric needs to be maximised or minimised, respectively. Gray-shaded regions refer to encoding layers not considered in this analysis.



Supplementary Fig. 4 | Linear evaluation performance of HuBERT-ECG BASE, after first and second pre-training iteration, HuBERT-ECG SMALL and HuBERT-ECG LARGE on *Ribeiro-dev*.



Supplementary Fig. 5 | HuBERT-ECG BASE linear evaluation performance after pre-training with different values of the hyper-parameter *p* (i.e. the percentage of ECG embeddings to mask). The red dashed line indicates the percentage of ECG embeddings that would be masked if we followed the masking strategy used to pre-train HuBERT[3].

Supplementary Tables

Supplementary Table 1: Diagnoses and corresponding abbreviations used.

Diagnosis	Abbreviation
2nd Degree Av Block	2AVB
Second Av Block Mobitx Type I	2AVB1
Mobitz Type II Atrioventricular Block	2AVB2
Av Block Varying Conduction	AVBVC
Av Block Advanced (High Grade)	AVBHG
Av Block Complete (Third Degree)	3AVB
1st Degree Av Block	1AVB
Anterior Myocardial Infarction	ANMI
Extensive Anterior Myocardial Infarction	EAMI
St Elevation	STE_
Inferoposterolateral Myocardial Infarction	IPLMI
Posterior Myocardial Infarction	PMI
Inferolateral Myocardial Infarction	ILMI
Inferoposterior Myocardial Infarction	IPMI
Anterolateral Myocardial Infarction	ALMI
Lateral Myocardial Infarction	LMI
Acute Myocardial Infarction	AMI
Anteroseptal Myocardial Infarction	ASMI
Inferior Myocardial Infarction	IMI

Ischemic In Anterior Leads (Subclass)	ISCA
Anterior Ischemia	ANMIS
Inferior Ischaemia	IIS
Subendocardial Injury In Inferior Leads	INJIN
Ischemic In Inferior Leads	ISCI
Subendocardial Injury In Inferolateral Leads	INJIL
Ischemic In Inferolateral	ISCIL
Subendocardial Injury In Lateral Leads	INJLA
Ischemic In Lateral Leads	ISCLA
Lateral Ischaemia	LIS
Subendocardial Injury In Anteroseptal Leads	INJAS
Ischemic In Anteroseptal Leads	ISCAS
Subendocardial Injury In Anterolateral Leads	INJAL
Ischemic In Anterolateral Leads	ISCAL
Left Atrial Abnormality	LAA
Left Atrial Hypertrophy	LAH
Right Atrial Abnormality	RAAB
Right Atrial Hypertrophy	RAH
Left Ventricular Hypertrophy	LVH
Voltage Criteria (QRS) For Left Ventricular Hypertrophy	VCLVH
Septal Hypertrophy	SEHYP
Right Ventricular Hypertrophy	RVH
ST-T Change Due To Ventricular Hypertropy	STTVH
Incomplete Right Bundle Branch Block	IRBBB
Complete Left Bundle Branch Block | Left Bundle Branch Block	CLBBB|LBBB
Incomplete Left Bundle Branch Block	ILBBB
Complete Right Bundle Branch Block | Right Bundle Branch Block	CRBBB|RBBB
Transient Ischemic Attack	TIA
Atrial Hypertrophy	AH
Myocardial Infarction	MI
Myocardial Ischemia	MIS
Ventricular Hypertrophy	VH
Coronary Heart Disease	CHD
Chronic Myocardial Ischemia	CMIS
Heart Failure	HF
Heart Valve Disorder	HVD
Left Ventricular Strain	LVS
Countercolockwise Rotation	-ROT
Clockwise Rotation	+ROT
Accelerated Atrial Escape Rhythm	AAR
Atrial Bigeminy	AB
Abnormal QRS	ABQRS
Atrial Escape Beat	AED
Atrial Fibrillation	AF
Atrial Fibrillation And Flutter	AFAFL
Atrial Flutter	AFL
Accelerated Idioventricular Rhythm	AIVR
Accelerated Junctional Rhythm	AJR
Suspect Arm Ecg Leads Reversed	ALR
Atrial Pacing Pattern	AP
Atrial Rhythm	ARH
Atrial Tachycardia	ATACH
Av Block	AVB

Atrioventricular Dissociation	AVD
Atrioventricular Junctional Rhythm	AVJR
Atrioventricular Node Reentrant Tachycardia	AVNRT
Atrioventricular Reentrant Tachycardia	AVRT
Bundle Branch Block	BBB
Blocked Premature Atrial Contraction	BPAC
Bradycardia	BRADY
Brugada	BRU
Brady Tachy Syndrome	BTS
Cardiac Dysrhythmia	CD
Congenital Incomplete Atrioventricular Heart Block	CIAHB
Clockwise Or Counterclockwise Vectorcardiographic Loop	CVCL/CCVCL
Diffuse Intraventricular Block	DIB
Early Repolarization	ERE
Fusion Beats	FB
Fqrs Wave	FQRS
High T-Voltage	HTV
Indeterminate Cardiac Axis	ICA
Idioventricular Rhythm	IR
Junctional Escape	JE
Junctional Premature Complex	JPC
Junctional Tachycardia	JTACH
Left Axis Deviation	LAD
Left Posterior Fascicular Block	LPFB
Prolonged Pr Interval	LPR
Low Qrs Voltages	LQRSV
Prolonged Qt Interval	LNGQT
Left Ventricular High Voltage	LVHV
Nonspecific Intraventricular Conduction Disorder	NSIVCB
Sinus Rhythm	NORM
Nonspecific St T Abnormality	NSSTTA
Old Myocardial Infarction	OLDMI
Premature Atrial Contraction | Supraventricular Premature Beats	PAC|SVPB
Prolonged P Wave	PPW
Pacing Rhythm	PR
Poor R Wave Progression	PRWP
Paroxysmal Supraventricular Tachycardia	PSVT
Paroxysmal Ventricular Tachycardia	PVT
P Wave Change	PWC
Qwave Abnormal	QAB
R Wave Abnormal	RAB
Right Axis Deviation	RAD
Rapid Atrial Fibrillation	RAF
Right Atrial High Voltage	RAHV
Sinus Arrhythmia	SARRH
Sinus Atrium To Atrial Wandering Rhythm	SAAWR
Sinoatrial Block	SAB
Sinus Arrest	SARR
Sinus Bradycardia	SBRAD
Sinus Node Dysfunction	SND
Shortened Pr Interval	SPRI
Decreased Qt Interval	SQT
Sinus Tachycardia	STACH

S T Changes	STC
St Depression	STD_
St Interval Abnormal	STIAB
Supraventricular Bigeminy	SVB
Supraventricular Tachycardia	SVT
T Wave Abnormal	TAB
T Wave Inversion	TINV
Tall P Wave	TPW
U Wave Abnormal	UAB
Ventricular Bigeminy	VBIG
Ventricular Ectopics	VEB
Ventricular Escape Beat	VESB
Ventricular Escape Rhythm	VESR
Ventricular Fibrillation	VF
Ventricular Flutter	VFL
Premature Ventricular Contractions | Ventricular Premature Beats	VPC|VPB
Ventricular Pre Excitation	VPEX
Ventricular Pacing Pattern	VPP
Paired Ventricular Premature Complexes	VPVC
Ventricular Tachycardia	VTACH
Ventricular Trigeminy	VTRIG
Wandering Atrial Pacemaker	WAP
Wolff-Parkinson-White Pattern	WPW
Low Voltage	LVOLT
TU fusion	TUF
atrial premature complexes non-conducted	PAC_NC
av conduction ration N:D	AVCR
left anterior fascicular block	LAFB
junctional escape complex(es)	JEC
ST deviation with T-wave change	STTC
left anterior/posterior fascicular block	LAFB/LPFB
non specific intraventricular conduction disturbance	IVCD
non specific t wave changes	NT_
sinus rhythm	SR
digitalis effect	DIG
premature complex(es)	PRC(S)
supraventricular arrhythmia	SVARR
trigeminal pattern (unknown origin SV or Ventricular)	TRIGU
low amplitude t waves	LOWT
electrolytic disturbance or drug (former EDIS)	EL
bigeminal pattern (unknown origin SV or Ventricular)	BIGU
normal functionning artificial pacemaker	PACE
non-diagnostic t abnormalities	NDT
ST-T changes compatible with ventricular aneurysm	ANEUR
non specific ST changes	NST_
non specific ischemic	ISC_
hypertrophy	HYP
high qrs voltage	HVOLT

Supplementary Table 2: Linear evaluation performance of pre-trained HuBERT-ECG BASE models when using labels generated by MFCC- and time-frequency based k-means models.

Label generator	Macro-averaged AUROC
MFCC-based k-means – C = 100	0.933
Time-frequency based k-means – C = 100	0.913

Supplementary Table 3: Architecture design and running configuration when experimenting with ECGs sampled at 50 and 100 Hz.

Architectural outline	Sampling rate	
	50 Hz	100 Hz
Convolutional embedder	kernels = (10, 3, 3, 2) strides= (4, 2, 2, 2) channels = 512	kernels = (10, 3, 3, 2, 2) strides= (4, 2, 2, 2, 2) channels = 512
Transformer encoder	BASE	
k-means	MFCC, C = 100	
Pre-training steps	130k	
Linear evaluation steps	80k	
Batch size	448	
P	33%	
LayerDrop	0.1	
Dropout	0.1	
Weight decay	0.01	

Supplementary Table 4: Linear evaluation performance of HuBERT-ECG on *Ribeiro-dev* after being pre-trained at 50 and 100 Hz. Linear evaluation steps necessary to let validation AUROC plateau are also reported.

ECG sampling rate	macro-averaged AUROC	Linear evaluation steps at plateau
50 Hz	0.9176	65k
100 Hz	0.9330	

Supplementary Table 5: Effects of the use of dynamic regularisation when pre-training HuBERT-ECG in terms of pre-training steps and upstream performance under the form of validation loss and validation accuracy. For each setup, linear evaluation performance on *Ribeiro-dev* is also reported.

Setups	Pre-training steps	Pre-training (validation loss, validation accuracy)	macro-averaged AUROC
Dynamic regularisation & Default regularisation Terms	80k	0.6379, 0.8205	0.933
No Dynamic regularisation & Default regularisation Terms	235k	0.6493, 0.8182	0.921
(C) Dynamic regularisation & Maximum regularisation Terms	72.5k	0.6775, 0.8111	0.9263
No Dynamic regularisation & Maximum regularisation Terms	90k	0.5860, 0.8312	0.9244

Supplementary Table 6: Linear evaluation performance of HuBERT-ECG BASE after being pre-trained in a multi-task learning framework in which tasks are represented by multiple k-means models composing an ensemble of label generators. The corresponding number of pre-training steps is reported for every task/cluster ensemble.

K-means ensemble	Macro-averaged AUROC	Pre-training steps
MFCC-based k-means, C = 100	0.933	80k
MFCC-based k-means, C = 100, 200	0.930	75k
MFCC-based k-means, C = 100, 200, 300	0.943	135k

Supplementary References

1. Ribeiro, A. H. *et al.* Automatic diagnosis of the 12-lead ECG using a deep neural network. *Nat Commun* 11, 1760 (2020).
2. Hsu, W.-N. *et al.* HuBERT: Self-Supervised Speech Representation Learning by Masked Prediction of Hidden Units. (2021).
3. Xu, M. *et al.* HMM-Based Audio Keyword Generation. in 566–574 (2004). doi:10.1007/978-3-540-30543-9_71.
4. Thorndike, R. L. Who belongs in the family? *Psychometrika* 18, 267–276 (1953).
5. Davies, D. L. & Bouldin, D. W. A Cluster Separation Measure. *IEEE Trans Pattern Anal Mach Intell* PAMI-1, 224–227 (1979).
6. Wagner, P. *et al.* PTB-XL, a large publicly available electrocardiography dataset. *Sci Data* 7, 154 (2020).
7. Liu, H. *et al.* A large-scale multi-label 12-lead electrocardiogram database with standardized diagnostic statements. *Sci Data* 9, 272 (2022).
8. Hong, S., Zhang, W., Sun, C., Zhou, Y. & Li, H. Practical Lessons on 12-Lead ECG Classification: Meta-Analysis of Methods From PhysioNet/Computing in Cardiology Challenge 2020. *Front Physiol* 12, (2022).
9. Rogers, A., Kovaleva, O. & Rumshisky, A. A Primer in BERTology: What we know about how BERT works. (2020).
10. Calinski, T. & Harabasz, J. A dendrite method for cluster analysis. *Commun Stat Theory Methods* 3, 1–27 (1974).