ECGBERT: Understanding Hidden Language of ECGs with Self-Supervised Representation Learning

Seokmin Choi[1,2,†] **Sajad Mousavi**[1,†*] **Phillip Si**[1,3,†]
Haben G. Yhdego[1] **Fatemeh Khadem**[1] **Fatemeh Afghah**[4]
[1]*CardioPhi LLC*, CA, USA
[2]University at Buffalo, SUNY, NY, USA
[3]Carnegie Mellon University, PA, USA
[4]Clemson University, SC, USA
{seokmin.choi,sajad.mousavi,phillip.si}@cardiophi.com
{haben.yhdego,fatemeh.khadem}@cardiophi.com
fatemeh.afghah@clemson.edu

Abstract

In the medical field, current ECG signal analysis approaches rely on supervised deep neural networks trained for specific tasks that require substantial amounts of labeled data. However, our paper introduces ECGBERT, a self-supervised representation learning approach that unlocks the underlying language of ECGs. By unsupervised pre-training of the model, we mitigate challenges posed by the lack of well-labeled and curated medical data. ECGBERT, inspired by advances in the area of natural language processing and large language models, can be fine-tuned with minimal additional layers for various ECG-based problems. Through four tasks, including Atrial Fibrillation arrhythmia detection, heartbeat classification, sleep apnea detection, and user authentication, we demonstrate ECGBERT's potential to achieve state-of-the-art results on a wide variety of tasks.

1 Introduction

The Centers for Disease Control (CDC) reported that heart disease is the leading cause of death in the United States [for disease control, 2022]. Specifically, one person dies every 34 seconds from cardiovascular disease and about 697,000 people died from heart disease in 2020, which is one in every five deaths. The electrocardiogram (ECG) is the most essential bio-signal used by cardiologists and physicians to keep track of heart activity and detect different heart-related diseases and is used by cardiologists and physicians. One of the most critical limitations of ECG signals is that it requires manual analysis and annotation. Furthermore, the interpretation of the ECG signals varies from physician to physician as different heart diseases are associated with complex patterns within the ECG which can be hard to detect. The resulting inconsistencies may affect diagnostic accuracy or the trust between the patient and the physician. Therefore, to mitigate the aforementioned limitations in regard to manual ECG interpretation, several studies have proposed alternative ECG analysis techniques to achieve higher accuracy in real-time. Among these, deep learning-based approaches have recently gained traction in this domain [Pyakillya et al., 2017, Mousavi et al., 2020]. Compared with machine learning-based approaches where features need to be extracted manually, deep learning-based approaches automatically extract relevant features [Rajpurkar et al., 2017, Isin and Ozdalili, 2017], allowing for improved performance given enough data and a sufficiently expressive model. As a result, deep learning techniques have been widely applied to the medical domain in recent years to

[*] Corresponding author. †Equal contribution.

Preprint. Under review.

solve different medical-related problems such as Atrial Fibrillation (AFIB) arrhythmia detection or heartbeat classification Mousavi and Afghah [2019], Andersen et al. [2019].

However, even with a large amount of data and sufficient computation, deep learning models are typically designed for specific tasks, limiting their applicability to one task at a time. Achieving optimal performance with deep learning models necessitates a substantial amount of data, which is a critical challenge in the medical field due to privacy constraints and data availability. While the curated and well-labeled ECG data is limited, there is a wealth of unlabeled ECG data that remains untapped. In addition, deep learning models, which are constructed by stacking multiple layers, require a large number of learnable parameters and extensive data as the model becomes deeper. Even when these complications have been resolved, the resulting models are tailored to their specific task and lack versatility for broader applications. Therefore, despite the fact that previous ECG-related models have shown promising results, the methods have limited applicability in the real world. To overcome these limitations of deep learning-based methods, it is necessary to design a more versatile and universal model that also resolves the data label or annotation issue.

A large language model (LLM) is a machine learning model within the field of natural language processing (NLP) that is capable of processing and emulating human language. LLMs are trained on a vast amount of text data and exhibit remarkable capabilities in diverse NLP tasks, including summarization [Cai et al., 2021], text generation [Kumar et al., 2021], sentiment analysis [Liu et al., 2012], or question-answering [Khot et al., 2020]. Among the most famous LLMs currently are BERT [Devlin et al., 2018], GPT and its variants [Radford et al., 2018], [Brown et al., 2020], and transformer-based architecture models [Vaswani et al., 2017]. Given the outstanding performance demonstrated by LLMs in NLP, researchers have explored the possibility of extrapolating LLM models to other domains. For example, Khan et al. [2022] introduced vision transformers which divide an image into smaller patches, the equivalent of words in NLP. Transformers have also been applied to a variety of robotics problems [Brohan et al., 2022], and LLMs were used in the field of law as well [Choi et al., 2023, Nay, 2023].

Recently, Mousavi et al. [2021] introduced a novel method called ECG language processing (ELP), which applies NLP-style models to analyze ECG signals. While traditional models commonly employed in ECG analysis are limited by their reliance on labeled data, the medical field possesses an extensive repository of unrefined and unlabeled ECG records. This parallelism with the corpus of textual data highlights the vast number of unrefined and unlabeled ECG records within the medical field, which remains untapped by conventional models designed for ECG analysis.

To address this challenge, we introduce ECGBERT, a novel LLM model framework inspired by BERT [Devlin et al., 2018]. ECGBERT capitalizes on large amounts of unlabeled ECG data during the pre-training stage to learn meaningful representations such that downstream tasks can be adapted efficiently with a minimal amount of labeled data. To achieve this, we integrate the framework for BERT (Bidirectional Encoder Representations from Transformers), an LLM used within the NLP domain, with the upgraded version of ELP paradigm, as shown in Figure 1. 2 illustrates a detailed overview of the model architecture.

The proposed framework allows us to create a versatile and potent tool applicable to various medical tasks including, but not limited to, heartbeat classification, cardiac arrhythmias detection, sleep Apnea detection, or even user authentication, all of which rely on ECGs as the input. Because we pre-train ECGBERT on a large amount of ECG data in an unsupervised manner, this enables the model to learn and represent the nuances, complexities, and latent patterns of ECG signals without the need for human supervision or annotation. This unique capability of ECGBERT facilitates a more efficient and effective analysis and interpretation of ECG signals.

Our proposed framework's contributions are multifold: Firstly, it analyzes and details the method of mapping between textual tokens within the NLP domain and **continuous** ECG signals. Secondly, it permits learning more general representations of ECG signals while still capturing subtler fine-grained pattern differences after the transformation into tokens. Thirdly, it offers embedded explainability and interpretability as it is treating ECG signals like text within the NLP domain, shedding light on the model's decision-making process. To our best understanding, this is the first paper that truly re-interprets ECG signals from an unsupervised LLM perspective.

Our paper is structured in the following manner:

- We interpret the time series ECG signals as integer-encoded ECG tokens by creating a wave vocabulary and wave assignment.
- We utilize a CNN encoder to create shift-invariant embeddings to capture the finer-grained characteristics of the ECG signals in combination with general ELP tokens.
- We introduce ECGBERT, a novel deep learning model that combines state-of-the-art ideas from both ECG and NLP domains, that can be applied to various downstream tasks after it is pre-trained in an unsupervised manner (i.e., unlabeled data).
- We show that ECGBERT performs competitively across four distinct downstream tasks, demonstrating its versatility and effectiveness.

2 Related Work

BERT, the primary inspiration for our work, is based on a transformer architecture that consists of attention blocks where each output is connected to its input and determines the importance of their relationship [Vaswani et al., 2017]. One of the most essential characteristics of BERT is the departure of heavy reliance on well-refined labels during pre-training, which is especially suitable for medical data due to the myriad privacy restrictions. Moreover, BERT was proposed to resolve one of the critical limitations of the previous language models: uni-directionality, which can only leverage the previous tokens in the attention layers [Radford et al., 2018], resulting in suboptimal performance on downstream tasks which requires utilizing context from both directions such as question answering.

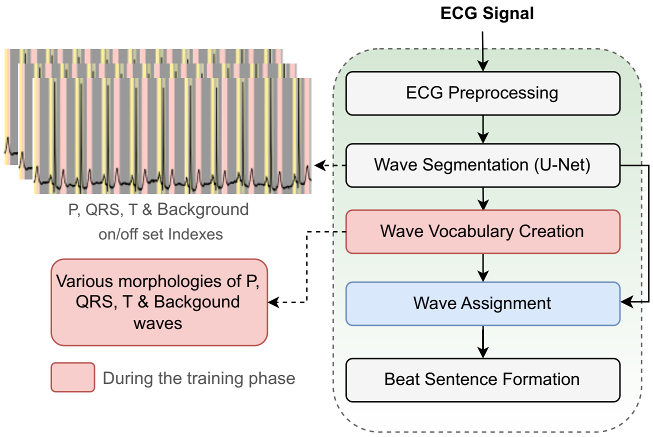

Figure 1: Holistic view of ECG language processing (ELP)

2.1 ECG with Deep Learning

As deep learning methods approach human-level accuracy while far surpassing their human counterparts in predictive speed, researchers have begun to apply deep learning techniques to resolve various tasks using ECG data, such as heartbeat classification [Mathews et al., 2018], heart arrhythmia detection [Singh et al., 2018], interpretable AFIB classification [Nankani and Baruah, 2022], cardiovascular disease diagnosis [Qiu et al., 2023], and sleep apnea [Feng et al., 2020], with each of them leveraging different architecture styles ranging from dense networks to RNNs to HMMs.

However, all of these studies are task-specific and do not seek to learn a general representation of the ECG language. In this study, we seek to create a framework that can extract good representations for ECG signals which can also utilize context effectively.

2.2 Large Language Model in the Medical Domain

Due to the LLMs' outstanding performance in NLP, researchers have tried deploying LLMs to the medical domain due to a range of benefits. Firstly, to improve the accuracy and reliability due to the ability of LLMs to learn the relationships between different medical conditions and symptoms. Secondly, helping interpret medical tests such as laboratory results with explainable attention weights.

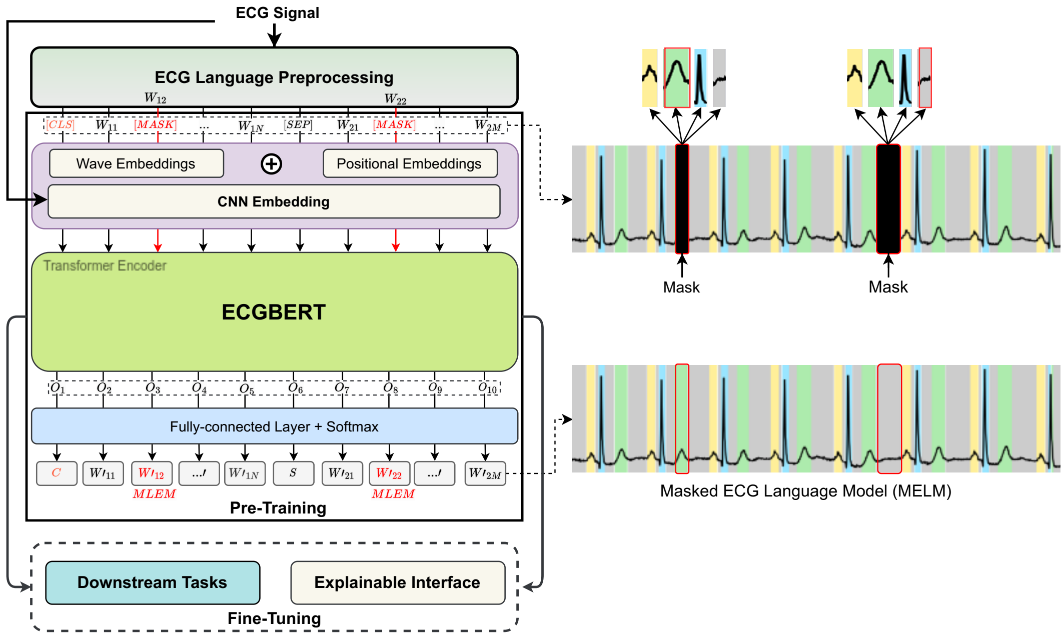

Figure 2: Illustration of the proposed ECGBERT architecture

Finally, LLMs can possibly be applied to identify clinical trials which may be relevant to the patient's health condition. GatorTron [Yang et al., 2022] develops a large clinical language model using more than 90 billion words of text, and performs five clinical NLP tasks such as medical relation extraction or semantic textual similarity. BEHRT [Li et al., 2020] introduces a deep sequence transduction model for electronic health records (EHR) which is capable of predicting the likelihood of 301 conditions in future visits. Moreover, MED-BERT [Rasmy et al., 2021] adopts the BERT framework to pre-train with structured EHR data and implements two disease prediction tasks from two clinical databases. BioBERT [Lee et al., 2020] leverages a biomedical domain corpora to pre-train the model and conducts different biomedical text mining tasks (e.g, question answering or next entity recognition) to validate the model's effectiveness. Although the aforementioned studies show promising results, all of them leveraged either the clinical notes or EHR data, which is still textual data. However, ECGBERT employs time series ECG signals, which are transformed into words and vectors to be fed into the proposed model.

2.3 Transfer Learning from Supervised ECG Data

Transfer learning has become an increasingly popular area of research in various domains of machine learning like NLP, and computer vision. It has also shown promising results for ECG analysis using supervised data such as ECG arrhythmia classification [Salem et al., 2018, Strodthoff et al., 2020, Weimann and Conrad, 2021] and transferable representation [Kachuee et al., 2018]. In addition, Li et al. [2021] proposed a photoplethysmography (PPG) and actigraphy-based sleep stage classification by applying a model trained on an ECG dataset, and Zhang et al. [2021] introduced heartbeat classification by adopting a transfer learning technique with a Residual Neural Network (ResNet) that was pre-trained on ImageNet to analyze ECG-signals which were transformed into 2-D time-frequency diagrams. It can be seen that ECG-related tasks are highly amenable to transfer-based learning. We adopt this in an unsupervised manner by pre-training ECGBERT and fine-tuning the model on different downstream tasks.

3 Methodology

ECGBERT pipeline is composed of two macroscopic tasks: ECG language processing and model training. In particular, the former is mainly composed of five modules: ECG preprocessing, wave segmentation, wave vocabulary creation, wave assignment, and beat sentence formation.

3.1 ECG Language Processing

3.1.1 ECG Preprocessing

Since raw ECG signals are contaminated during the recording by various factors such as motion artifacts or powerline interference, it is essential to polish the raw data to be cleaner by applying different signal processing techniques. Specifically, in the ECG signals, power-line interference and baseline wander are the two major factors that induce ECG signal corruption [Mian Qaisar, 2020]. First, to eliminate the powerline interference from the ECG signals, we apply a second-order Butterworth band-stop filter with cut-off frequencies at 50 Hz and 60 Hz. After mitigating the powerline interference issue, we remove the baseline wander by leveraging discrete wavelet transforms (DWT). The objective of the DWT is to decompose the signal into different resolutions using high-pass and low-pass components. In this study, we decompose the signal into one level using the Daubechies 4 wavelet family. We shift and calculate the energy of the detail coefficients, and the baseline is then reconstructed from this level using low-pass signals, as baseline wander is a low-frequency artifact.

3.1.2 Wave Segmentation

Segmenting heartbeats of an ECG signal can drastically affect the diagnosis accuracy. Specifically, in this study, we divide ECG signals into P waves, T waves, QRS complexes, and background waves. Even if parts of an ECG signal belong to the same wave, due to different external factors such as motion artifacts and noises, it can be challenging to extract the correct segmented waves. To segment the ECG signals into a set of different waves, we first adopt the Hamilton algorithm [Hamilton, 2002] to clean the ECG signals to improve the quality of R-peak detection. We then apply DWT to separate the different waves of the cardiac cycle: 1) the P wave, which is a depolarization wave that spreads throughout the atria, 2) the QRS complex, which has a larger amplitude than other waves and shows a rapid depolarization of both verticles, 3) the T wave, which reflects the ventricular repolarization of the ventricles, and 4) background waves, which don't belong to any of three waves. Based on the time-frequency analysis of the ECG signals, wave segmentation will return the onset and offset indices of each wave.

3.1.3 Wave Vocabulary Creation

ECG Patterns within a given wave group may exhibit variations, even though they belong to the same group. Within the same wave group, different wave types represent different phases of heart activities: electrical activity towards or away from a lead that causes an upward or downward deflection respectively. In particular, heart-related diseases can be identified by interpreting the unique morphology of the waves.

According to a previous study [Rawshani], P waves can be represented by five different morphologies, QRS waves can be represented by twelve morphologies and T waves can be categorized as seven morphologies. As the proposed system is trained to interpret and analyze different wave types to learn the representations of the ECG signals, it is critical to categorize the waves into different groups. So, we employ a clustering model to create a comprehensive ECG wave vocabulary. However, it will be time and labor-intensive if clustering the waves requires the morphology labels. Therefore, we adopt an unsupervised clustering algorithm (i.e., Kmeans) that employs Dynamic Time Warping (DTW) [Berndt and Clifford, 1994] to complete this step. DTW was introduced to measure the similarity between two given temporal sequences which may vary in speed. One of the main advantages of applying DTW is that it shows reliable time alignment when there are two similar patterns with different duration. This is especially crucial for the ECG signals considering the similarity in patterns despite the difference in each patient's heartbeat cycle. To cluster the ECG waves, we train four different clustering models to categorize the P waves into 12 clusters, QRS waves into 19 clusters, T waves into 14 clusters, and background waves into 25 clusters (70 clusters in total). Since different wave morphologies carry different chunks of information about the corresponding ECG signal, we reinterpret the waves as an analogous component to words within the NLP domain. Therefore, the clustering algorithm could be viewed as a method that groups words together based on their semantic similarity to create a wave vocabulary.

3.1.4 Wave Assignment

Afterward, each segmented wave is fed into a different clustering model depending on the wave type to assign the wave to the proper corresponding cluster. This encodes the ECG signal into a sequence of integer tokens.

3.1.5 Beat Sentence Formulation

Based on the encoded beat waves, we then construct a sentence that consists of one or more heartbeats. In particular, to enhance the generalizability during pre-training, a sentence is formulated by either one, two, six, or eight consecutive heartbeats randomly, where a heartbeat is comprised of multiple consecutive waves. When constructing a sentence composed of one or more heartbeats, we need to determine the onset and offset of the heartbeat wave. Normally for the ECG interpretations, first the P-wave is interpreted, followed by the PR wave in between the offset of the P wave and the onset of the next QRS wave. In this study, the PR wave is categorized as one of the background waves. Afterward, the QRS wave is interpreted, and then the T wave. In summary, the order of a beat sentence starts with P, QRS, and T, with background clusters in between the defined wave clusters.

3.2 Bidirectional Transformer

3.2.1 Input Embedding Representations

After constructing an ECG sentence from the clustered waves, these sentences are then fed into the Bidirectional Transformer part of the pipeline, which is shown in Figure 2. For the input representations, we include positional embeddings, token embeddings, and CNN embeddings.

Positional Embeddings Positional embeddings are added to assign orders to the non-recurrent multihead attention, creating a temporal context for the tokens. This is especially critical to ECGBERT which tokenizes the heartbeat wave because the ECG record loses some relevant temporal information during the tokenizing step.

Token Embeddings Token embeddings convert heartbeat waves into different tokens. In particular, for the tokenizer, we leverage our own predicted clusters from the wave vocabulary creation module to represent a heartbeat sentence in a single set of tokens. Furthermore, the `[SEP]` token is added at the end of the sentence to serve as a marker indicating the end of the sentence.

CNN Embeddings Unfortunately, directly tokenizing the ECG signals and applying BERT's schematic suffers from the inability to capture finer-grained details of ECG signals. While token embeddings provide general representations for each token in input text by capturing the meaning of each word in context and positional embeddings reflect the position of each token in the sequence, the continuous time-series data structure of ECG signals requires much more refined input representations to fully capture subtle differences in various ECG patterns. To address this limitation, we introduce a CNN token embedding that serves as a feature extractor of raw ECG signals. Specifically, we adopt a U-Net architecture with two downsampling and upsampling blocks, along with skip connections and batch normalization layers. The output of the CNN feature extractor is then segmented based on the onset and offset indices of each wave and added with token and positional embeddings.

3.2.2 Model Architecture

The model portion of ECGBERT adopts a transformer-encoder-style architecture. The main rationale behind applying attention from transformer architectures is to focus on appropriate parts of the ECG sequence and determine the important neighboring components by applying scaled dot-product operation. In addition, we also adopt multi-headed attention layers in this framework.

3.2.3 Pre-training ECGBERT

Similar to BERT, we apply Masked Language Modeling (MLM) unsupervised technique during the pre-training stage. We apply MLM to train a bidirectional representation of the ECG signals which helps the model learn the temporal relationships between nearby heartbeats. As it can be seen from the two right figures in Figure 2, MLM is implemented by masking 15% of the wave tokens randomly

Table 1: AFIB Arrhythmia Detection Performance Comparison with State-of-the-Art Methods (RRI: RR-interval)

Model	**Paradigm**	**Signal Length**	**Performance**			
			Accuracy	***Specificity***	***Sensitivity***	***PPV***
Tuboly et al. [2021]	Intra-patient	60s	0.980	0.987	0.974	0.988
ResNet	Inter-patient	10s	0.884	0.951	0.846	0.969
Andersen et al. [2019]	Inter-patient	30 RRIs	**0.978**	**0.989**	0.969	0.957
Pereira and Andreão [2022]	Inter-patient	10	0.908	0.910	0.915	-
ECGBERT	Inter-patient	10s	0.973	0.976	**0.970**	**0.981**

and asking ECGBERT to predict the masked waves. We didn't include Next Sentence Prediction (NSP) task, as a necessity of the NSP is not as crucial as originally thought which was discussed in some of the previous studies[Cui et al., 2021, Tinn et al., 2023]. In particular, the NSP task only requires the model to predict whether two given sentences are consecutive or not, which may not be a good representation of the subtle but complex patterns that exist in ECG signals.

3.2.4 Pre-training Datasets

Since ECGBERT adopts an unsupervised learning approach during the pre-training stage where the model learns the general representations of ECG signals, utilizing ECG datasets without labels is one of the main contributions of our proposed system which can reduce the cost and time. For the pre-training purpose, we adopt MIMIC-III waveform [Moody et al., 2020], PTB-XL [Wagner et al., 2020], Georgia [Alday et al., 2020], CPSC-2018 [Liu et al., 2018] datasets which can be downloaded from the Physionet website [Goldberger et al., 2000]. We then pretrain our proposed ECGBERT using two RTX 2080Ti GPUs on a local machine, with around 236 hours of data due to the limited computing power. With more data and computing power, we expect the ECGBERT to learn even better representations than the model we used for our downstream tasks.

3.2.5 Fine-Tuning ECGBERT

Following the pre-training phase, the pre-trained ECGBERT model is utilized in diverse downstream applications, extending beyond medical tasks such as AFIB or heart arrhythmia detection. The potential of deploying ECGBERT for alternative tasks like sleep apnea detection or even user authentication has also been investigated. This low-cost fine-tuning process involves adding one or two additional dense layers on top of the pre-trained ECGBERT model, which can be achieved by adjusting the model weights. This process will be further detailed in the subsequent section.

4 Experiments

This section presents four different downstream tasks using a pre-trained ECGBERT from the previous section. For evaluation metrics, we adopt accuracy, specificity, sensitivity, and positive predictive value (PPV). For the datasets for which other papers have different data setups (e.g. different input lengths or inter/intra patient setups), we benchmark against a ResNet model containing three residual blocks with ReLU activations before a final linear layer. Batchnorms are applied after every convolutional layer within each of the blocks. The details for the publicly available datasets used for training and evaluation can be found in Appendix A.1.2. In this study, we focus on the inter-patient evaluation scheme, which reflects a more realistic scenario compared to the commonly employed intra-patient paradigm. In an inter-patient experimental design, train and test data are divided at a patient level before splitting into sub-segments. On the other hand, an intra-patient scheme divides the data into small segments first before randomly assigning these small segments to train or test, which allows data from the same patient to be in both train and test sets. As a result, train and test distributions are much more similar to an intra-patient schematic, which is unrealistic in real-world scenarios. By avoiding biases introduced by training and testing on the same patient's samples, our results reflect a more reliable comparison with existing methods [De Chazal et al., 2004].

Table 2: Confusion Matrix and Per-Class Performance of Heartbeat Classification Implemented by ECGBERT on the MIT-BIH Database.

		Predicted				**Per-class Performance**			
		N	**S**	**V**	**Q**	***Accuracy***	***Specificity***	***Sensitivity***	***PPV***
True	**N**	38538	1483	1941	1119	0.86	0.45	0.89	0.94
	S	187	26	39	7	0.95	0.99	0.10	0.01
	V	1778	201	1280	277	0.91	0.94	0.36	0.38
	Q	451	77	25	2445	0.96	0.99	0.82	0.64

Table 3: Performance on User Verification Task

Model	**Performance**			
	Accuracy	***Specificity***	***Sensitivity***	***PPV***
ResNet	0.993	0.995	0.811	0.882
ECGBERT	**0.999**	**1.000**	**1.000**	**0.972**

4.1 AFIB Arrhythmia Detection

For this downstream task, we explore the performance of the pre-trained ECGBERT model on the detection of AFIB rhythms by leveraging the MIT-BIH Atrial Fibrillation database [Moody, 1983].

In this study, we fine-tune the model with two extra dense layers on top of the ECGBERT with a learning rate of 0.001 and a batch size of 64 for 13 epochs. As in Table 1, ECGBERT achieves an accuracy of 0.973, a specificity of 0.976, a sensitivity of 0.970, and a PPV of 0.981. Note that we only compared the performance with previous studies that evaluated the performance based on the inter-patient approach (i.e., [Andersen et al., 2019, Pereira and Andreão, 2022]). Even though some studies may outperform ECGBERT (e.g., [Tuboly et al., 2021]) their approaches seem to be evaluated on intra-patient data, which is unrealistic as similar distributions at train and test time are not guaranteed under real-world circumstances. Despite these noted advantages, however, ECGBERT can still achieve comparable performance to those studies.

4.2 Heartbeat Classification

Next, we fine-tune ECGBERT to perform Heartbeat Classification by leveraging the MIT-BIH Arrhythmia database [Moody and Mark, 2001] with an inter-patient paradigm. Among different beat types, we divide the beats into five classes: normal (N), unknown (Q), supraventricular ectopic (S), ventricular ectopic (V), and fusion (F) heartbeat groups recommended by the American Association of Medical Instrumentation (AAMI) [ANSI-AAMI, 1998-2008]. For finetuning layers, we add two residual blocks on top of BERT to capture a variable number of labels to predict. We finetune the model with a learning rate of 1e-4 for 20 epochs.

From Table 2, we can see that ECGBERT offers a reasonable performance which still has room for improvement since beat classification tasks need to have appropriate masking and data organization in order to classify a variable number of beats within a 10s ECG segment.

4.3 User Verification and Identification with ECG signals

Besides the aforementioned cardiovascular-related tasks, we also compare a downstream task that is not related to cardiovascular disease, user authentication. Previous studies have demonstrated that ECG signals can serve as a biometric authentication method [Odinaka et al., 2012, Labati et al., 2019]. We analyze our results on the MIT-BIH arrhythmia dataset with 48 patient records. We label the data according to the user ID. We fine-tune the model for 20 epochs with a learning rate of 1e-4 and a batch size of 128. On a held-out test set, ECGBERT achieved an accuracy of 0.999 and a PPV of 0.972, as compared to an accuracy of 0.993 and a PPV of 0.882 for the ResNet, as it can be seen from Table 3.

Furthermore, we also test on a user identification task, where the model is to predict which user a certain ECG is from. Once again, ECGBERT outperforms the ResNet baseline as in Table 4. Due to the dataset being balanced across all classes, we only report the test accuracy for both models.

Table 5: Inter-patient Performance on Sleep Apnea Detection

Model	**Performance**			
	Accuracy	***Specificity***	***Sensitivity***	***PPV***
ResNet	0.709	**0.744**	0.653	0.606
ECGBERT	**0.725**	0.626	**0.831**	**0.678**

4.4 Sleep Apnea Detection

Table 4: Performance on User Identification Task

Model	***Accuracy***
ResNet	0.920
ECGBERT	**0.938**

Obstructive sleep Apnea (OSA) plays a crucial role in health because it can potentially cause life-threatening problems such as heart failure or cognitive impairments [Beaudin et al., 2021]. Therefore, we chose to explore sleep Apnea detection with the PhysioNet Apnea-ECG Database v1.0.0 [Penzel et al., 2000]. We fine-tuned the ECGBERT with two extra dense layers with a batch size of 64 and a learning rate of 0.005, respectively for 5 epochs. From Table 5, our baseline ResNet model achieved an accuracy of 0.709, a specificity of 0.744, a sensitivity of 0.653, and a PPV of 0.606. On the other hand, ECGBERT achieved an accuracy of 0.725, specificity of 0.626, sensitivity of 0.831, and PPV of 0.678. We believe the low performance is due to the noisiness of the signals compared with other datasets, making it difficult for the model to learn the specific representations. Moreover, since the length of the input data is 60 seconds, it makes it even more challenging since we only feed in 10-second sub-sequences of the Apnea data, while the labels are given per 60-second segment. Compared to the ResNet baseline, ECGBERT shows lower specificity but much higher sensitivity, along with higher overall accuracy. Sensitivity, where correctly identifying positive instances, is a very critical concept in a medical study because high sensitivity demonstrates that the model is effective in capturing AFIB cases correctly.

5 Limitations and Future Work

We acknowledge that this work has some notable limitations. First, we do not conduct more experiments such as an ablation study, comparison of model parameters, or comparison with other works, which leaves the results with room for improvement.

Second, pre-training ECGBERT to learn more general representations of the ECG signals requires much more time and computing. Since this is a proof-of-concept study to whether the ECG signals can be interpreted as a language, we only trained on a minor subset, about 236 hours. In addition, we only employ the most widely-adopted lead-II ECG signals due to the fact that multi-lead downstream datasets currently are few and far between, and different conventions would require that ECGBERT fits all these paradigms. Further modifications could possibly be made to ECGBERT to take in a variable number of leads, with a specialized masking method so that the transformer only uses the non-masked signals. This could result in a multifaceted and more robust language model, though at the same time, it would require a vastly greater amount of computing.

Finally, ECG language processing steps in Section 3.1 require the segmentation model to be able to correctly segment the ECG signal into appropriate small waves. In the presence of low-quality ECG signals, the wave segmentation module inconsistently and inaccurately segments the small waves for the abnormal ECG signals, which results in noisier outputs that affect the ECGBERT performance. This can be resolved with a more robust segmentation model with deep learning approaches to help better segment the noisy and abnormal ECG signals.

6 Conclusion

In this study, we propose the ECGBERT, a novel framework that can interpret the ECG signals and perform different downstream tasks with a single pre-trained model. The proposed approach consists of two main steps: 1) ECG language processing, and 2) large language modeling. After ECGBERT is pre-trained with a large amount of unlabeled data, it can be deployed to any downstream task that uses ECG signals as inputs. As we hypothesized, ECGBERT successfully learns general representations of different types of ECG signals, as shown by its stable performance on a variety of downstream tasks. We view this paper as the first foray into effectively utilizing a large amount of unlabeled, uncurated ECG data present online to solve various tasks efficiently without relying on internal and restricted datasets. This provides a non-block-box, repeatable model framework which allows for more accountability and repeatability on the part of deep learning researchers in the medical domain.

References

Erick Andres Perez Alday, Annie Gu, Amit Shah, Chad Robichaux, An-Kwok Ian Wong, Chengyu Liu, Feifei Liu, Ali Bahrami Rad, Andoni Elola, Salman Seyedi, Qiao Li, Ashish Sharma, Gari D. Clifford, and Matthew A. Reyna. Classification of 12-lead ecgs: the physionet/computing in cardiology challenge 2020. *medRxiv*, 2020. doi: 10.1101/2020.08.11.20172601. URL `https://www.medrxiv.org/content/early/2020/08/14/2020.08.11.20172601`.

Rasmus S Andersen, Abdolrahman Peimankar, and Sadasivan Puthusserypady. A deep learning approach for real-time detection of atrial fibrillation. *Expert Systems with Applications*, 115: 465–473, 2019.

ANSI-AAMI. Testing and reporting performance results of cardiac rhythm and st segment measurement algorithms. *American National Standards Institute, Inc. (ANSI), Association for the Advancement of Medical Instrumentation (AAMI), ANSI/AAMI/ISO*, 1998-2008.

Andrew E Beaudin, Jill K Raneri, Najib T Ayas, Robert P Skomro, Nurit Fox, AJ Marcus Hirsch Allen, Matthew W Bowen, Andrhea Nocon, Emma J Lynch, Meng Wang, et al. Cognitive function in a sleep clinic cohort of patients with obstructive sleep apnea. *Annals of the American Thoracic Society*, 18(5):865–875, 2021.

Donald J Berndt and James Clifford. Using dynamic time warping to find patterns in time series. In *KDD workshop*, volume 10, pages 359–370. Seattle, WA, USA:, 1994.

Anthony Brohan, Noah Brown, Justice Carbajal, Yevgen Chebotar, Joseph Dabis, Chelsea Finn, Keerthana Gopalakrishnan, Karol Hausman, Alex Herzog, Jasmine Hsu, et al. Rt-1: Robotics transformer for real-world control at scale. *arXiv preprint arXiv:2212.06817*, 2022.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

Xiaoyan Cai, Sen Liu, Junwei Han, Libin Yang, Zhenguo Liu, and Tianming Liu. Chestxraybert: A pretrained language model for chest radiology report summarization. *IEEE Transactions on Multimedia*, 2021.

Jonathan H Choi, Kristin E Hickman, Amy Monahan, and Daniel Schwarcz. Chatgpt goes to law school. *Available at SSRN*, 2023.

Yiming Cui, Wanxiang Che, Ting Liu, Bing Qin, and Ziqing Yang. Pre-training with whole word masking for chinese bert. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29:3504–3514, 2021.

Philip De Chazal, Maria O'Dwyer, and Richard B Reilly. Automatic classification of heartbeats using ecg morphology and heartbeat interval features. *IEEE transactions on biomedical engineering*, 51 (7):1196–1206, 2004.

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

Kaicheng Feng, Hengji Qin, Shan Wu, Weifeng Pan, and Guanzheng Liu. A sleep apnea detection method based on unsupervised feature learning and single-lead electrocardiogram. *IEEE Transactions on Instrumentation and Measurement*, 70:1–12, 2020.

Center for disease control. Heart disease facts, October 2022. URL `https://www.cdc.gov/heartdisease/facts.htm#:~:text=Coronary%20Artery%20Disease,-Coronary%20heart%20disease&text=About%2020.1%20million%20adults%20age,have%20CAD%20(about%207.2%25)`.

Ary L Goldberger, Luis AN Amaral, Leon Glass, Jeffrey M Hausdorff, Plamen Ch Ivanov, Roger G Mark, Joseph E Mietus, George B Moody, Chung-Kang Peng, and H Eugene Stanley. Physiobank, physiotoolkit, and physionet: components of a new research resource for complex physiologic signals. *circulation*, 101(23):e215–e220, 2000.

Pat Hamilton. Open source ecg analysis. In *Computers in cardiology*, pages 101–104. IEEE, 2002.

Ali Isin and Selen Ozdalili. Cardiac arrhythmia detection using deep learning. *Procedia computer science*, 120:268–275, 2017.

Mohammad Kachuee, Shayan Fazeli, and Majid Sarrafzadeh. Ecg heartbeat classification: A deep transferable representation. In *2018 IEEE international conference on healthcare informatics (ICHI)*, pages 443–444. IEEE, 2018.

Salman Khan, Muzammal Naseer, Munawar Hayat, Syed Waqas Zamir, Fahad Shahbaz Khan, and Mubarak Shah. Transformers in vision: A survey. *ACM computing surveys (CSUR)*, 54(10s):1–41, 2022.

Tushar Khot, Peter Clark, Michal Guerquin, Peter Jansen, and Ashish Sabharwal. Qasc: A dataset for question answering via sentence composition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 8082–8090, 2020.

Sachin Kumar, Eric Malmi, Aliaksei Severyn, and Yulia Tsvetkov. Controlled text generation as continuous optimization with multiple constraints. *Advances in Neural Information Processing Systems*, 34:14542–14554, 2021.

Ruggero Donida Labati, Enrique Muñoz, Vincenzo Piuri, Roberto Sassi, and Fabio Scotti. Deep-ecg: Convolutional neural networks for ecg biometric recognition. *Pattern Recognition Letters*, 126: 78–85, 2019.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240, 2020.

Qiao Li, Qichen Li, Ayse S Cakmak, Giulia Da Poian, Donald L Bliwise, Viola Vaccarino, Amit J Shah, and Gari D Clifford. Transfer learning from ecg to ppg for improved sleep staging from wrist-worn wearables. *Physiological measurement*, 42(4):044004, 2021.

Yikuan Li, Shishir Rao, José Roberto Ayala Solares, Abdelaali Hassaine, Rema Ramakrishnan, Dexter Canoy, Yajie Zhu, Kazem Rahimi, and Gholamreza Salimi-Khorshidi. Behrt: transformer for electronic health records. *Scientific reports*, 10(1):1–12, 2020.

Feifei Liu, Chengyu Liu, Lina Zhao, Xiangyu Zhang, Xiaoling Wu, Xiaoyan Xu, Yulin Liu, Caiyun Ma, Shoushui Wei, Zhiqiang He, et al. An open access database for evaluating the algorithms of electrocardiogram rhythm and morphology abnormality detection. *Journal of Medical Imaging and Health Informatics*, 8(7):1368–1373, 2018.

Kun-Lin Liu, Wu-Jun Li, and Minyi Guo. Emoticon smoothed language models for twitter sentiment analysis. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 26, pages 1678–1684, 2012.

Sherin M Mathews, Chandra Kambhamettu, and Kenneth E Barner. A novel application of deep learning for single-lead ecg classification. *Computers in biology and medicine*, 99:53–62, 2018.

Saeed Mian Qaisar. Baseline wander and power-line interference elimination of ecg signals using efficient signal-piloted filtering. *Healthcare technology letters*, 7(4):114–118, 2020.

B Moody, G Moody, M Villarroel, G Clifford, and I Silva. Mimic-iii waveform database (version 1.0)," physionet, 2020. *Online, url: https://doi. org/10.13026/c2607m*, 2020.

George Moody. A new method for detecting atrial fibrillation using rr intervals. *Proc. Comput. Cardiol.*, 10:227–230, 1983.

George B Moody and Roger G Mark. The impact of the mit-bih arrhythmia database. *IEEE engineering in medicine and biology magazine*, 20(3):45–50, 2001.

Sajad Mousavi and Fatemeh Afghah. Inter-and intra-patient ecg heartbeat classification for arrhythmia detection: a sequence to sequence deep learning approach. In *ICASSP 2019-2019 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 1308–1312. IEEE, 2019.

Sajad Mousavi, Fatemeh Afghah, and U Rajendra Acharya. Han-ecg: An interpretable atrial fibrillation detection model using hierarchical attention networks. *Computers in biology and medicine*, 127:104057, 2020.

Sajad Mousavi, Fatemeh Afghah, Fatemeh Khadem, and U Rajendra Acharya. Ecg language processing (elp): A new technique to analyze ecg signals. *Computer methods and programs in biomedicine*, 202:105959, 2021.

Deepankar Nankani and Rashmi Dutta Baruah. Atrial fibrillation classification and prediction explanation using transformer neural network. In *2022 International Joint Conference on Neural Networks (IJCNN)*, pages 01–08. IEEE, 2022.

John J Nay. Large language models as fiduciaries: A case study toward robustly communicating with artificial intelligence through legal standards. *arXiv preprint arXiv:2301.10095*, 2023.

Ikenna Odinaka, Po-Hsiang Lai, Alan D Kaplan, Joseph A O'Sullivan, Erik J Sirevaag, and John W Rohrbaugh. Ecg biometric recognition: A comparative analysis. *IEEE Transactions on Information Forensics and Security*, 7(6):1812–1824, 2012.

Thomas Penzel, George B Moody, Roger G Mark, Ary L Goldberger, and J Hermann Peter. The apnea-ecg database. In *Computers in Cardiology 2000. Vol. 27 (Cat. 00CH37163)*, pages 255–258. IEEE, 2000.

Rafael Pereira and Rodrigo Varejão Andreão. Inter-patient detection of atrial fibrillation in short ecg segments based on lstm network with multiple input layers. *Research on Biomedical Engineering*, 38(2):465–476, 2022.

Boris Pyakillya, Natasha Kazachenko, and Nikolay Mikhailovsky. Deep learning for ecg classification. In *Journal of physics: conference series*, volume 913, page 012004. IOP Publishing, 2017.

Jielin Qiu, William Han, Jiacheng Zhu, Mengdi Xu, Michael Rosenberg, Emerson Liu, Douglas Weber, and Ding Zhao. Transfer knowledge from natural language to electrocardiography: Can we detect cardiovascular disease through language models? *arXiv preprint arXiv:2301.09017*, 2023.

Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever, et al. Improving language understanding by generative pre-training. 2018.

Pranav Rajpurkar, Awni Y Hannun, Masoumeh Haghpanahi, Codie Bourn, and Andrew Y Ng. Cardiologist-level arrhythmia detection with convolutional neural networks. *arXiv preprint arXiv:1707.01836*, 2017.

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

Rawshani. Ecg interpretation: Characteristics of the normal ecg (p-wave, qrs complex, st segment, t-wave). `https://ecgwaves.com/topic/ecg-normal-p-wave-qrs-complex-st-segment-t-wave-j-point/`.

Milad Salem, Shayan Taheri, and Jiann-Shiun Yuan. Ecg arrhythmia classification using transfer learning from 2-dimensional deep cnn features. In *2018 IEEE biomedical circuits and systems conference (BioCAS)*, pages 1–4. Ieee, 2018.

Shraddha Singh, Saroj Kumar Pandey, Urja Pawar, and Rekh Ram Janghel. Classification of ecg arrhythmia using recurrent neural networks. *Procedia computer science*, 132:1290–1297, 2018.

Nils Strodthoff, Patrick Wagner, Tobias Schaeffter, and Wojciech Samek. Deep learning for ecg analysis: Benchmarks and insights from ptb-xl. *IEEE Journal of Biomedical and Health Informatics*, 25(5):1519–1528, 2020.

Robert Tinn, Hao Cheng, Yu Gu, Naoto Usuyama, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. Fine-tuning large neural language models for biomedical natural language processing. *Patterns*, 4(4), 2023.

Gergely Tuboly, György Kozmann, Orsolya Kiss, and Béla Merkely. Atrial fibrillation detection with and without atrial activity analysis using lead-i mobile ecg technology. *Biomedical Signal Processing and Control*, 66:102462, 2021.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

Patrick Wagner, Nils Strodthoff, Ralf-Dieter Bousseljot, Dieter Kreiseler, Fatima I Lunze, Wojciech Samek, and Tobias Schaeffter. Ptb-xl, a large publicly available electrocardiography dataset. *Scientific data*, 7(1):154, 2020.

Kuba Weimann and Tim OF Conrad. Transfer learning for ecg classification. *Scientific reports*, 11 (1):1–12, 2021.

Xi Yang, Aokun Chen, Nima PourNejatian, Hoo Chang Shin, Kaleb E Smith, Christopher Parisien, Colin Compas, Cheryl Martin, Anthony B Costa, Mona G Flores, et al. A large language model for electronic health records. *npj Digital Medicine*, 5(1):194, 2022.

Yatao Zhang, Junyan Li, Shoushui Wei, Fengyu Zhou, and Dong Li. Heartbeats classification using hybrid time-frequency analysis and transfer learning based on resnet. *IEEE Journal of Biomedical and Health Informatics*, 25(11):4175–4184, 2021.

A Appendix

A.1 Detailed Experimental Setup

A.1.1 Detailed Descriptions for Pretraining Datasets

- **MIMIC-III Matched Subset Dataset** includes 22,317 waveform records and 22,247 records for 10,282 distinct intensive care unit (ICU) patients. With a sample rate of 125 Hz, each record contains up to eight channels of signals such as ECG, arterial blood pressure (ABP), or PPG. It also contains additional contextual values such as respiration rate, SpO2, and blood pressure. We used the ECG lead-II signals for pre-training ECGBERT after resampling the signals to 250 Hz and splitting the signals into multiple 10-second segments. Given our limited computational resources, we randomly selected a subset of 13 records from the dataset. Correspondingly, within each selected record, we further employed a random sampling technique to select 3,000 segments to be used in the ECGBERT pre-training phase.
- **PTB-XL Dataset** contains 21,837 clinical 12 lead ECG signals from 18,869 patients with a length of 10 seconds. This dataset provides two different sampling rates at 100 Hz and 500 Hz and was annotated by two cardiologists, resulting in detailed diagnosis information across five different classes per record: Normal, Myocardial Infarction, ST/T Change, Conduction Disturbance, and Hypertrophy. During the pre-training phase, lead II ECG signals with a sampling rate of 500 Hz were chosen and then resampled to a rate of 250 Hz.
- **CPSC Dataset** has 6,877 12-lead ECG records with a 500 Hz sampling frequency. The duration of these records ranges from 6 to 60 seconds. The dataset contains nine different types of cardiac states, including atrial fibrillation (AFIB), ST-segment elevation (STE), ST-segment depression (STD), premature ventricular contraction (PVC), premature atrial contraction (PAC), normal heartbeat (Normal), left bundle branch block (LBBB), right bundle branch block (RBBB), and intrinsic paroxysmal atrioventricular block (IAVB). To facilitate the pre-training of the ECGBERT, any records exceeding a duration of 10 seconds were subdivided into multiple segments, each segment spanning 10 seconds.
- **Georgia Dataset** contains 20,672 ECG records where the duration of each record is between 5 and 10 seconds long with a 500 Hz sampling rate. After the records were resampled to 250 Hz, we only selected the 10-second ECG records for pre-training the ECGBERT.

A.1.2 Detailed Descriptions for Downstream Task Datasets

- **MIT-BIH Atrial Fibrillation Dataset** contains 23 ECG recordings which were recorded for 10 hours and consist of two ECG signals, ECG1 and ECG2, sampled at 250 Hz with 12-bit resolution over a range of 10 millivolts, respectively. This dataset contains five rhythm classes that are manually annotated: atrial fibrillation (AFIB), atrial flutter (AFL), AV junctional rhythm (J), and normal (N) rhythms. We label AFIB as one and the remaining labels as zero to make this a binary classification task. We excluded 5 out of the 25 recordings due to missing signals. Herein, we used the ECG1, corresponding to lead-2, with a resampled frequency of 250 Hz and split each ECG signal into data segments of 10 seconds, and annotated a label for each of the data segments based on the majority voting system. If a data segment contained more AFIB heartbeats in 10-second ECG signals, it is labeled as AFIB, otherwise, it was labeled as non-AFIB arrhythmia. Please note that other downstream task datasets follow the same resample rate and labeling process otherwise mentioned explicitly.
- **MIT-BIH Arrhythmia Dataset** includes various beat types from 48 records of 47 subjects between 1975 and 1979 from the BIH Arrhythmia Laboratory. ECG recording duration is 30 minutes with 360 Hz sampling rate and a bandpass filter is applied to the signal at 0.1 Hz – 100 Hz. In this study, we leveraged a modified limb lead-II signal, where each record has two channels (i.e., II or occasionally V5). Specifically, we split the data into multiple segments where each segment is 10 seconds of ECG signals and categorized different beat types into five classes, which are normal beat (N), ventricular ectopic beat (V), supraventricular ectopic beat (S), fusion beat (F), and unknown beat (Q) using a mixture of features.
- **Apnea-ECG Dataset** is comprised of 70 recordings, which are split into two sets: 35 records for the model training and another 35 for the testing. These recordings are available on the

PhysioNet website. The ECG signals were captured at a sample rate of 100 Hz, offering a 16-bit resolution with 200 A/D units per mV. Each record spans a duration of approximately seven to ten hours, and annotations are provided to indicate normal and apnea occurrences per minute. In this study, the data was segmented into multiple non-overlapping segments, with each segment representing a duration of 10 seconds.

-