Interpretation of Intracardiac Electrograms Through Textual Representations

William Jongwon Han   WJHAN@ANDREW.CMU.EDU
Carnegie Mellon University, USA

Diana Gomez   DGGOMEZ@ANDREW.CMU.EDU
Carnegie Mellon University, USA

Avi Alok   AAVI@ANDREW.CMU.EDU
Carnegie Mellon University, USA

Chaojing Duan   CHAOJIND@ANDREW.CMU.EDU
Allegheny Health Network, USA

Michael A. Rosenberg   MICHAEL.A.ROSENBERG@CUANSCHUTZ.EDU
University of Colorado, USA

Douglas Weber   DWEBER2@ANDREW.CMU.EDU
Carnegie Mellon University, USA

Emerson Liu   EMERSONLIU@MSN.COM
Allegheny Health Network, USA

Ding Zhao   DINGZHAO@ANDREW.CMU.EDU
Carnegie Mellon University, USA

Abstract

Understanding the irregular electrical activity of atrial fibrillation (AFib) has been a key challenge in electrocardiography. For serious cases of AFib, catheter ablations are performed to collect intracardiac electrograms (EGMs). EGMs offer intricately detailed and localized electrical activity of the heart and are an ideal modality for interpretable cardiac studies. Recent advancements in artificial intelligence (AI) has allowed some works to utilize deep learning frameworks to interpret EGMs during AFib. Additionally, language models (LMs) have shown exceptional performance in being able to generalize to unseen domains, especially in healthcare. In this study, we are the first to leverage pretrained LMs for finetuning of EGM interpolation and AFib classification via masked language modeling. We formulate the EGM as a textual sequence and present competitive performances on AFib classification compared against other representations. Lastly, we provide a comprehensive interpretability study to provide a multi-perspective intuition of the model's behavior, which could greatly benefit the clinical use.

Data and Code Availability We recorded intracardiac electrograms in the left atrium of two patients, one healthy and the other afflicted with atrial fibrillation (AFib), via an Octoray catheter from Biosense Webster Inc. The recording was done at Allegheny General Hospital in Pittsburgh, Pennsylvania. Our recorded data will not be released. We also experiment with the Intracardiac Atrial Fibrillation Database (Goldberger et al., 2000) accessible through this link: https://doi.org/10.13026/C23S33/. They used a decapolar catheter with 7mm spacing between bipoles to record the right atria from 8 patients in atrial fibrillation or flutter. More details about the data is provided in Section 4. We have released the code at the following link: https://github.com/willxxy/Text-EGM.

Institutional Review Board (IRB) An official IRB document has been received from Allegheny Health Network deeming our work with the collected data as exempt status under the Code of Regulations: 45 CFR 6.104 (d) Exempt 4. The Intracardiac Atrial Fibrillation Database is publicly available and de-identified, thus not needing an IRB approval.

1. Introduction

Atrial fibrillation (AFib) is one of the most common types of arrhythmia globally, affecting more than 60 million people over the last thirty years (Elliott et al., 2023). AFib is characterized by the chaotic, irregular, and often rapid, beating pattern of the heart's upper chambers, known as the atria. This causes the atria to become out of sync with the heart's lower chambers, called the ventricles. Due to these debilitating properties, people with AFib are at risk of stroke, heart failure, and a number of other cardiac diseases.

Catheter ablations are performed for patients with more serious cases of AFib, during which intracardiac electrograms (EGMs) representative of electrical activity along the endocardial surface of the atria are collected by means of navigational catheters embedded with multiple electrodes. Because EGMs have distinct and spatially specific signatures for different types of cardiac arrhythmia, they are the ideal modality for highly interpretable cardiac studies, in particular for AFib.

Due to their locality and richness in information, there are works utilizing deep learning methods with EGMs for AFib classification (Alhusseini et al., 2020), patient-specific therapy (John et al., 2022), and surface electrocardiogram (ECG) reconstruction (Zhang et al., 2021). In addition, the representation of EGMs are carefully considered (Alhusseini et al., 2020; Tang et al., 2022) and are most commonly represented as visual or time series modalities (Kong et al., 2023; Zhang et al., 2021; Alhusseini et al., 2020). Despite the impressive performances of these EGM algorithms, the model's interpretability is a critical component that requires further exploration before deployment in the clinical setting. Although some works attempt to interpret the model's decisions, to our best knowledge, only a uniperspective interpretability metric is provided, such as visualizing the Grad-CAM heatmap (Alhusseini et al., 2020), which can often be misleading (Ribeiro et al., 2016; Donoso-Guzmán et al., 2023).

In this study, we introduce a tokenization schema, inspired by Chen et al. (2022), that represents an EGM as a textual sequence. The tokenization process discretizes continuous EGM signal amplitudes and maps each to a unique token ID, similarly to textual data. Representing EGM signals as a textual sequences allows us to leverage powerful Language Models (LMs) pretrained on billions of texts. Specifi-

cally, we use a Masked Language Model (MLM) that is optimized for predicting randomly masked positions in a given sequence. As seen in Figure 1, the AFib label is also tokenized input. By masking 75% of the signal tokens and the AFib label token, the model is able to interpolate the masked EGM values and simultaneously classify the EGM as AFib or not. Additionally, we provide a multi-perspective interpretability procedure to expose the intuition behind the model's decisions at the token level.

In summary, our main contributions are the following:

- To our best knowledge, this is the first work to represent EGMs as a textual sequence. We introduce an effective tokenization schema that maintains the low level information of the original signal.

- We utilize a MLM pretrained on **textual data** to finetune for interpreting EGM signals by interpolation and classification for AFib.

- We also perform a comprehensive interpretability procedure via attention maps, integrated gradients (Sundararajan et al., 2017), and counterfactual analysis to provide clarity of the model's decisions for clinicians.

2. Related Work

2.1. Language Models for Healthcare

Language Models (LMs) have been an extremely popular medium for many different downstream tasks in healthcare. Qiu et al. (2023a) utilizes LMs to translate surface ECG signals into descriptive clinical notes. Choi et al. (2023) first cluster surface ECG signals into 70 groups and uses them to create a wave vocabulary for pretraining a MLM. They then employ the pretrained MLM model for various downstream tasks, such as AFib classification, heartbeat classification, and user identification (Choi et al., 2023). There has also been large efforts in pretraining LMs on solely clinical text, such as Alsentzer et al. (2019) and Li et al. (2022). Mehandru et al. (2023) observed that as the role of LMs become larger in the medical world, they can act as agents by providing clinical decision support and stakeholder interactions. Therefore, they propose new evaluation frameworks, termed Artificial-intelligence Structured Clinical Examinations (AI-SCI), to assess LMs in real-world clinical tasks (Mehandru et al., 2023).
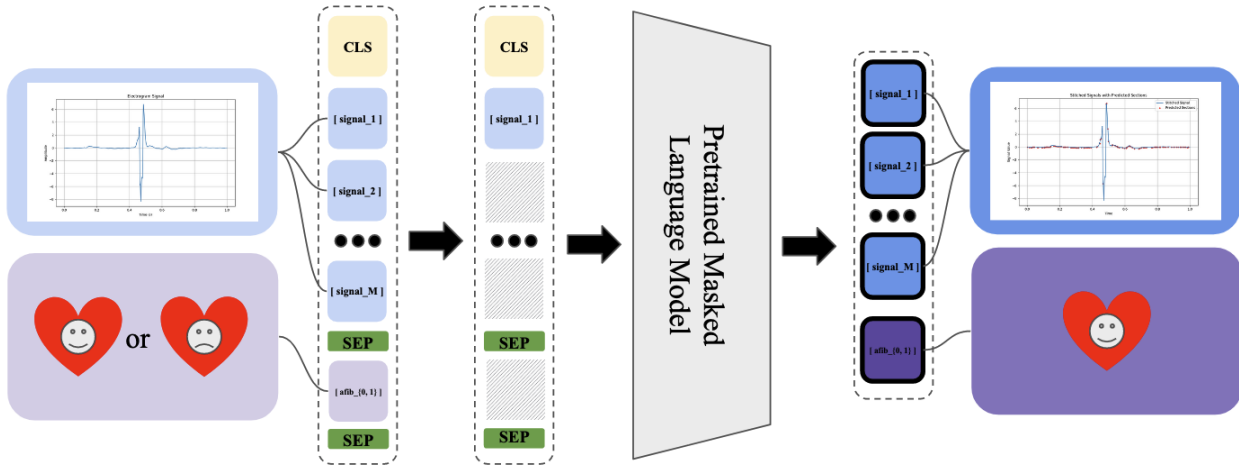
Figure 1: The overall pipeline of our model. We formulate EGM signal interpolation and AFib classification as a masked language modeling task. ▨ denotes the mask.

2.2. Machine Learning for EGM

Machine learning in electrogram (EGM) analysis, though limited, has seen notable advancements. Alhusseini et al. (2020) used a 64-electrode catheter to create 8x8 spatial activation heatmaps from atrial EGM signals, feeding them into a CNN for atrial fibrillation (AFib) classification. Zhang et al. (2021) developed RT-RCG, a neural network tailored for ECG data reconstruction from EGMs, incorporating a Differentiable Acceleration Search for efficient hardware accelerator optimization. Duque et al. (2017) presented a genetic algorithm and K-NN based classifier to categorize EGM signals, enhancing AFib ablation therapy guidance. Tang et al. (2022) employed a multimodal approach, integrating surface ECG, EGM signals, and clinical data in a CNN, improving AFib classification accuracy. Our study extends these works by applying LMs, specifically MLMs, pretrained on general text to interpret complex EGM signals in AFib cases.

2.3. Interpretability for Machine Learning in Healthcare

Interpretability in healthcare machine learning is crucial but challenging for clinical use (Amann et al., 2020). Jin et al. (2021) outline four attribution methods for machine learning interpretability: backpropagation, feature perturbation, attention, and model distillation. Notably, Alhusseini et al. (2020) and Vaid et al. (2022) use Grad-CAM heatmaps and attention saliency maps, respectively, to interpret CNN and Vision Transformer models. Shi and Norgeot (2022) advocates for interpretability in healthcare models, with Sanchez et al. (2022) and Prosperi et al. (2020) emphasizing counterfactual analysis and addressing biases in machine learning. Our study merges several interpretability methods (e.g., attention, integrated gradients, counterfactual analysis) for a more holistic understanding of model decisions.

3. Methods

Our method consists of three main steps: (1) translating the EGM signals from time series to text through tokenization, (2) utilizing MLMs to predict the masked portions of the input sequence, and (3) interpreting the decisions of the models through attention weights, integrated gradients, and counterfactual analysis.

3.1. Tokenization

Our tokenization approach, inspired by Chen et al. (2022), discretizes the continuous amplitudes of EGM signals to create a textual representation of the time series information. In this section, we intricately explain the process of tokenization for a single sequence. Let $S = \{s_1, s_2, \ldots, s_M\} \in \mathbb{R}^M$ be a continuous EGM signal, where $M = 1000$ for a one second signal.

The normalization of S scales the amplitude values to $[0, 1]$. For the sequence S and each s_i, this is given

by:

$$S_{\text{norm}} = \left\{ \frac{s_i - s_{\min}}{s_{\max} - s_{\min}} : s_i \in S \right\}$$

where $s_{\min} = \min(S)$, $s_{\max} = \max(S)$, and S_{norm} is the normalized sequence.

Quantization converts S_{norm} into discrete levels V, where $V = 250$. The quantized sequence S_{quant} is:

$$S_{\text{quant}} = \{\lfloor S_{\text{norm}} \times V \rfloor : s_i \in S_{\text{norm}}\}$$

where $\lfloor \cdot \rfloor$ denotes the floor function. The decision of the value of V were purposely chosen through experimentation. We observed that 250 was the minimum amount of discrete levels that represents the fine-grained details of an EGM signal as well as obtain the best performance for the signal interpolation task.

We then map S_{quant} to a unique token ID and expand the tokenizer's embedding table by adding the new tokens. For signals, we define:

$$T_{\text{S}} = \{\text{``signal\_''} + \text{str}(q) : q \in S_{\text{quant}}\}$$

where $\text{str}(\cdot)$ converts numbers to strings. We also create the token IDs for the AFib label by introducing the following tokens:

$$T_{\text{A}} = \{\text{``afib\_''} + \text{str}(a) : a \in 0, 1\}$$

where 0 and 1 denotes a normal and AFib signal respectively.

3.2. Masked Language Model

The Masked Language Models (MLMs) we use are BigBird (Zaheer et al., 2021), LongFormer (Beltagy et al., 2020), Clinical BigBird (Li et al., 2022), and Clinical LongFormer (Li et al., 2022), which are optimized to handle sequences of lengths up to 4,096, for finetuning for EGM signal interpolation and AFib classification. Clinical BigBird and Clinical Longformer uses the same architecture as their predecessors, however, they are pretrained on 2 million clinical notes from the MIMIC-III dataset (Johnson et al., 2016) instead of a general web-crawled text dataset. More details on the models are available in the appendix.

3.3. Masking Strategy

To prepare our input for the MLM, we adopt a high masking strategy. We randomly mask out 75% of T_S. We choose 75%, as suggested in He et al. (2021), to ensure the model learns a significant portion of

the EGM waveform and to diminish the possibility of interpolating from the surrounding unmasked areas. We always mask out T_A to ensure the model is not being fed the classification label. After applying the masks, our input is formulated as follows:

$$T_I = [\text{CLS}] \oplus T_S \oplus [\text{SEP}] \oplus T_A \oplus [\text{SEP}]$$

where $\oplus$ is the concatenation operation, [CLS] is the classification token indicating the start of the sequence, and [SEP] is the separator token denoting both the end of a sequence as well as distinguishing boundaries for different types of input (Devlin et al., 2019).

3.4. Learning Objectives

Masked Language Model Loss Let M be the set of indices of the masked tokens in T_I. For each masked token index $i \in M$, the original token at this position is t_i. The MLM objective is to minimize the following loss function:

$$L_{MLM}(\theta) = -\sum_{i \in M} \log P(t_i | T_I; \theta),$$

where θ represents the parameters of the MLM model, $P(t_i | T_I; \theta)$ is the probability predicted by the model for the token t_i given the masked input sequence T_I, and the sum is over all masked tokens in the sequence.

Atrial Fibrillation Classification Loss We add another loss L_{AFib} based on the prediction of T_A within the sequence T_I. The cross-entropy loss for this classification task is computed as follows:

$$L_{AFib}(\theta) = -\sum_{c=1}^{C} y_c \log(p_c),$$

where C is the number of classes for AFib classification, y_c is a binary indicator (0 or 1) if class label c is the correct classification for T_A, and p_c is the predicted probability for the token T_A being of class c.

Final Learning Objective The total loss for the model is given by:

$$L(\theta) = \alpha_1 L_{MLM}(\theta) + \alpha_2 L_{AFib}(\theta),$$

where $\alpha_i \in [0, 1], i \in \{1, 2\}$.

3.5. Model Interpretability

All of our model interpretability procedures have been conducted using the BigBird (Zaheer et al., 2021) model with the final learning objective.

Attention Visualizing the attention map is an extremely common technique for interpretability in Transformers (Vaswani et al., 2023). There has been arguments in the interpretability community that debate whether attention maps are reliable or not (Jain and Wallace, 2019; Wiegreffe and Pinter, 2019). The general consensus seems to be that although attention can be informative, it can also be misleading if it is the only metric of interpretability (Wen et al., 2023). Therefore, we visualize the attention maps of our models, while providing other perspectives as well. In our study, we visualize the averaged attention map over all heads and layers.

Integrated Gradients Integrated Gradients (IG), introduced by (Sundararajan et al., 2017), is a method for attributing a model's prediction to its input features, providing insight into the model's behavior. Given an input x and a baseline x', IG computes the gradient of the model's output with respect to the input along the path α from the baseline to the input. The attribution of each feature i is given by:

$$IG_i(x) = (x_i - x_i') \times \int_{\alpha=0}^{1} \frac{\partial F(x' + \alpha \times (x - x'))}{\partial x_i} d\alpha$$

where F is the model function. IG satisfies two key axioms: Sensitivity, ensuring non-zero attribution for features that change the output, and Implementation Invariance, guaranteeing consistent attributions for functionally equivalent models. In our setting, the input x is our masked, tokenized sequence, T_I. The baseline x' is defined to be a vector of padding tokens [PAD] of the same dimension as T_I.

Counterfactual Analysis Counterfactual analysis observes the change of the predicted outcome by directly manipulating the input (Feder et al., 2021; Pearl, 2009). We individually conduct three different counterfactual analysis methods: Token Substitution, Token Addition, and Label Flipping. We view the effects of the introduced counterfactuals in two settings. The first setting is where we finetune the MLM without the counterfactuals and inference with modified and unmodified inputs. The second setting is where we finetune the MLM with the counterfactuals alongside unmodified inputs, and inference on both modified and unmodified inputs. In both settings, we randomly choose 25% of the batch to be modified with the counterfactuals.

Token Substitution Given our input signal T_S, we apply token substitution by replacing the original

sequence with a smoothed out version via a moving average filter. The intuition behind this is aligned from the clinicians' perspective such that the sharp oscillations in a given signal is where most of the information is stored for the model or clinician to determine whether it is a normal heartbeat or AFib. By smoothing out the oscillations via a moving average filter, we want to observe the model's robustness against this counterfactual as well as its attention and gradients.

Token Addition Given our input signal T_S, we apply token addition by introducing the following V new augmentation signal tokens and adding them to tokenizer's embedding table:

$$T_{\text{AUG}} = \{\text{"augsig\_"} + \text{str}(q) : q \in S_{\text{quant}}\}$$

Please note that T_{AUG} is still derived from the same S_{quant} as T_S. We then randomly choose 25% (a 250 length segment) of T_{AUG} and append it to T_S therefore getting the input sequence

$$T_I' = [\text{CLS}] \oplus T_S \oplus T_{\text{AUG}} \oplus [\text{SEP}] \oplus T_A \oplus [\text{SEP}]$$

These augmented tokens will not be masked. The ground truth label to these augmented tokens will be the padding token [PAD]. Essentially, the augmented tokens will simply serve as noise to T_S to observe whether the model leverages the noise or still attends to the meaningful, non-augmented portion of the signal.

Label Flipping Given our masked input AFib label T_A, we apply label flipping by replacing the ground truth label with its opposite. This is used to understand the model's behavior under adversarial examples.

Please note that the top section of the table are results reported on different datasets used by the corresponding authors. We present these results to show the current state of AFib classification of EGMs with deep learning.

4. Experiments

4.1. Dataset and Preprocessing

Our dataset consists of 20 different placements of catheter ablation for a patient with normal heartbeat rhythm and 45 different placements of catheter ablation for a patient with AFib. The catheter ablation was performed with the Octoray catheter from Biosense Webster Inc. The catheter consists of 8 splines with 6 electrodes on each spline, summing up

Table 1: AFib classification results with different baselines and representations.

| Method	Representation	Sensitivity %	Specificity %	PPV %	NPV %	Accuracy %
CatBoost (Tang et al., 2022)	Time Series	88.5	62.7	-	-	70.1
CNN (Alhusseini et al., 2020)	Image	97.0	93.0	93.1	97.0	95.0
K-Means (Alhusseini et al., 2020)	Image	77.0	82.3	83.5	75.5	79.4
KNN (Alhusseini et al., 2020)	Image	75.3	84.0	87.1	70.3	78.9
LDA (Alhusseini et al., 2020)	Image	85.0	74.6	76.4	83.7	79.7
SVM (Alhusseini et al., 2020)	Image	82.9	76.7	77.4	82.3	79.7
ViT (Dosovitskiy et al., 2021)	Image	**1.0**	99.1	99.1	**1.0**	**99.7**
BEiT (Bao et al., 2022)	Image	99.9	99.3	98.8	99.9	99.5
BigBird (Zaheer et al., 2021)	Time Series	0.0	1.0	0.0	63.6	63.6
LongFormer (Beltagy et al., 2020)	Time Series	0.0	1.0	0.0	63.6	63.6
BigBird (Zaheer et al., 2021)	Text	97.8	96.1	93.5	98.7	96.7
LongFormer (Beltagy et al., 2020)	Text	69.0	98.4	96.1	84.7	87.7
Clinical BigBird (Li et al., 2022)	Text	12.8	97.3	72.8	66.1	66.6
Clinical LongFormer (Li et al., 2022)	Text	85.2	97.9	95.8	92.1	93.3
Ours BigBird (Zaheer et al., 2021)	Text	96.7	**99.8**	99.6	98.2	99.2
Ours LongFormer (Beltagy et al., 2020)	Text	**1.0**	99.4	99.0	99.8	99.5
Ours Clinical BigBird (Li et al., 2022)	Text	99.0	99.0	98.1	99.4	99.0
Ours Clinical LongFormer (Li et al., 2022)	Text	99.9	99.6	**99.9**	99.3	**99.7**
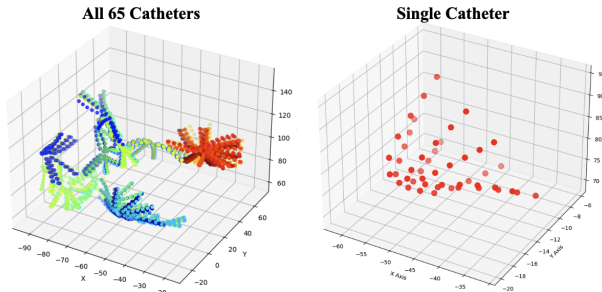
Figure 2: Visualization of all 65 (left) Octoray catheters and a single Octoray Catheter (right) placements inside the left atrium.

to 48 electrodes. The catheter ablation performed on the patient with AFib was for 30 seconds and for the patient with a normal heartbeat, the ablation was for 29 seconds. Both studies were sampled at a rate of 1000 Hz. It is important to note that our dataset is from two patients. However, this limitation is overcome by the diversity of catheter placement locations as seen in Fig 2. We treat each electrode on all 65 catheters placements as a separate instance. In summary, we have a total number of $59 * 48 * 65 = 184,080$ unique samples we finetune on.

Additionally, we finetune and inference on the Intracardiac Atrial Fibrillation Database (Goldberger et al., 2000) to show the generalizability of our method to external datasets. This dataset consists of endocardial recordings from the right atria of 8 patients in atrial fibrillation or flutter (Goldberger et al., 2000). They utilize a decapolar catheter with a 7 millimeter spacing between bipoles at a sampling rate of 1000 Hz. Four separate catheter placements of the heart were recorded, and for each region, 5 bipolar signals were recorded.

Let $\mathbf{X} \in \mathbb{R}^{I \times J \times K}$ represent the normal and AFib data tensor, with dimensions I, J and K denoting the time length, number of electrodes, and number of catheter placements, respectively. Z-score normalization is applied across the recording time and electrodes for each placement k. The normalized tensor $\mathbf{Z}$ is computed as:

$$Z_{i,j,k} = \frac{X_{i,j,k} - \mu_k}{\sigma_k}$$

where μ_k and σ_k are the mean and standard deviation of the elements in $\mathbf{X}$ across the first two dimensions for each k. We then segment $Z_{i,j,k}$ into a sequence of non-overlapping segments of length M, arriving at $S = \{s_1, s_2, \ldots, s_M\}$. Unless specified otherwise, we report the results where $M = 1000$.

4.2. Experimental Setting

We finetuned the masked language model utilizing the AdamW optimizer (Kingma and Ba, 2017) with a learning weight of $1e-4$, weight decay rate of $1e-2$

Table 2: Results on inferencing on the external dataset (Goldberger et al., 2000) where we finetuned with our dataset (✓) or external dataset (✗) for AFib Classification and interpolation.

Model	Finetuned	Interpolation		AFib Classification
		MSE ↓	MAE ↓	Accuracy %
Ours BigBird Zaheer et al. (2021)	✓	0.77	0.29	80.3
Ours LongFormer Beltagy et al. (2020)	✓	**0.75**	0.29	74.08
Ours Clinical BigBird Li et al. (2022)	✓	0.80	0.23	**81.8**
Ours Clinical LongFormer Li et al. (2022)	✓	**0.75**	**0.21**	73.83
Ours BigBird Zaheer et al. (2021)	✗	**0.86**	0.30	99.8
Ours LongFormer Beltagy et al. (2020)	✗	1.02	0.25	99.8
Ours Clinical BigBird Li et al. (2022)	✗	**0.86**	**0.24**	**99.9**
Ours Clinical LongFormer Li et al. (2022)	✗	1.02	0.26	99.7

(Loshchilov and Hutter, 2017). We conducted all experiments with a batch size of 8 and 1 for finetuning and inference, respectively. For our final learning objective L, $\alpha_1 = \alpha_2 = 1$. Our experiments were conducted on 2 NVIDIA A6000 and 2 NVIDIA A5000 GPUs.

During inference, we evaluate our models on two tasks, namely EGM signal interpolation and AFib classification. For EGM signal interpolation, we use the Mean Squared Error and Mean Absolute Error. For AFib classification, we use the sensitivity, specificity, positive predicted values (PPV), negative predicted values (NPV), and accuracy metrics for evaluation. Additionally, all reported results are utilizing our own dataset unless specified otherwise.

5. Results

5.1. AFib Classification

We compare our results with different representations and baselines in Table 1. We divide our table into three sections. The top section contains prior works that do AFib classification with EGMs. It is important to note that the datasets used in this top section are **different from ours** (i.e., different patients, catheter, sampling rate), therefore they are not directly comparable. However, we decide to include them in this table to report the current state of AFib classification using EGMs.

The middle section contains the results from different base models and representations with our dataset. For ViT (Dosovitskiy et al., 2021) and BEiT (Bao et al., 2022), we transform the EGMs into image representations via Markov Transition Field (MTF), Gramian Angular Field (GAF) (Wang and

Oates, 2014), and Recurrence Plot (RP) (Eckmann et al., 1987), which have proven to be excellent image representations of time series signals (Qiu et al., 2023b). We load in the pretrained weights of ViT and BEiT, 'google/vit-base-patch16-224-in21k' and 'microsoft/beit-base-patch16-224-pt22k' respectively, and apply the 75% masking rate on the images during finetuning. For the time series representation results using LongFormer (Beltagy et al., 2020) and BigBird (Zaheer et al., 2021), we simply utilize the normalized signal amplitudes as input embeddings by projecting them through a linear layer before passing them into their respective models. For the text representation results, we report the results for only utilizing the pretrained base model with masked language model loss.

In the bottom section of the table, we report our results for utilizing the textual representation and full learning objective. We can observe that our method of representing the complex EGM signal as a simple textual sequence achieves competitive results with the image modality.

Lastly, we report the results of AFib classification on the Intracardiac Atrial Fibrillation Database (Goldberger et al., 2000) in Table 2. Table 2 shows the AFib classification results under two settings: 1) we finetune on our own collected dataset and inference on the Intracardiac Atrial Fibrillation Database (Goldberger et al., 2000) and 2) we finetune and inference on the external dataset. Observably, simply inferencing on the external dataset after finetuning on our dataset yields competitive results, with Clinical BigBird (Li et al., 2022) achieving 81.8% accuracy. When we finetune our method on the exter-

nal dataset, we can see that our method achieves near perfect results, with Clinical BigBird (Li et al., 2022) achieving 99.9% accuracy. These results emphasize the generalizability of our method to external datasets under the two settings.

Table 3: Results on interpolation.

Method	Interpolation	
	MSE ↓	MAE ↓
BigBird (Zaheer et al., 2021)	0.44	0.16
LongFormer (Beltagy et al., 2020)	0.37	**0.13**
Clinical BigBird (Li et al., 2022)	0.50	0.17
Clinical LongFormer (Li et al., 2022)	**0.36**	**0.13**
Ours BigBird (Zaheer et al., 2021)	0.80	0.37
Ours LongFormer (Beltagy et al., 2020)	0.40	0.14
Ours Clinical BigBird (Li et al., 2022)	0.44	0.15
Ours Clinical LongFormer (Li et al., 2022)	0.40	0.14

5.2. EGM Interpolation

We report the results of the EGM interpolation task on our own dataset and the external dataset (Goldberger et al., 2000) in Table 3 and Table 2, respectively. We visualize the comparison of the reconstructed (blue) and ground truth (yellow) signal in the appendix. From Table 3, we can see the models with only the MLM learning objective opposed to our final learning objective achieves slightly lower MSE and MAE scores. We can see a clear trade-off between the interpolation results and the AFib classification results in adding the Cross Entropy (CE) loss on top of the MLM loss. However, this trade-off is acceptable due to the increase in performance for AFib classification.

5.3. Ablation Study

Pretrained vs Non-Pretrained We observe the effects of utilizing pretrained vs non-pretrained versions of the MLM in Table 4. For clarification, pretrained means we load in the checkpoint that was pretrained on the large text corpus and conduct finetuning. Conversely, non-pretrained means we do not load the checkpoint that was pretrained on the text corpus and conduct finetuning. Although the BigBird model (Zaheer et al., 2021) and LongFormer model (Beltagy et al., 2020) were pretrained on a text corpora that was not directly related to clinical subjects, we see that the pretrained setting greatly outperforms the non-pretrained setting. From this, we can observe the characteristic of pretrained models being able to leverage the language knowledge to well adapt to unseen domains.

5.4. Interpretability

In our study, we analyzed AFib EGM signal attention maps and attribution scores during inference (Figure 3), examining both before and after finetuning, and under different masking conditions (T_A Masked, $T_A + T_S$ Masked).

Attention maps pre-finetuning showed slightly more oscillation compared to post-finetuning, though the overall distribution remained largely unchanged in both settings and masking conditions. Notably, the model focused more on the start and end of sequences, a behavior attributed to its pretraining on natural language data, where salient information is often found at these positions. This tendency seems to have generalized to EGM data.

Regarding attribution scores, a marked difference pre- and post-finetuning was observed in score stability. Pre-finetuning, scores were random across the signal when masking only the T_A token and somewhat concentrated at the start but still randomly distributed with our original masking strategy. Post-finetuning, the distribution changed significantly. With only the T_A token masked, the scores peaked around signal oscillations, aligning with clinical significance. For full masking, scores were evenly distributed across the sequence, with notable attribution at the start, reflecting the MLM's use of the entire sequence for contextual understanding, including predicting the T_A token. We visualize more random samples of attribution scores overlayed on both normal and AFib EGMs in the appendix.

5.5. Counterfactual Analysis

We first analyzes model performance with three counterfactuals (Token Substitution, Token Addition, Label Flipping) across two finetuning settings (i.e., with/without counterfactuals) and three representations (i.e., text, image, time series), using ViT and BigBird architectures for image and time series respectively. Textual representations exhibit higher robustness to counterfactuals in interpolation tasks compared to others.

Attribution scores from checkpoints finetuned on respective counterfactuals were compared against non-counterfactual methods (Figure 4). For Token Substitution, full masking results in typical attribution score distributions, while partial masking (T_A token) shows random distributions, indicating a forced learning of a generalized signal distribution. Token Addition demonstrates a focused attribution on oscillating signal parts when T_A is masked, reflect-

Table 4: Ablation study on pretrained (✓) vs non-pretrained (✗) MLM.

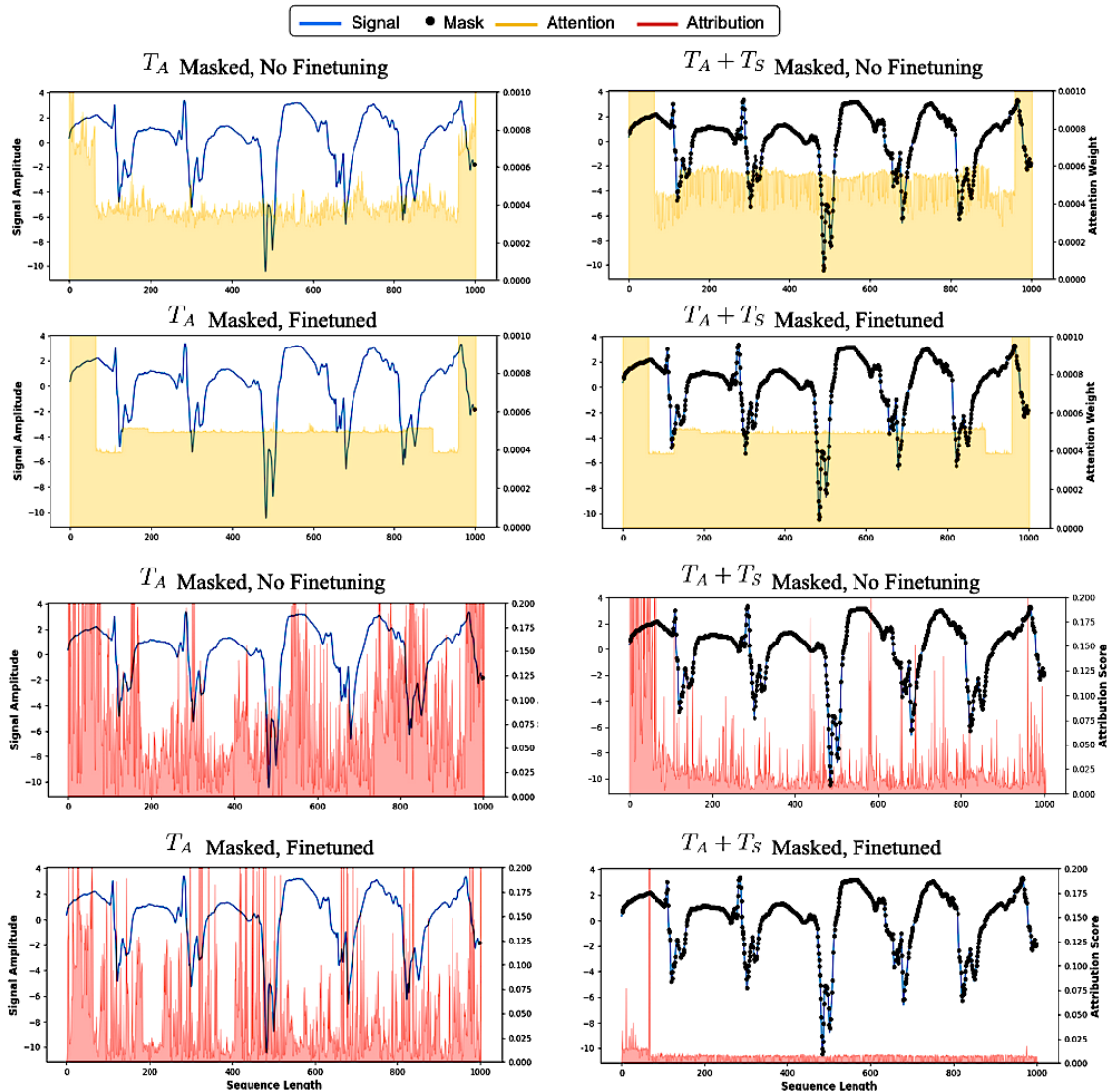| Model | Pretrained | Interpolation | | AFib Classification |
		MSE ↓	MAE ↓	Accuracy %
BigBird Zaheer et al. (2021)	✗	5.42	1.42	63.6
BigBird Zaheer et al. (2021)	✓	**0.44**	**0.16**	**96.7**
LongFormer Beltagy et al. (2020)	✗	6.32	2.21	60.9
LongFormer Beltagy et al. (2020)	✓	0.37	0.13	87.7

Figure 3: The averaged attention weights (yellow) and attribution scores (red) of an AFib EGM signal during inference. 'T_A Masked' refers to running inference when only the T_A token is masked. '$T_A + T_S$ Masked' denotes our regular paradigm of masking out 75% of the T_S tokens and always masking out the T_A token.
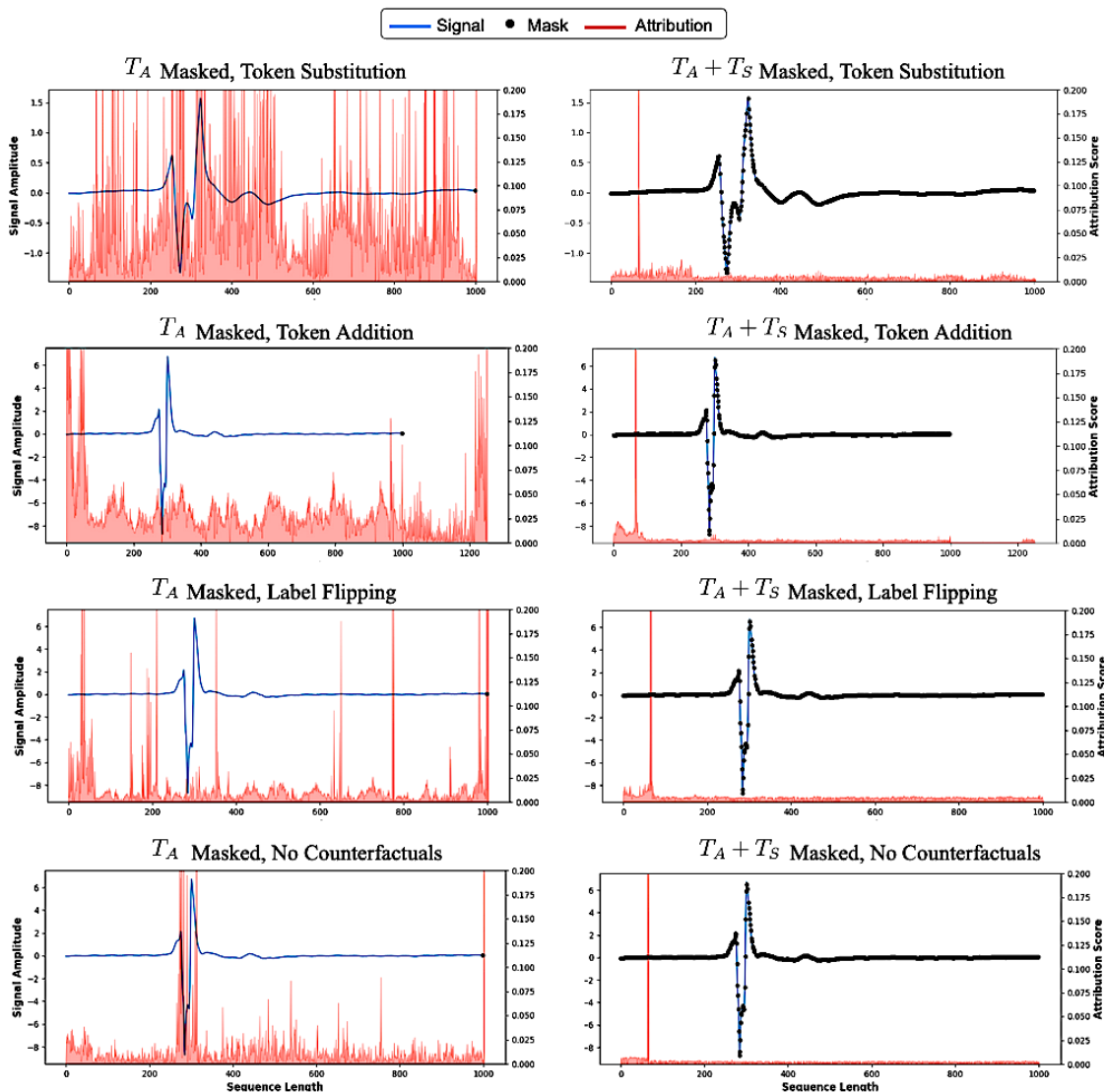
Figure 4: The averaged attribution scores for checkpoints that have been finetuned on their respective counterfactuals. We also include the visualization of the average attribution score for the checkpoint that has been finetuned without counterfactuals for comparison.

ing model adaptation to identify key sequence elements amidst noisy tokens, enhancing interpolation and AFib classification (see Table 5).

For Label Flipping, the attribution scores under full masking align with our standard method, indicating robustness to adversarial examples. However, partial masking shows arbitrary peaks across the sequence, diverging from the focused attribution of our non-counterfactual method. This suggests inherent

model resilience against incorrect labels without specific adversarial training objectives.

6. Discussion and Conclusion

In this study, we demonstrated the effectiveness of using pretrained MLMs with textual representations for EGM signal interpretation. Our approach yielded impressive results in AFib classification and interpolation on our own internal dataset as well as the external dataset (Goldberger et al., 2000), achieving 99.7%

Table 5: Results for finetuning with (✓) and without (✗) Token Substitution, Token Addition, and Label Flipping, where we highlight the best results for each counterfactual in **orange**, **blue**, and **red**, respectively.

Representations	Counterfactuals	Finetuned	Interpolation MSE ↓	MAE ↓	AFib Classification Accuracy %
Image	Token Substitution	✓	0.57	0.46	**99.4**
		✗	0.72	0.55	**99.0**
	Token Addition	✓	1.12	0.64	**99.1**
		✗	1.92	1.24	73.8
	Label Flipping	✓	2.31	1.12	74.0
		✗	2.51	2.08	74.5
Time Series	Token Substitution	✓	11.63	6.40	37.1
		✗	11.96	6.34	36.4
	Token Addition	✓	8.81	5.10	54.2
		✗	9.92	5.50	50.7
	Label Flipping	✓	7.22	4.51	56.9
		✗	10.41	9.56	56.1
Text	Token Substitution	✓	**0.29**	**0.26**	96.3
		✗	**0.25**	**0.25**	74.7
	Token Addition	✓	**0.70**	**0.34**	99.0
		✗	**0.61**	**0.19**	**77.4**
	Label Flipping	✓	**0.47**	**0.17**	**93.5**
		✗	**0.81**	**0.38**	**91.3**

accuracy, 0.40 MSE, and 0.14 MAE, and 99.9% accuracy, 0.86 MSE, and 0.24 MAE, respectively. We compared the use of a general text corpus pretrained MLM versus a non-pretrained MLM for EGM interpretation. The pretrained model significantly outperformed the non-pretrained model in both AFib classification and interpolation, showing a 38.8% increase in accuracy and decreases of 5.92 in MSE and 2.07 in MAE. Our analysis included visualizing the attention and attribution scores. While attention weights were not insightful regarding the model's decision-making process, the attribution scores under full masking indicated substantial reliance on pretrained knowledge for interpreting EGM signals. Interestingly, partial masking revealed the model's focus on oscillating signal parts for AFib classification, aligning with clinical interpretations. Further, we demonstrated our method's robustness by finetuning it with predefined counterfactuals, which showed superior performance over other representations. Analyzing the attribution scores of these finetuned models revealed distinct attribution distributions for each counterfactual. The distinct behaviors observed in the attribution scores of these finetuned models not only confirm the adaptability of our approach but also open avenues for future research. Specifically, these findings encour-age further exploration into optimizing the model's focus on critical signal segments, maintaining accuracy even when confronted with varying counterfactual scenarios. Additionally, this work not only contributes to the field of EGM signal interpretation via deep learning but also sets a foundation for future studies aimed at expanding the capabilities of LM-based approaches in EGM analysis.

Limitations While our study highlights the potential of utilizing general pretrained MLMs for interpreting complicated EGM signals during AFib, there are notable limitations worth mentioning. The most obvious limitation of this study is the data. Due to the invasive and complex nature of collecting intracardiac EGM data, it is difficult to find and conduct comparable baselines. However, we try to mitigate this limitation in our work by conducting several baseline models with different representations and external datasets, namely the Intracardiac Atrial Fibrillation Database (Goldberger et al., 2000).

Acknowledgments

References

Mahmood I. Alhusseini, Firas Abuzaid, Albert J. Rogers, Junaid A.B. Zaman, Tina Baykaner, Paul Clopton, Peter Bailis, Matei Zaharia, Paul J. Wang, Wouter-Jan Rappel, and Sanjiv M. Narayan. Machine learning to classify intracardiac electrical patterns during atrial fibrillation. *Circulation: Arrhythmia and Electrophysiology*, 13, 08 2020. doi: 10.1161/circep.119.008160.

Emily Alsentzer, John R. Murphy, Willie Boag, Wei-Hung Weng, Di Jin, Tristan Naumann, and Matthew B. A. McDermott. Publicly available clinical bert embeddings, 2019.

Julia Amann, Alessandro Blasimme, Effy Vayena, Dietmar Frey, and Vince Madai. Explainability for artificial intelligence in healthcare: a multidisciplinary perspective. *BMC Medical Informatics and Decision Making*, 20, 11 2020. doi: 10.1186/s12911-020-01332-6.

Hangbo Bao, Li Dong, Songhao Piao, and Furu Wei. Beit: Bert pre-training of image transformers, 2022.

Iz Beltagy, Matthew E. Peters, and Arman Cohan. Longformer: The long-document transformer, 2020.

Ting Chen, Saurabh Saxena, Lala Li, David J. Fleet, and Geoffrey Hinton. Pix2seq: A language modeling framework for object detection, 2022.

Seokmin Choi, Sajad Mousavi, Phillip Si, Haben G. Yhdego, Fatemeh Khadem, and Fatemeh Afghah. Ecgbert: Understanding hidden language of ecgs with self-supervised representation learning, 2023.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding, 2019.

Ivania Donoso-Guzmán, Jeroen Ooge, Denis Parra, and Katrien Verbert. Towards a comprehensive human-centred evaluation framework for explainable ai, 2023.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale, 2021.

S.I. Duque, A. Orozco-Duque, V. Kremen, D. Novak, C. Tobón, and J. Bustamante. Feature subset selection and classification of intracardiac electrograms during atrial fibrillation. *Biomedical Signal Processing and Control*, 38:182–190, 2017. ISSN 1746-8094. doi: https://doi.org/10.1016/j.bspc.2017.06.005. URL https://www.sciencedirect.com/science/article/pii/S1746809417301088.

Jean-Pierre Eckmann, Sylvie Kamphorst, and D. Ruelle. Recurrence plots of dynamical systems. *Europhysics Letters (epl)*, 4:973–977, 11 1987. doi: 10.1209/0295-5075/4/9/004.

Adrian D. Elliott, Melissa E. Middeldorp, Isabelle C. Van Gelder, Christine M. Albert, and Prashanthan Sanders. Epidemiology and modifiable risk factors for atrial fibrillation. *Nature Reviews Cardiology*, page 1–14, 01 2023. doi: 10.1038/s41569-022-00820-8. URL https://www.nature.com/articles/s41569-022-00820-8#ref-CR1.

Amir Feder, Nadav Oved, Uri Shalit, and Roi Reichart. Causalm: Causal model explanation through counterfactual language models. *Computational Linguistics*, page 1–54, May 2021. ISSN 1530-9312. doi: 10.1162/coli_a_00404. URL http://dx.doi.org/10.1162/coli_a_00404.

Ary L. Goldberger, Luis A. N. 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. *Circulation*, 101, 06 2000. doi: 10.1161/01.cir.101.23.e215.

Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners, 2021.

Sarthak Jain and Byron C. Wallace. Attention is not explanation, 2019.

Di Jin, Elena Sergeeva, Wei-Hung Weng, Geeticka Chauhan, and Peter Szolovits. Explainable deep learning in healthcare: A methodological survey from an attribution view, 2021.

Mathews M. John, Anton Banta, Allison Post, Skylar Buchan, Behnaam Aazhang, and Mehdi Razavi.

Artificial intelligence and machine learning in cardiac electrophysiology. *Texas Heart Institute Journal*, 49, 03 2022. doi: 10.14503/thij-21-7576.

Alistair E. W. Johnson, Tom J. Pollard, Lu Shen, Li wei H. Lehman, Mengling Feng, Mohammad Mahdi Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G. Mark. Mimic-iii, a freely accessible critical care database. *Scientific Data*, 3, 2016. URL https://api.semanticscholar.org/CorpusID:33285731.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization, 2017.

Xiangzhen Kong, Vasanth Ravikumar, Siva K. Mulpuru, Henri Roukoz, and Elena G. Tolkacheva. A data-driven preprocessing framework for atrial fibrillation intracardiac electrocardiogram analysis. *Entropy*, 25(2), 2023. ISSN 1099-4300. doi: 10.3390/e25020332. URL https://www.mdpi.com/1099-4300/25/2/332.

Yikuan Li, Ramsey M. Wehbe, Faraz S. Ahmad, Hanyin Wang, and Yuan Luo. Clinical-longformer and clinical-bigbird: Transformers for long clinical sequences, 2022.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

Nikita Mehandru, Brenda Y. Miao, Eduardo Rodriguez Almaraz, Madhumita Sushil, Atul J. Butte, and Ahmed Alaa. Large language models as agents in the clinic, 2023.

Judea Pearl. *Causality*. Cambridge University Press, Cambridge, UK, 2 edition, 2009. ISBN 978-0-521-89560-6. doi: 10.1017/CBO9780511803161.

Mattia Prosperi, Yi Guo, Matt Sperrin, James S. Koopman, Jae S. Min, Xing He, Shannan Rich, Mo Wang, Iain E. Buchan, and Jiang Bian. Causal inference and counterfactual prediction in machine learning for actionable healthcare. *Nature Machine Intelligence*, 2:369–375, 07 2020. doi: 10.1038/s42256-020-0197-y.

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?, 2023a.

Jielin Qiu, Jiacheng Zhu, Shiqi Liu, William Han, Jingqi Zhang, Chaojing Duan, Michael Rosenberg, Emerson Liu, Douglas Weber, and Ding Zhao. Automated cardiovascular record retrieval by multimodal learning between electrocardiogram and clinical report, 2023b.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. "why should i trust you?": Explaining the predictions of any classifier, 2016.

Pedro Sanchez, Jeremy P. Voisey, Tian Xia, Hannah I. Watson, Alison Q. O'Neil, and Sotirios A. Tsaftaris. Causal machine learning for healthcare and precision medicine. *Royal Society Open Science*, 9, 08 2022. doi: 10.1098/rsos.220638.

Jingpu Shi and Beau Norgeot. Learning causal effects from observational data in healthcare: A review and summary. *Frontiers in Medicine*, 9, 07 2022. doi: 10.3389/fmed.2022.864882.

Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks, 2017.

Siyi Tang, Orod Razeghi, Ridhima Kapoor, Mahmood Alhusseini, Muhammad Fazal, Albert J Rogers, Miguel Rodrigo Bort, Paul Clopton, Paul J Wang, Daniel L Rubin, Sanjiv M Narayan, and Tina Baykaner. Machine learning–enabled multimodal fusion of intra-atrial and body surface signals in prediction of atrial fibrillation ablation outcomes. *Circulation-arrhythmia and Electrophysiology*, 15, 08 2022. doi: 10.1161/circep.122.010850.

Akhil Vaid, Joy Jiang, Ashwin Sawant, Stamatios Lerakis, Edgar Argulian, Yuri Ahuja, Joshua Lampert, Alexander Charney, Hayit Greenspan, Benjamin Glicksberg, Jagat Narula, and Girish Nadkarni. Heartbeit: Vision transformer for electrocardiogram data improves diagnostic performance at low sample sizes, 2022.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need, 2023.

Zhiguang Wang and Tim Oates. Encoding time series as images for visual inspection and classification using tiled convolutional neural networks. 2014. URL https://api.semanticscholar.org/CorpusID:16409971.

Kaiyue Wen, Yuchen Li, Bingbin Liu, and Andrej Risteski. Transformers are uninterpretable with myopic methods: a case study with bounded dyck grammars, 2023.

Sarah Wiegreffe and Yuval Pinter. Attention is not not explanation, 2019.

Manzil Zaheer, Guru Guruganesh, Avinava Dubey, Joshua Ainslie, Chris Alberti, Santiago Ontanon, Philip Pham, Anirudh Ravula, Qifan Wang, Li Yang, and Amr Ahmed. Big bird: Transformers for longer sequences, 2021.

Yongan Zhang, Anton Banta, Yonggan Fu, Mathews M. John, Allison Post, Mehdi Razavi, Joseph Cavallaro, Behnaam Aazhang, and Yingyan Lin. Rt-rcg: Neural network and accelerator search towards effective and real-time ecg reconstruction from intracardiac electrograms, 2021.

Appendix A. Additional Visualizations

Visualization of Tokenized Sequence We compare a three second tokenized signal when $V = \{50, 100, 250\}$ to its original representation (far left) in Figure 6.

Visualization of Reconstructed EGM Signal We visualize the reconstructed EGM signal and compare between using only the MLM learning objective versus our full learning objective in Figure 5. The reconstructed EGM signal using only the MLM learning objective follows the wave morphology more closely but results in lower AFib classification results. The reconstructed EGM signal using our full learning objective still closely resembles the signal's overall morphology, thus we accept this trade-off for the improvement in AFib classification.

Visualization of Attribution Scores We randomly sampled 4 normal and 4 AFib EGMs and visualize the attribution score distributions in Figure 7. The visualizations are based off of the fine-tuned checkpoint of our full learning objective with the BigBird model (Zaheer et al., 2021). We also visualize the attribution score when the T_A token is only masked out, due to the clinical significance. From Figure 7, we can observe that for normal EGMs, the attribution scores are precisely localized around the fluctuating sections of the signal, which is aligned with the clinician's perspective. For AFib EGMs, we

can still observe a pattern formulating where the rise and fall of the attribution scores are aligned with the irregularly oscillating portions as well.

Appendix B. Details on Masked Language Models

BigBird The BigBird model employs a sparse attention mechanism to efficiently process long sequences. This approach is characterized by three types of attention patterns: random, windowed, and global. The random pattern enables long-range interactions modeled after the Erdős-Rényi graph. The windowed attention captures local context within a fixed-size window, while the global tokens are included to maintain overall sequence context. These elements are integrated into a modified attention calculation, reducing the complexity from quadratic to linear with respect to sequence length, thereby facilitating scalability for datasets with long textual information. We use the pretrained weights 'google/bigbird-roberta-base' provided by the authors (Zaheer et al., 2021).

LongFormer The Longformer model is also optimized for processing long sequences of data. Beltagy et al. (2020) introduce a sparse attention mechanism that combines sliding window and global attention, reducing complexity from quadratic to linear and enabling handling of up to 4096 tokens. The sliding window attention applies standard self-attention within a fixed-size window, ensuring local context is maintained. Global attention allows designated tokens to interact with the entire sequence. We use the pretrained weights 'allenai/longformer-base-4096' provided by the authors (Beltagy et al., 2020).

Clinical BigBird and LongFormer Clinical Big-Bird and Clinical LongFormer utilize the same architectures as BigBird and LongFormer, respectively. The difference among these two models is that they are pretrained on 2 million clinical notes from the MIMIC-III dataset (Johnson et al., 2016). We use the pretrained weights 'yikuan8/Clinical-BigBird' and 'yikuan8/Clinical-Longformer' for Clinical BigBird and Clinical LongFormer respectively, provided by the authors (Li et al., 2022).

Appendix C. Additional Experiments

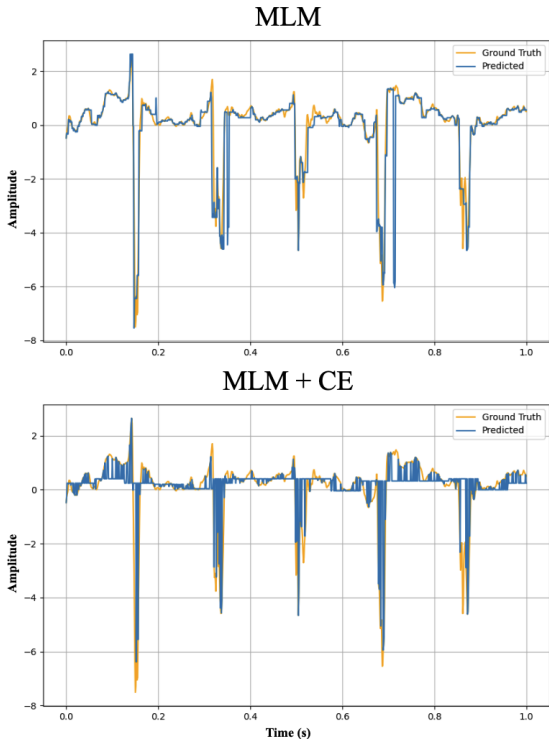**Only MLM vs MLM + Classification head** We compare our method with a traditional formu-

Figure 5: Comparison of L_{MLM} and $L_{MLM} + L_{CE}$ reconstructed AFib EGM signals with the BigBird model (Zaheer et al., 2021) using our final learning objective.

Table 6: Ablation Study on Signal Length M

M	Interpolation		AFib Classification
	MSE ↓	MAE ↓	Accuracy %
4000	0.88	0.40	98.3
3000	0.92	0.42	**99.3**
2000	1.18	1.04	97.4
1000	**0.80**	**0.37**	99.2

lation of doing classification with a MLM. The traditional formulation is that given some input, in this case $[CLS] \oplus T_S \oplus [SEP]$, to the MLM, the output of it will be inputted into an additional linear layer for classification. In our traditional formulation, we first finetune the MLM for the interpolation task then further finetune for AFib classification with an added linear layer. In Table 7, we can see that our method is able to outperform the traditional formulation. Therefore, we highlight the efficiency and supe-

rior performance of our method since we only require the MLM to be finetuned once.

Table 7: Ablation study on MLM vs MLM + classification head

Method	AFib Classification Accuracy %
BigBird (Zaheer et al., 2021) + Linear Layer	98.7
Ours BigBird (Zaheer et al., 2021)	**99.2**

Performance on Varying Signal Length M We compare the performance of our model when we vary the signal length M in Table 6, using the BigBird model (Zaheer et al., 2021) with our final learning objective. We observe that for interpolation, $M = 1000$ gives the best results, whereas for AFib classification, $M = 3000$ does slightly better. In general, our model seems to be robust to varying signal length inputs.

Performance on Different Losses L We compare the performance of our model with three different loss L formulations in Table 8 : Our full learning objective $L_{MLM} + L_{AFib}$, only the MLM loss L_{MLM}, and only the classification loss L_{AFib}. Our reported results are utilizing the BigBird model (Zaheer et al., 2021). We can see a clear performance tradeoff between the different learning objectives. Notably, when we add the L_{AFib} with L_{MLM} we can see that the accuracy rises while the MSE and MAE scores increases. Additionally, we visualize in Figure 5 and observe that the reconstructed EGM signal using our full learning objective is comparable to only using L_{MLM} while increasing the accuracy, thus we deem this tradeoff acceptable.

Table 8: Ablation study on different losses L.

L	Interpolation		AFib Classification
	MSE ↓	MAE ↓	Accuracy %
L_{AFib}	36.79	4.17	63.0
L_{MLM}	**0.44**	**0.16**	96.7
$L_{MLM} + L_{AFib}$	0.80	0.37	**99.2**

Performance on Different Bin Levels V. We compare the performance of our model with different bin levels V in Table 9. We can observe that although the MSE and MAE scores are very good for lower V, the accuracy takes a toll. Through this empirical analysis, we found that $V = 250$ to be the most
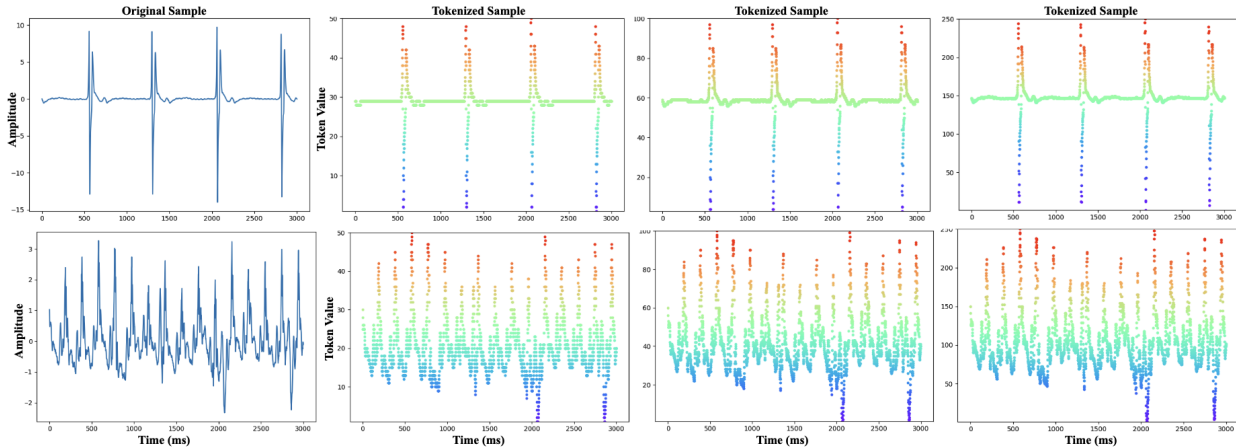
Figure 6: The tokenized representation of an EGM signal for a patient with a normal heartbeat (Top) and AFib (Bottom) when $V = \{50, 100, 250\}$ levels from left to right and starting with the the original time series EGM representation.

appropriate number to maintain good performance across interpolation and classification. Our reported results are utilizing the BigBird model (Zaheer et al., 2021).

Table 9: Ablation study on different bin levels V.

V	Interpolation MSE ↓	MAE ↓	AFib Classification Accuracy %
50	**0.40**	0.24	64.4
100	0.42	0.19	64.2
150	0.42	**0.17**	68.43
200	0.45	**0.17**	75.41
250	0.80	0.37	**99.2**

Performance on Fitting Previous Methods to Our Data For a fairer comparison to other methods, we implemented previous methods proposed in Alhusseini et al. (2020) and Tang et al. (2022) to our data and present the results in Table 10. Although, we do want to note that for Alhusseini et al. (2020), they utilize a basket catheter with 64 electrodes to collect their data. Additionally, they create a spatial heatmap that represents a 8 x 8 grid of the 64 electrodes as input to the CNN. In our setting, we only use one out of the 48 electrodes of our Octoray catheter as input to the CNN for a fair comparison. We can see from the table that our method is able to perform better compared with previous methods.

Table 10: Results on classification task utilizing models proposed in Alhusseini et al. (2020) and Tang et al. (2022).

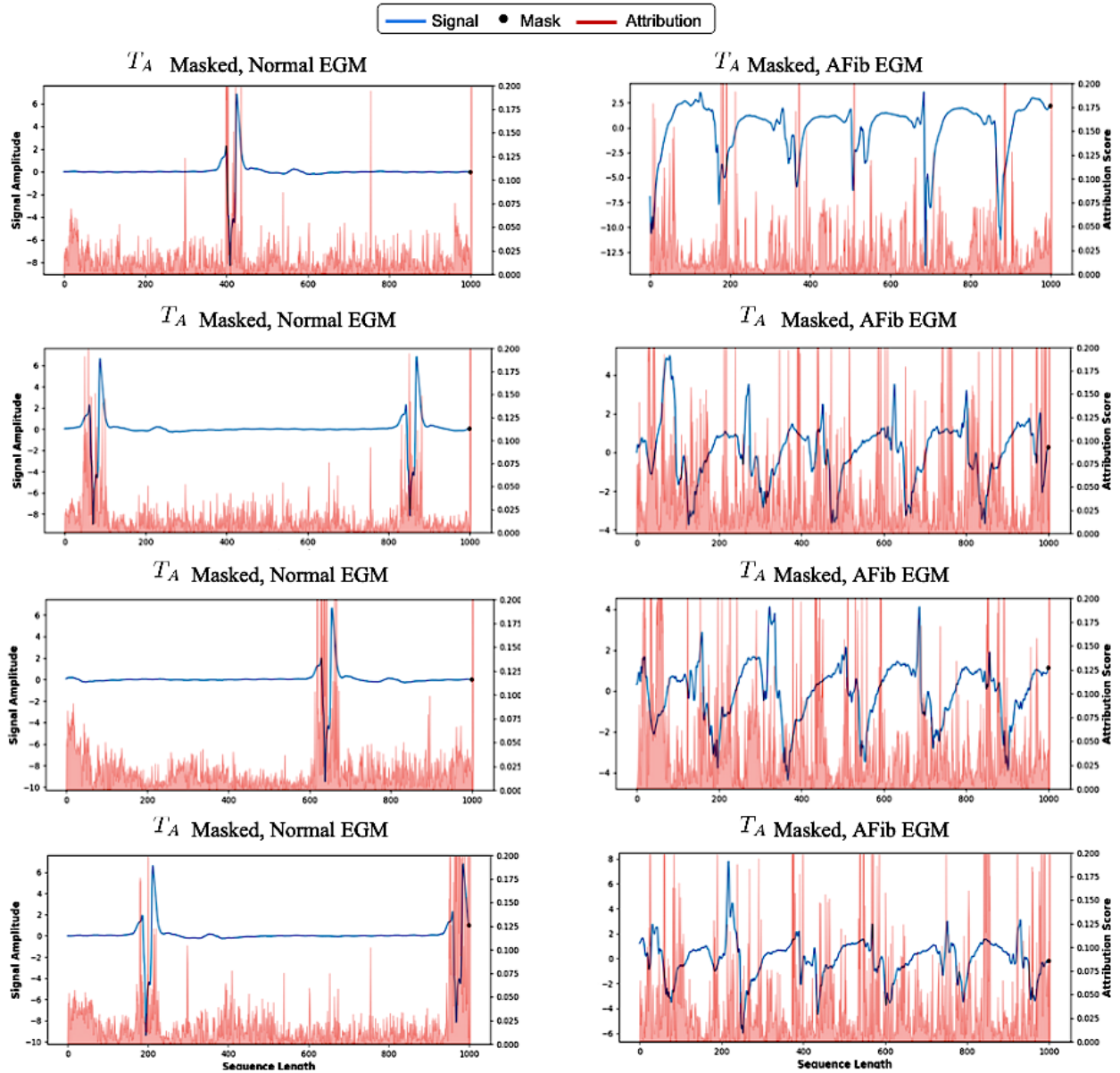| Method	AFib Classification Accuracy %
CatBoost (Tang et al., 2022)	90.6
CNN (Alhusseini et al., 2020)	91.6
K-Means (Alhusseini et al., 2020)	66.5
KNN (Alhusseini et al., 2020)	82.5
LDA (Alhusseini et al., 2020)	65.5
SVM (Alhusseini et al., 2020)	64.6
Ours BigBird (Zaheer et al., 2021)	**99.2**

Figure 7: We visualize the attribution scores of the finetuned checkpoint of our full learning objective using the BigBird model (Zaheer et al., 2021). We visualize randomly sampled normal and AFib EGMs with only masking the T_A token for clinical significance.