Automated Cardiovascular Record Retrieval by Multimodal Learning between Electrocardiogram and Clinical Report

Jielin Qiu[*1] JIELINQ@CS.CMU.EDU
Jiacheng Zhu[*2] ZJC@MIT.EDU
Shiqi Liu[1] SHIQILIU@ANDREW.CMU.EDU
William Han[1] WJHAN@ANDREW.CMU.EDU
Jingqi Zhang[1] JINGQIZ@ANDREW.CMU.EDU
Chaojing Duan[3] CHAOJIND@ANDREW.CMU.EDU
Michael A. Rosenberg[4] MICHAEL.A.ROSENBERG@CUANSCHUTZ.EDU
Emerson Liu[3] EMERSONLIU@MSN.COM
Douglas Weber[1] DOUGWEBER@CMU.EDU
Ding Zhao[1] DINGZHAO@CMU.EDU
[1] *Carnegie Mellon University*
[2] *MIT CSAIL*
[3] *Allegheny Health Network*
[4] *University of Colorado Denver*

[*] marked as equal contribution

Abstract

Automated interpretation of electrocardiograms (ECG) has garnered significant attention with the advancements in machine learning methodologies. Despite the growing interest, most current studies focus solely on classification or regression tasks which overlook a crucial aspect of clinical cardio-disease diagnosis: the diagnostic report generated by experienced human clinicians. In this paper, we introduce a novel approach to ECG interpretation, leveraging recent breakthroughs in Large Language Models (LLMs) and Vision-Transformer (ViT) models. Rather than treating ECG diagnosis as a classification or regression task, we propose an alternative method of automatically identifying the most similar clinical cases based on the input ECG data. Also, since interpreting ECG as images is more affordable and accessible, we process ECG as encoded images and adopt a vision-language learning paradigm to jointly learn vision-language alignment between encoded ECG images and ECG diagnosis reports. Encoding ECG into images can result in an efficient ECG retrieval system, which will be highly practical and useful in clinical applications. More importantly, our findings could serve as a crucial resource for providing diagnostic services in underdevelopment regions.

Keywords: Cardiovascular, Retrieval, Multimodal, Electrocardiogram

1. Introduction

Cardiovascular diseases, such as heart attacks and strokes, account for the majority of global deaths. ECG is a vital tool in cardiology and electrophysiology, providing valuable information about the heart's structure, electrical activity, and potential systemic conditions through waveform changes in timing and morphology. Accurate interpretation of clinical ECGs is critical, as it remains a primary method for identifying cardiac abnormalities and screening populations at risk of heart-related issues.

The precise interpretation of ECGs is essential for providing timely, efficient, and cost-effective interventions for acute cardiac conditions. Machine learning (ML) algorithms have been used to assist ECG interpretation, including disease classification (Nonaka and Seita, 2021; Khurshid et al., 2021; Raghunath et al., 2021; Giudicessi et al., 2021; Strodthoff et al., 2021), adversarial attack (Han et al., 2020a; Hossain et al., 2021; Chen et al., 2020a), data augmentation (Raghu et al., 2022; Nonaka and Seita, 2020), contrastive learning (Gopal et al., 2021), and the application of transformer models (Che et al., 2021; Natarajan et al., 2020; Behinaein et al., 2021).

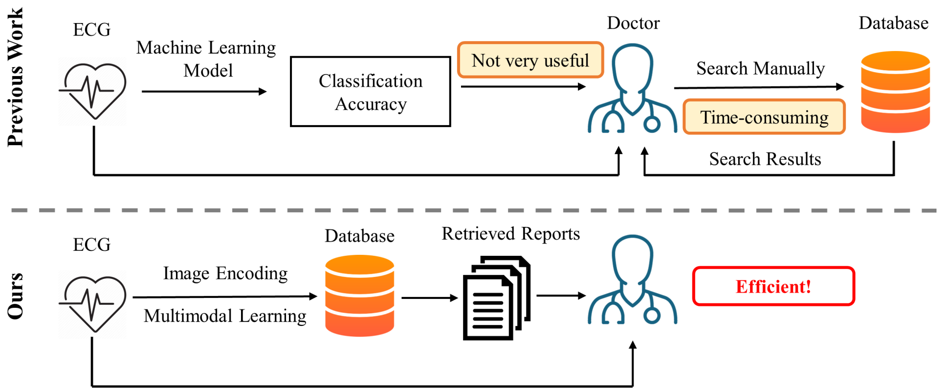

Figure 1: Prior studies only provided disease prediction accuracy for machine learning models, leaving doctors to cross-reference databases for precise diagnoses. In contrast, our method retrieves relevant past studies in the database, greatly reducing doctors' manual search efforts and improving patient care efficiency.

Most current machine learning ECG interpretation frameworks have practical limitations. They mainly use the ECG signal as input and diagnosis as a label, adapting to supervised learning as in other fields. However, ECG diagnosis is multifaceted, involving a complex hierarchy of disorders. For instance, the ST/T changes class can be divided into subclasses like ischemic in anterior leads (ISCA), ischemic in inferior leads (ISCI), non-specific ischemic (ISC), and non-specific ST changes (NST) (Guglin and Thatai, 2006). In practice, physicians commonly provide detailed ECG reports (Wagner et al., 2020) containing nuanced signal features and categorical diagnoses. AI-powered ECG frameworks, however, assume digital ECG processing via advanced systems. In reality, paper-printed ECGs (Zhang et al., 2023) from ECG monitor machines (Olson et al., 2013) are predominantly used by patients and doctors. Notably, paper-printed ECGs are the sole protocols in **underdeveloped regions**.

To overcome the aforementioned limitations, our goal is to enhance ECG interpretation automation by addressing the following challenge: *Can we automatically match it with the most similar ECG records in the database?*? This involves leveraging joint inference between the ECG signal and expert-written reports. This functionality can greatly aid in diagnosing common diseases like arrhythmia (Hong et al., 2020; Fu et al., 2021), reducing physicians' workload (Hannun et al., 2019). Additionally, this ECG data retrieval system can assist in diagnosing complex conditions such as atrial fibrillation and contractile dysfunction (Attia et al., 2019b,a), which pose challenges for supervised learning networks due to limited training data.

To realize this goal, we introduce an ECG-Text retrieval system that employs a multimodal information retrieval framework to automatically fetch expert-written reports along with corresponding ECG records. From a pragmatic standpoint, we treat ECG data as image input and employ various featurization methods. Our model is designed to discern the similarity score between these two modalities, enabling automatic identification of correspondences between ECG images and human language descriptions. Our contributions are outlined as follows:

- Our approach aims to improve ML-driven ECG automatic diagnosis by tackling the multimodal retrieval challenge and training to align the two modalities.
- We present a robust framework that provides clinicians with a practical and efficient method to automatically search and identify similar ECG records for newly acquired ECG data.
- Building upon progress in image-text alignment research, we highlight the treatment of ECG signals as images and introduce diverse preprocessing methods. This strategy makes our approach practical and readily applicable, given the widespread use of commercial ECG machines.

2. Related Work

Multimodal Learning in Healthcare The computational field of machine learning has faced the multimodal nature of clinical expert decision-making. Kline et al. (2022) summarized the current studies in multimodal learning in healthcare applications and identified topics ripe for future research. Amal et al. (2022) reviewed multimodal data fusion and machine learning in cardiovascular medicine. For example, the detection of cardiac amyloidosis can benefit from fusing ECG signals and echocardiograms with convolution neural networks (Goto et al., 2021). The multimodal approach also helps as combining salient physiological signals and EHR data can effectively predict the onset of hemodynamic decompensation (Hernandez et al., 2021). However, our study is the first to investigate the multimodal properties between ECG and natural language data.

Encode Time Series Signals into Images Deep learning has been successfully applied to automate ECG diagnosis (Han et al., 2020b). These methods are usually based on raw ECG signal data and corresponding features (Kiranyaz et al., 2015; Zhu et al., 2022). However, traditionally, ECG data is transformed into printed images with waveforms and interpreted by trained clinicians (Sangha et al., 2022a). To harness recent advancements in deep learning and computer vision, making ECG interpretation more practical and accessible, machine learning approaches that treat ECG data as image features have been investigated. An early approach combined either ECG images or signals (Sangha et al., 2022a) as inputs for cardiac disease diagnosis by a convolutional neural network based on the EfficientNet architecture. The idea of interpreting printed ECG papers has also been shown to be effective for diagnosing left ventricular (LV) systolic dysfunction (Sangha et al., 2022b). Additionally, digitizing printed ECG papers by scanning and processing raw printed images (Wu et al., 2022) is a critical task. Similarly, an automated ECG diagnostic pipeline employing paper-ECG images can facilitate accessible diagnostic services in regions with limited healthcare information systems.

3. Methods

Our approach comprises two key components: (1) the conversion of ECG time series signals into images, and (2) the utilization of these encoded ECG images and their corresponding clinical reports to construct an ECG record retrieval system. We delve into the specifics in Section 3.1 and Section 3.2 respectively.

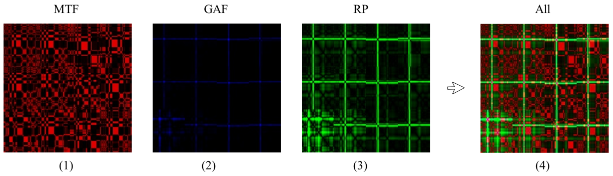

Figure 2: Examples of encoded ECG images by (1) MTF; (2) GAF; (3) RP; and (4) combine all three methods in three channels.

3.1. Encode ECG Signals into Images

In our study, we employed three different encoding methods to convert ECG time series signals into visual formats: Markov Transition Field (MTF), Gramian Angular Field (GAF), and Recurrence Plot (RP). Each technique's detailed explanation is provided in the subsequent sections, with further particulars available in Appendix B due to space limit.

3.1.1. Markov Transition Field (MTF)

Markov Transition Field (MTF) is a method of transforming time series data, such as ECG signals, into visual representations. MTF works by calculating transition probabilities between adjacent data points in a time series, and then using these probabilities to generate a matrix of color-coded pixels. Given a ECG time series X, the Q quantile bins are identified, and each data point x_i is assigned to its corresponding bin $q_j (j \in [1, Q])$. The resulting weighted adjacency matrix W, constructed using a first-order Markov chain model along the time axis, reflects the transition probabilities among the quantile bins. The frequency with which a data point in quantile bin q_j is followed by a point in bin q_i determines the value of the corresponding entry $w_{i,j}$ in W. Although W represents the Markov transition matrix after normalization by $\sum_j w_{ij} = 1$, it is insensitive to the distribution of X and the temporal dependencies between time steps t_i, resulting in a loss of information. To address this issue, the Markov Transition Field (MTF) M is defined as follows:

$$\begin{bmatrix} w_{ij|x_1 \in q_i, x_1 \in q_j} & w_{ij|x_1 \in q_i, x_2 \in q_j} & \cdots & w_{ij|x_1 \in q_i, x_n \in q_j} \\ w_{ij|x_2 \in q_i, x_1 \in q_j} & w_{ij|x_2 \in q_i, x_2 \in q_j} & \cdots & w_{ij|x_2 \in q_i, x_n \in q_j} \\ \vdots & \vdots & \ddots & \vdots \\ w_{ij|x_n \in q_i, x_1 \in q_j} & w_{ij|x_n \in q_i, x_2 \in q_j} & \cdots & w_{ij|x_n \in q_i, x_n \in q_j} \end{bmatrix} \tag{1}$$

It involves building a $Q \times Q$ Markov transition matrix W by dividing the time series data into Q quantile bins, where q_i and $q_j (q \in [1, Q])$ represent the quantile bins that contain the data at time stamps

i and j along the temporal axis. The MTF matrix M encodes the transition probabilities of the time series by spreading out the transition probability values from matrix W along the magnitude axis to M while taking into consideration the temporal positions. At each pixel M_{ij}, the probability of transitioning from the quantile at time step i to the quantile at time step j is assigned. In this way, the MTF matrix M captures the multi-span transition probabilities of the time series. The entry $M_{i,j||i-j|=k}$ in M represents the transition probability between points with a time interval of k, where $M_{i,j||j-i} = 1$ represents the transition process along the time axis with a skip step. The main diagonal M_{ii} in M is a special case when $k = 0$ and captures the probability of transitioning from each quantile to itself, i.e., the self-transition probability, at time step i.

3.1.2. Gramian Angular Field (GAF)

Gramian Angular Field (GAF) (Wang and Oates, 2014) is another method for transforming ECG time series signals into visual representations. GAF generates a matrix of cosine and sine values based on the pairwise differences between the original data points in the time series. This matrix is then transformed into an image, where each pixel corresponds to a particular combination of cosine and sine values. Similar to MTF, the resulting image captures important features of the original ECG signal, such as patterns and trends, which can aid in the interpretation and analysis of the data.

The Gramian Angular Field (GAF) (Wang and Oates, 2014) method represents time series data in a polar coordinate system instead of using the traditional Cartesian coordinates. In the Gramian matrix of GAF, each element corresponds to the cosine of the summation of angles. The rescaled time series $\tilde{X}$ of n real-valued observations are transformed to fall within the range of $[-1, 1]$ or $[0, 1]$ using the formula:

$$\tilde{x}^i_{-1} = \frac{(x_i - max(X) + (x_i - min(X))}{max(X) - min(X)} \tag{2}$$

$$or \quad \tilde{x}^i_0 = \frac{x_i - min(X)}{max(X) - min(X)} \tag{3}$$

Then, by encoding the value as the angular cosine and the time stamp as the radius, we represent the rescaled time series $\tilde{X}$ in polar coordinates as follows:

$$\phi = arccos(\tilde{x}_i), \quad -1 \leq \tilde{x}_i \leq 1, \quad \tilde{x}_i \in \tilde{X}, \quad r = \frac{t_i}{N}, \quad t_i \in N \tag{4}$$

Here, t_i is the time stamp, and N is a constant factor that regulates the span of the polar coordinate system. This encoding technique is a novel way to visualize time series data, where the values transform among different angular positions on the spanning circles as time passes, resembling water rippling. The encoding map is bijective, and it preserves absolute temporal relations, unlike Cartesian coordinates. The angular cosine function is monotonic for $\phi \in [0, \pi]$, producing a unique result in the polar coordinate system with a one-to-one inverse map.

We utilize the angular perspective of the polar coordinate system to examine temporal correlations between different time intervals by calculating the trigonometric sum/difference between each point. Specifically, we define the Gramian Summation Angular Field (GASF) and Gramian Difference Angular Field (GADF) as follows:

$$GASF = [cos(\phi_i + \phi_j)] = \tilde{X}' \cdot \tilde{X} - \sqrt{I - \tilde{X}^2}' \cdot \sqrt{I - \tilde{X}^2} \tag{5}$$

$$GADF = [sin(\phi_i - \phi_j)] = \sqrt{I - \tilde{X}^2}' \cdot \tilde{X} - \tilde{X}' \cdot \sqrt{I - \tilde{X}^2} \tag{6}$$

Here, I is the unit row vector $[1, 1, ..., 1]$. After transforming the time series into the polar coordinate system, we treat each time step as a 1-D metric space. Defining the inner product as follows:

$$< x, y >_1 = x \cdot y - \sqrt{1 - x^2} \cdot \sqrt{1 - y^2} \tag{7}$$

$$< x, y >_2 = \sqrt{1 - x^2} \cdot y - x \cdot \sqrt{1 - y^2} \tag{8}$$

The two types of Gramian Angular Fields (GAFs) are actually quasi-Gramian matrices $[< \tilde{x_1}, \tilde{x_1} >]$.

The Gramian Angular Fields (GAFs) offer multiple benefits. First, they enable the retention of temporal relationships, as the position's movement from the top-left to the bottom-right corresponds to the increase in time. The GAFs incorporate temporal correlations since $G_{i,j||i-j|=k}$ symbolizes the relative correlation due to the superimposition/difference of directions concerning time interval k. The main diagonal $G_{i,i}$ is a special case for $k = 0$, containing the original angular/value information.

3.1.3. Recurrence Plot (RP)

Recurrence Plot (RP) (Eckmann et al., 1987) is a non-linear time series analysis technique that can also be applied to transform ECG time series signals into visual representations. RP generates a square matrix that reflects the similarity between all pairs of data points in the time series. The matrix is constructed by measuring the distance between each pair of data points and comparing them to a predefined threshold value. RP has been shown to be effective in capturing complex patterns in ECG signals, such as P-waves and QRS complexes, which are important for the accurate diagnosis of cardiovascular diseases.

Given a time series $(x_1, \ldots, x_n)$, we can extract trajectories from it as follows:

$$\boldsymbol{x}i = (x_i, xi + \tau, \ldots, x_{i+(m-1)\tau}), \quad \forall i \in 1, \ldots, n-(m-1)\tau \tag{9}$$

Here, m denotes the dimension of the trajectories, and τ is the time delay. Once we have extracted the trajectories, we can create a recurrence plot, denoted by R, which is essentially the pairwise distance between the trajectories. Formally, we define $R_{i,j}$ as:

$$R_{i,j} = \Theta(\varepsilon - |\boldsymbol{x}_i - \boldsymbol{x}_j|), \quad \forall i, j \in 1, \ldots, n-(m-1)\tau \tag{10}$$

Here, Θ is the Heaviside step function, and ε is the threshold. The recurrence plot helps us visualize the structure and patterns of the time series by preserving the temporal dependencies and revealing the relative correlations between the extracted trajectories.

3.2. Retrieval System

This section commences with an overview of the model architecture, followed by a detailed account of the training objectives. An elaborate illustration of the model architecture is presented in Figure 3. Our model follows Li et al. (2021, 2022), which includes a vision encoder responsible for processing visual information, a language encoder dedicated to understanding textual data, and a multimodal encoder that integrates information from both the vision and language encoders to form a robust representation.

Vision Encoder Our current vision encoder architecture is based on a visual transformer (Dosovitskiy et al., 2021), which implements a patch-based processing approach that encodes an input image into a sequence of embeddings. This is achieved by dividing the image into patches and then performing a sequence of encoding operations on each patch. In addition, an extra [CLS] token is included to represent the global image feature. This approach has been shown to be more computation-friendly than using pre-trained object detectors for visual feature extraction (Chen et al., 2020b) and has been adopted by more recent methods such as ALBEF and ViLT (Li et al., 2021; Kim et al., 2021). Specifically, given an input image I, the ViT-based vision encoder generates a sequence of embeddings: $\boldsymbol{v}_{\text{cls}}, \boldsymbol{v}_1, ..., \boldsymbol{v}N$. Here, $\boldsymbol{v}$cls represents the embedding of the [CLS] token, and the remaining $\boldsymbol{v}_i$ represents the patch embeddings. It is worth noting that this patch-based processing approach allows the vision encoder to capture fine-grained details of the input image, which can be important for downstream tasks that require a high level of visual understanding.

Language Encoder Our text encoder is based on the highly effective BERT architecture (Devlin et al., 2019b), which employs a [CLS] token appended to the beginning of the input text to provide a summary of the sentence. The encoder also utilizes a bi-directional self-attention mechanism to generate representations for each of the input tokens. This approach is highly effective for capturing the context and meaning of each token in the input text, enabling the model to better understand the overall meaning of the text. When processing an input text T, the text encoder generates a sequence of embeddings $\boldsymbol{w}_{\text{cls}}, \boldsymbol{w}_1, ..., \boldsymbol{w}_N$, where $\boldsymbol{w}_{\text{cls}}$ represents the embedding of the [CLS] token, and the remaining $\boldsymbol{w}_i$ represent the embeddings of the individual input tokens. This sequence of embeddings is then passed to the multimodal encoder to be combined with the visual embeddings generated by the vision encoder.

Multimodal Encoder The multimodal encoder is a complex module that plays a critical role in enabling the model to learn the relationships between the visual and textual inputs. To achieve this, it incorporates an additional cross-attention (CA) layer that sits between the self-attention (SA) layer and the feed-forward network (FFN) for each transformer block of the text encoder. By doing so, the model can attend to both the textual and visual inputs and build better representations of the image-text pair. To create a multimodal representation of the image-text pair, the text input is modified by appending a task-specific [Encode] token at the end of the sequence, which is then fed into the multimodal encoder. The output embedding of this token is used as the final representation of the image-text pair. The embedding layers, CA layers, and FFN share similar functionality between encoding and decoding tasks, which means that they can be shared to improve training efficiency and benefit from multi-task learning. Additionally, the cross-attention layer introduces another set of attention weights to the model, which requires additional computation and increases the number of parameters to be learned. However, this additional complexity is necessary to enable the model to learn

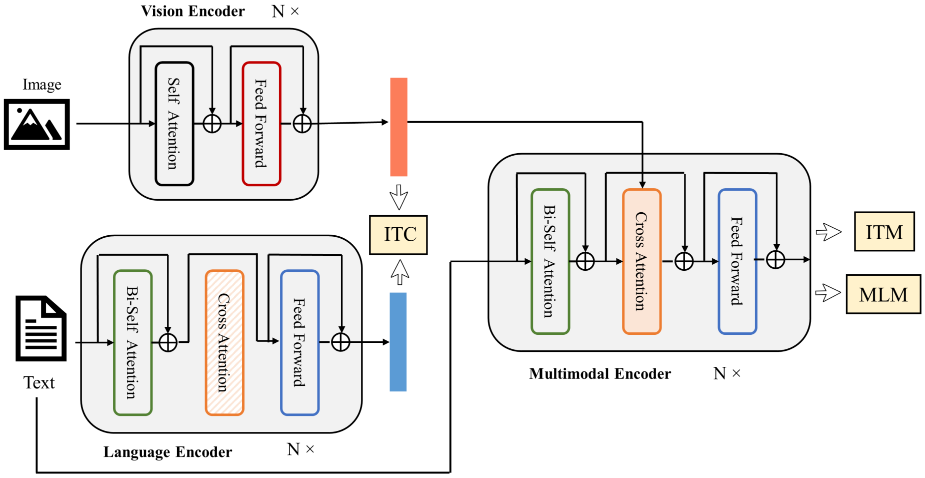

Figure 3: The overall architecture of our model, which comprises a vision encoder responsible for processing visual data, a language encoder that focuses on comprehending textual information, and a multimodal encoder that combines the input from both the vision and language encoders to fuse comprehensive representations.

the relationships between the visual and textual inputs and to achieve state-of-the-art performance on various image-text tasks.

3.3. Loss Objectives

There are three objectives during learning, including Image-Text Contrastive (ITC) Loss, Image-Text Matching (ITM) Loss, and Mask Language Modeling (MLM) Loss. An overview of each loss is provided in the subsequent sections. More details can be found in the Appendix due to the page limit.

Image-Text Contrastive Loss (ITC) To compute the ITC loss, we follow the approach proposed by Li et al. (2021), which introduces a momentum encoder to generate features and creates soft labels from the momentum encoder to serve as training targets. The soft labels help account for the potential positive samples in the negative pairs and improve the quality of the learned representations. The model learns a similarity function represented by $s = g_v(\boldsymbol{v}_{\text{cls}})^\top g_w(\boldsymbol{w}_{\text{cls}})$, which aims to increase the similarity scores for matching image-text pairs. Here, g_v and g_w refer to linear transformations that convert the `[CLS]` embeddings into lower-dimensional, normalized (256-d) representations. Following the MoCo approach (He et al., 2020), we use two queues to store the most recent M image-text representations obtained from the momentum unimodal encoders. The features obtained from the momentum encoders are normalized and denoted by $g'v(\boldsymbol{v}'_{\text{cls}})$ and $g'w(\boldsymbol{w}'_{\text{cls}})$.

To calculate the similarity score between an image-text pair and a text-image pair, we define $s(I,T) = g_v(\boldsymbol{v}_{\text{cls}})^\top g'w(\boldsymbol{w}'_{\text{cls}})$ and $s(T,I) = g_w(\boldsymbol{w}_{\text{cls}})^\top g'v(\boldsymbol{v}'_{\text{cls}})$, respectively. We use the softmax-normalized image-to-text and text-to-image similarity to calculate each image and text. This is represented by the equations below, where τ is a temperature parameter that can be learned:

$$p_m^{\text{i2t}}(I) = \frac{\exp(s(I,T_m)/\tau)}{\sum_{m=1}^{M}\exp(s(I,T_m)/\tau)}, \tag{11}$$

$$p_m^{\text{t2i}}(T) = \frac{\exp(s(T,I_m)/\tau)}{\sum_{m=1}^{M}\exp(s(T,I_m)/\tau)} \tag{12}$$

We represent the ground-truth one-hot similarity as $\boldsymbol{y}^{\text{i2t}}(I)$ and $\boldsymbol{y}^{\text{t2i}}(T)$, where negative pairs have a probability of 0, and the positive pair has a probability of 1. The image-text contrastive loss is defined as the cross-entropy H between $\boldsymbol{p}$ and $\boldsymbol{y}$, which is shown in the following equation:

$$\mathcal{L}\text{itc} = \frac{1}{2}\mathbb{E}(I,T) \sim D\big[\mathrm{H}(\boldsymbol{y}^{\text{i2t}}(I), \boldsymbol{p}^{\text{i2t}}(I)) + \mathrm{H}(\boldsymbol{y}^{\text{t2i}}(T), \boldsymbol{p}^{\text{t2i}}(T))\big] \tag{13}$$

Image-Text Matching Loss (ITM) ITM is a binary classification task where the model predicts whether an image-text pair is positive (matched) or negative (unmatched) based on its multimodal feature. The ITM head, which is a linear layer, is used to make this prediction. To obtain the joint representation of the image-text pair, we use the output embedding of the `[CLS]` token from the multimodal encoder, and then append a fully-connected

(FC) layer followed by softmax to predict a two-class probability p^{itm}. The ITM loss is defined as: $\mathcal{L}\text{itm} = \mathbb{E}(I, T) \sim DH(\boldsymbol{y}^{\text{itm}}, \boldsymbol{p}^{\text{itm}}(I, T))$, where $\boldsymbol{y}^{\text{itm}}$ is a 2-dimensional one-hot vector representing the ground-truth label. To improve the selection of negative pairs, we employ a strategy called hard negative mining, as proposed by Li et al. (2021). This strategy involves selecting negative pairs that have a higher contrastive similarity within a batch.

Mask Language Modeling Loss (MLM) The Mask Language Modeling Loss (MLM) is used to predict masked words using both the image and contextual text. In this loss, we randomly mask out input tokens with a probability of 15% and replace them with the special token `[MASK]`, with 10% random tokens, 10% unchanged, and 80% `[MASK]` replacements following the BERT approach. The predicted probability of a masked token is denoted by $\boldsymbol{p}^{\text{msk}}(I, \hat{T})$, where $\hat{T}$ represents the masked text. The cross-entropy loss is used to minimize the difference between the predicted and ground-truth distributions, which is expressed as follows:

$$\mathcal{L}\text{mlm} = \mathbb{E}(I, \hat{T}) \sim DH(\boldsymbol{y}^{\text{msk}}, \boldsymbol{p}^{\text{msk}}(I, \hat{T})) \tag{14}$$

$\boldsymbol{y}^{\text{msk}}$ represents a one-hot vocabulary distribution and the ground-truth token has a probability of 1.

4. Experiments

4.1. Dataset and Preprocessing

Our experiments were conducted using the PTB-XL dataset (Wagner et al., 2020), which comprises clinical 12-lead ECG signals that are 10 seconds in length. The dataset includes five different conditions: Normal ECG (NORM), Myocardial Infarction (MI), ST/T Change (STTC), Conduction Disturbance (CD), and Hypertrophy (HYP). The waveform files are stored in the WaveForm DataBase (WFDB) format and have a precision of 16 bits at a resolution of 1μV/LSB, with a sampling frequency of 100Hz. The raw waveform data was annotated by up to two cardiologists who assigned one or more ECG statements to each record, resulting in a total of 71 different ECG statements that conform to the SCP-ECG standard. These statements cover diagnostic, form, and rhythm-related information. Additionally, the dataset contains extensive metadata on demographics, infarction characteristics, likelihoods for diagnostic ECG statements, as well as annotated signal properties. To convert the time series data into a spectrum, we leveraged the WFDB library (Xie et al.) to read the raw data and performed Fast Fourier Transform (FFT). In order to eliminate noise, we implemented n-points window filtering, and to eliminate power frequency interference, which occurs at 50Hz, we employed notch processing with a quality factor of 30 (Qiu et al., 2023b).

4.2. Experimental Setting

The initialization of our visual encoder, text encoder, and multimodal encoder was carried out using the image encoder, text encoder, and image-grounded text encoder from Li et al. (2022), respectively. Specifically, the visual encoder was based on ViT (Dosovitskiy et al., 2020) pre-trained on ImageNet (He et al., 2015), while the other two encoders were initialized from BERT (Devlin et al., 2019a). All three encoders were trained on a dataset consisting of 14M images from COCO (Lin et al., 2014), Visual Genome (Krishna et al., 2016), Conceptual Captions (Sharma et al., 2018), Conceptual 12M (Changpinyo et al., 2021), and SBU captions (Ordonez et al., 2011), as described in Li et al. (2022).

Next, we fine-tuned the three encoders on our ECG image data using the AdamW optimizer (Loshchilov and Hutter, 2017) with a weight decay of 0.05. The learning rate was warmed-up to 3e-4 (for ViT-B) / 2e-4 (for ViT-L) and decayed linearly with a rate of 0.85. During fine-tuning, we randomly cropped images to a resolution of 384 × 384. Our experiments were conducted on 4 NVIDIA A6000. We evaluate our models using the recall at K (R@K) metric, where K $= 1, 5, 10$, and report the RSUM, which is the sum of the recall metrics at K $= 1, 5, 10$ for both image and text retrieval tasks.

4.3. Experimental Results and Discussions

Table 1 presents the results of our experiments comparing different image encoding methods. We conducted experiments in various settings to obtain a comprehensive understanding of the methods:
(1) The “Simple-plot” method serves as a straightforward baseline, where we plotted the ECG time series signals of 12 leads, selecting one ECG pause from each lead and putting them in a 4 × 3 layout.
(2) Using each encoding method individually, we formulated three baseline approaches referred to as “MTF-only”, “GAF-only”, and “RP-only”.
(3) Two encoding methods were randomly selected from the three and concatenated in different image

channels.
(4) Instead of encoding each lead independently, we concatenated the 12 leads of one ECG pause into a single vector, which we then visualized using all three encoding methods. This approach is referred to as "All-Concat".
(5) Finally, we gridded all three encoding methods in three image channels under both zero-shot and fine-tune settings, referred to as "All-Grid (zero-shot)" and "All-Grid (fine-tune)", respectively.

Table 1 presents the experimental findings, which indicate that for single encoding comparison, RP encoding significantly outperforms both MTF and GAF encoding techniques. Moreover, the combination of GAF and RP encoding demonstrates superior performance compared to the other two combinations. Remarkably, the "All-Grid (fine-tune)" method exhibits the best overall performance among all the baseline methods. A detailed analysis of the zero-shot and fine-tuning results shows that fine-tuning has a considerable impact on improving performance.

The "All-Grid (fine-tune)" method utilizes a fine-tuning approach to improve the performance of the models by iteratively adjusting the parameters of the network. This method achieves the best overall performance by effectively leveraging the available data. The analysis of the zero-shot and fine-tuning results indicates that fine-tuning significantly enhances the performance of the models, highlighting the importance of optimizing the network parameters to improve the accuracy of the predictions.

4.4. Ablation Study

In any modeling exercise, a vast array of parameters and settings can be adjusted to optimize performance. However, it is not always clear which of these factors has the most significant impact on the final output. In order to gain a better understanding of the inner workings of our model and the effect that each individual parameter has on its performance, we conducted a series of ablation studies.

To gain insight into the impact of batch size on the training and testing of our model, we conducted an ablation study. In this study, we systematically varied the training and testing batch sizes, and the results are presented in Table 2. Surprisingly, we found that a smaller training batch size led to better performance. This observation may seem counterintuitive, as larger batch sizes are typically favored in deep learning to accelerate training. However, our results suggest that a smaller training batch size may help the model converge more quickly and reduce overfitting. In addition, we observed that a smaller testing batch size also contributed to improved performance, when the training batch size was kept the same. This finding highlights the importance of matching the testing batch size to the training batch size, to ensure that the model is evaluated on a representative sample of data. By carefully selecting the appropriate training and testing batch sizes, we can optimize our model and achieve better results.

In addition, we conducted an ablation study on the selection of the visual encoder. There are two choices of visual encoder selection: ViT-base and ViT-large. ViT-base has 12 transformer layers, about 85 million parameters, and is trained on images resized to 224x224 pixels. It is a relatively smaller model and is suitable for smaller datasets or where memory or computational resources are limited. ViT-large has 24 transformer layers, about 307 million parameters, and is trained on images resized to 384x384 pixels. It is a more complex and larger model, which typically results in better performance on large-scale datasets.

The results of the ablation study are summarized in Table 3. Notably, we observed that for both the All-Grid and All-Concat settings, ViT-base outperformed ViT-large in terms of classification accuracy. These results are surprising, as ViT-large has more parameters and has been shown to perform better than ViT-base on pretraining tasks. However, we posit that the reason for this discrepancy lies in the size of the ECG image dataset. Specifically, since the dataset is relatively small, the fine-tuned embeddings of the ViT-base can more quickly adapt to the unique features of ECG images. In contrast, ViT-large contains more parameters and may require a larger dataset for effective fine-tuning. These findings have important implications for the use of ViT in medical image analysis. While ViT has shown great promise in a variety of visual recognition tasks, it is important to carefully consider its performance when applied to medical imaging datasets. Our results suggest that ViT-base may be a better choice for small medical imaging datasets, while ViT-large may be more effective for larger datasets. Additionally, our study highlights the importance of conducting careful evaluations of visual encoders in the medical imaging domain to ensure that they perform well on the specific imaging modality in question.

Table 1: Experimental results.

Method	Report Retrieval			Image Retrieval			
	R@1	R@5	R@10	R@1	R@5	R@10	RSUM
Simple-plot	5.51	17.16	25.64	4.98	17.08	26.05	96.42
MTF-only	1.42	5.67	10.72	1.97	6.07	10.80	36.65
GAF-only	2.69	10.53	17.04	3.22	11.01	17.60	62.09
RP-only	5.27	16.84	25.68	5.63	17.20	26.44	97.06
MTF+GAF	4.02	14.03	21.91	4.10	14.42	21.91	80.39
GAF+RP	6.30	21.04	30.36	6.93	20.69	30.65	115.97
PR+MTF	5.12	17.26	25.84	4.96	17.10	26.56	96.84
All-Concat	1.58	7.25	12.77	1.73	7.33	13.71	4.37
All-Grid (zero-shot)	0.21	1.06	1.91	0.43	1.06	1.91	6.58
All-Grid (fine-tune)	**7.88**	**24.51**	**34.04**	**8.27**	**23.96**	**34.91**	**133.57**

Table 2: Ablation study on learning batch.

Batch	Report Retrieval			Image Retrieval			
	R@1	R@5	R@10	R@1	R@5	R@10	RSUM
Training 8 + Testing 32	**7.88**	**24.51**	**34.04**	**8.27**	**23.96**	**34.91**	**133.57**
Training 16 + Testing 32	7.88	22.54	32.23	7.32	21.91	33.41	125.29
Training 32 + Testing 32	7.01	20.88	32.23	6.78	21.04	32.47	120.41

Table 3: Ablation study on vision encoder.

Vision Encoder	Report Retrieval			Image Retrieval			
	R@1	R@5	R@10	R@1	R@5	R@10	RSUM
All-concat (ViT-base)	2.52	8.91	13.79	2.52	7.25	12.77	47.76
All-concat (ViT-large)	1.58	7.25	12.77	1.73	7.33	13.71	44.37
All-Grid (ViT-base)	**7.88**	**24.51**	**34.04**	**8.27**	**23.96**	**34.91**	**133.57**
All-Grid (ViT-large)	3.47	12.32	21.43	4.73	14.26	22.14	78.35

5. Discussion and Conclusion

Based on the experiments above, we have observed the strong potential of transforming ECG time series signals into images. Furthermore, by incorporating state-of-the-art advancements in vision-language learning, additional advantages can be gained from these encoded images. Our proposed model suggests that jointly learning the encoded ECG images and doctor's reports can yield improved representations. These representations hold promise for various clinical applications, including retrieving relevant previous diagnosis reports from a database. This support and reference can greatly assist doctors, leading to enhanced patient treatment outcomes. Given the critical nature of healthcare, enhancing patient care remains of utmost importance. Introducing the proposed model into clinical applications has the potential to reshape the healthcare landscape and significantly influence patient outcomes. Therefore, we believe that our proposed model holds substantial practical value in the realm of clinical applications, offering significant advantages for patients, doctors, and the broader healthcare ecosystem.

Limitations While our study has illuminated the potential of MTF, GAF, and RP methods for ECG data analysis, it's important to acknowledge that other encoding techniques might also yield favorable outcomes. Furthermore, the dataset size used in our study might not comprehensively cover the variability and intricacy of ECG signals. Additionally, the accuracy of the doctor's report, a component of multimodal learning, could pose limitations due to interobserver variability, potentially impacting the quality of learned representations. Moreover, variables like patient demographics, medical history, and comorbidities were not considered, suggesting an avenue for future exploration to enhance the generalizability of our findings. Hence, more extensive research, utilizing larger and more diverse datasets, encompassing various analysis techniques, and accounting for confounding factors, is crucial to fully delve into the potential of our approach.

Acknowledgements

The research is partially supported by the DARPA ADAPTER program, and partially supported by the Allegheny Health Network and Mario Lemieux Center for Innovation and Research in EP.

References

Demilade A. Adedinsewo, Habeeba Siddiqui, Patrick W. Johnson, Erika J Douglass, Michal Cohen-Shelly, Zachi I. Attia, Paul A. Friedman, Peter A. Noseworthy, and Rickey E. Carter. Digitizing paper based ecg files to foster deep learning based analysis of existing clinical datasets: An exploratory analysis. *Intelligence-Based Medicine*, 2022.

Miquel Alfaras, Miguel C. Soriano, and Silvia Ortín. A fast machine learning model for ecg-based heartbeat classification and arrhythmia detection. *Frontiers in Physics*, 2019.

Saeed Amal, Lida Safarnejad, Jesutofunmi A. Omiye, Ilies Ghanzouri, John H Cabot, and Elsie Gyang Ross. Use of multi-modal data and machine learning to improve cardiovascular disease care. *Frontiers in Cardiovascular Medicine*, 9, 2022.

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

Zachi I Attia et al. An artificial intelligence-enabled ecg algorithm for the identification of patients with atrial fibrillation during sinus rhythm: a retrospective analysis of outcome prediction. *The Lancet*, 394(10201):861–867, 2019b.

Yehualashet Megersa Ayano, Friedhelm Schwenker, Bisrat Derebssa Dufera, and Taye Girma Debelee. Interpretable machine learning techniques in ecg-based heart disease classification: A systematic review. *Diagnostics*, 13, 2022.

Saira Aziz, Sajid Ahmed, and Mohamed-Slim Alouini. Ecg-based machine-learning algorithms for heartbeat classification. *Scientific Reports*, 11, 2021.

Silvio Barra, Salvatore M. Carta, Andrea Corriga, Alessandro Sebastian Podda, and Diego Reforgiato Recupero. Deep learning and time series-to-image encoding for financial forecasting. *IEEE/CAA Journal of Automatica Sinica*, 7:683–692, 2020.

Behnam Behinaein, Anubha Bhatti, Dirk Rodenburg, Paul C. Hungler, and Ali Etemad. A transformer architecture for stress detection from ecg. *2021 International Symposium on Wearable Computers*, 2021.

Jianwei Bi, Hui Li, and Zhiyuan Fan. Tourism demand forecasting with time series imaging: A deep learning model. *Annals of Tourism Research*, 90: 103255, 2021.

Andriana S. L. O. Campanharo, M. Irmak Sirer, R. Dean Malmgren, Fernando M. Ramos, and Luis A. Nunes Amaral. Duality between time series and networks. In *PloS one*, 2011.

Soravit Changpinyo, Piyush Kumar Sharma, Nan Ding, and Radu Soricut. Conceptual 12m: Pushing web-scale image-text pre-training to recognize long-tail visual concepts. *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3557–3567, 2021.

Chao Che, Peiliang Zhang, Min Zhu, Yue Qu, and Bo Jin. Constrained transformer network for ecg signal processing and arrhythmia classification. *BMC Medical Informatics and Decision Making*, 21, 2021.

Huangxun Chen, Chenyu Huang, Qianyi Huang, Qian Zhang, and Wei Wang. Ecgadv: Generating adversarial electrocardiogram to misguide arrhythmia classification system. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 3446–3453, 2020a.

Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. UNITER: universal image-text representation learning. In *ECCV*, volume 12375, pages 104–120, 2020b.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *ArXiv*, abs/1810.04805, 2019a.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: pre-training of deep bidirectional transformers for language understanding. In Jill Burstein, Christy Doran, and Thamar Solorio, editors, *NAACL*, pages 4171–4186, 2019b.

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. *ArXiv*, abs/2010.11929, 2020.

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. In *ICLR*, 2021.

Jean-Pierre Eckmann, Sylvie Oliffson Kamphorst, and David Ruelle. Recurrence plots of dynamical systems. *EPL*, 4:973–977, 1987.

Zhaoji Fu, Shenda Hong, Rui Zhang, and Shaofu Du. Artificial-intelligence-enhanced mobile system for cardiovascular health management. *Sensors*, 21(3): 773, 2021.

John R. Giudicessi et al. Artificial intelligence-enabled assessment of the heart rate corrected qt interval using a mobile electrocardiogram device. *Circulation*, 2021.

Bryan Gopal, Ryan Han, Gautham Raghupathi, Andrew Ng, Geoff Tison, and Pranav Rajpurkar. 3kg: contrastive learning of 12-lead electrocardiograms using physiologically-inspired augmentations. In *Machine Learning for Health*, pages 156–167. PMLR, 2021.

Shinichi Goto, Keitaro Mahara, Lauren Beussink-Nelson, Hidehiko Ikura, Yoshinori Katsumata, Jin Endo, Hanna K Gaggin, Sanjiv J Shah, Yuji Itabashi, Calum A MacRae, et al. Artificial intelligence-enabled fully automated detection of cardiac amyloidosis using electrocardiograms and echocardiograms. *Nature communications*, 12(1): 2726, 2021.

Maya E Guglin and Deepak Thatai. Common errors in computer electrocardiogram interpretation. *International journal of cardiology*, 106(2):232–237, 2006.

Xintian Han, Yuxuan Hu, Luca Foschini, Larry Chinitz, Lior Jankelson, and Rajesh Ranganath. Deep learning models for electrocardiograms are susceptible to adversarial attack. *Nature medicine*, 26(3):360–363, 2020a.

Xintian Han, Yuxuan Hu, Luca Foschini, Larry Chinitz, Lior Jankelson, and Rajesh Ranganath. Deep learning models for electrocardiograms are susceptible to adversarial attack. *Nature medicine*, 26(3):360–363, 2020b.

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

Nima Hatami, Yann Gavet, and Johan Debayle. Classification of time-series images using deep convolutional neural networks. In *International Conference on Machine Vision*, 2017.

Kaiming He, X. Zhang, Shaoqing Ren, and Jian Sun. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. *2015 IEEE International Conference on Computer Vision (ICCV)*, pages 1026–1034, 2015.

Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *CVPR*, 2020.

Larry Hernandez et al. Multimodal tensor-based method for integrative and continuous patient monitoring during postoperative cardiac care. *Artificial Intelligence in Medicine*, 113:102032, 2021.

Shenda Hong, Zhaoji Fu, Rongbo Zhou, Jie Yu, Yongkui Li, Kai Wang, and Guanlin Cheng. Cardiolearn: a cloud deep learning service for cardiac disease detection from electrocardiogram. In *Companion Proceedings of the Web Conference 2020*, pages 148–152, 2020.

Khondker Fariha Hossain, Sharif Amit Kamran, Alireza Tavakkoli, Lei Pan, Xingjun Ma, Sutharshan Rajasegarar, and Chandan Karmaker. Ecg-adv-gan: Detecting ecg adversarial examples with conditional generative adversarial networks. In *2021 20th IEEE International Conference on Machine Learning and Applications (ICMLA)*, pages 50–56. IEEE, 2021.

Isaak Kavasidis, Simone Palazzo, Concetto Spampinato, Daniela Giordano, and Mubarak Shah. Brain2image: Converting brain signals into images.

Proceedings of the 25th ACM international conference on Multimedia, 2017.

Shaan Khurshid et al. Electrocardiogram-based deep learning and clinical risk factors to predict atrial fibrillation. *Circulation*, 2021.

Sungsoo Kim, Sohee Kwon, Mia K. Markey, Alan Conrad Bovik, Sung Hwi Hong, Junyong Kim, Hye Jin Hwang, Boyoung Joung, Hui Nam Pak, Moon-Hyeong Lee, and Junbeom Park. Machine learning based potentiating impacts of 12-lead ecg for classifying paroxysmal versus non-paroxysmal atrial fibrillation. *International Journal of Arrhythmia*, 23:1–9, 2022.

Wonjae Kim, Bokyung Son, and Ildoo Kim. Vilt: Vision-and-language transformer without convolution or region supervision. *arXiv preprint arXiv:2102.03334*, 2021.

Serkan Kiranyaz, Turker Ince, Ridha Hamila, and M. Gabbouj. Convolutional neural networks for patient-specific ecg classification. *2015 37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pages 2608–2611, 2015.

Adrienne S. Kline, Hanyin Wang, Yikuan Li, Saya Dennis, Meghan R. Hutch, Zhenxing Xu, Fei Wang, Feixiong Cheng, and Yuan Luo. Multimodal machine learning in precision health: A scoping review. *NPJ Digital Medicine*, 5, 2022.

Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A. Shamma, Michael S. Bernstein, and Li Fei-Fei. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International Journal of Computer Vision*, 123:32–73, 2016.

Junnan Li, Ramprasaath R. Selvaraju, Akhilesh Deepak Gotmare, Shafiq Joty, Caiming Xiong, and Steven Hoi. Align before fuse: Vision and language representation learning with momentum distillation. In *NeurIPS*, 2021.

Junnan Li, Dongxu Li, Caiming Xiong, and Steven C. H. Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International Conference on Machine Learning*, 2022.

Tsung-Yi Lin, Michael Maire, Serge J. Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft coco: Common objects in context. In *European Conference on Computer Vision*, 2014.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2017.

Giovanna Martínez-Arellano, Germán Terrazas, and Svetan M. Ratchev. Tool wear classification using time series imaging and deep learning. *The International Journal of Advanced Manufacturing Technology*, 104:3647 – 3662, 2019.

Siddhartha Mishra, Gaurav Khatwani, Rupali Patil, Darshan Sapariya, Vruddhi Shah, Darshna Parmar, Sharath Dinesh, Prathamesh Daphal, and Ninad Dileep Mehendale. Ecg paper record digitization and diagnosis using deep learning. *Journal of Medical and Biological Engineering*, 41:422 – 432, 2020.

Annamalai Natarajan, Yale Chang, Sara Mariani, Asif Rahman, Gregory Boverman, Shruti Gopal Vij, and Jonathan Rubin. A wide and deep transformer neural network for 12-lead ecg classification. *2020 Computing in Cardiology*, pages 1–4, 2020.

Naoki Nonaka and Jun Seita. Electrocardiogram classification by modified efficientnet with data augmentation. In *2020 Computing in Cardiology*, pages 1–4. IEEE, 2020.

Naoki Nonaka and Jun Seita. In-depth benchmarking of deep neural network architectures for ecg diagnosis. In *Proceedings of the 6th Machine Learning for Healthcare Conference*, Proceedings of Machine Learning Research, pages 414–439. PMLR, 2021.

Janet E Olson, Euijung Ryu, Kiley J Johnson, Barbara A Koenig, Karen J Maschke, Jody A Morrisette, Mark Liebow, Paul Y Takahashi, Zachary S Fredericksen, Ruchi G Sharma, et al. The mayo clinic biobank: a building block for individualized medicine. In *Mayo Clinic Proceedings*, volume 88, pages 952–962. Elsevier, 2013.

Vicente Ordonez, Girish Kulkarni, and Tamara L. Berg. Im2text: Describing images using 1 million captioned photographs. In *NIPS*, 2011.

Abdolrahman Peimankar and Sadasivan K. Puthusserypady. Dens-ecg: A deep learning approach for ecg signal delineation. *ArXiv*, abs/2005.08689, 2020.

Zhen Qin, Yibo Zhang, Shuyu Meng, Zhiguang Qin, and Kim-Kwang Raymond Choo. Imaging and fusing time series for wearable sensor-based human activity recognition. *Inf. Fusion*, 53:80–87, 2020.

Jielin Qiu, Jiacheng Zhu, Michael Rosenberg, Emerson Liu, and D. Zhao. Optimal transport based data augmentation for heart disease diagnosis and prediction. *ArXiv*, abs/2202.00567, 2022.

Jielin Qiu, William Jongwon 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*, abs/2301.09017, 2023a. URL https://api.semanticscholar.org/CorpusID:256105761.

Jielin Qiu, Jiacheng Zhu, Mengdi Xu, Peide Huang, Michael Rosenberg, Douglas Weber, Emerson Liu, and Ding Zhao. Cardiac disease diagnosis on imbalanced electrocardiography data through optimal transport augmentation. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2023b.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. *arXiv preprint arXiv:2103.00020*, 2021.

Aniruddh Raghu, Divya Shanmugam, Eugene Pomerantsev, John Guttag, and Collin M Stultz. Data augmentation for electrocardiograms. In *Proceedings of the Conference on Health, Inference, and Learning*, Proceedings of Machine Learning Research, pages 282–310. PMLR, 2022.

Sushravya Raghunath et al. Deep neural networks can predict new-onset atrial fibrillation from the 12-lead ecg and help identify those at risk of atrial fibrillation–related stroke. *Circulation*, 143:1287 – 1298, 2021.

Veer Sangha, Bobak J Mortazavi, Adrian D Haimovich, Antônio H Ribeiro, Cynthia A Brandt, Daniel L Jacoby, Wade L Schulz, Harlan M Krumholz, Antonio Luiz P Ribeiro, and Rohan Khera. Automated multilabel diagnosis on electrocardiographic images and signals. *Nature communications*, 13(1):1583, 2022a.

Veer Sangha, Arash A Nargesi, Lovedeep S Dhingra, Akshay Khunte, Bobak J Mortazavi, Antônio H Ribeiro, Evgeniya Banina, Oluwaseun Adeola, Nadish Garg, Cynthia A Brandt, et al. Detection of left ventricular systolic dysfunction from electrocardiographic images. *medRxiv*, pages 2022–06, 2022b.

Aya Nabil Sayed, Yassine Himeur, and Fayçal Bensaali. From time-series to 2d images for building occupancy prediction using deep transfer learning. *Eng. Appl. Artif. Intell.*, 119:105786, 2023.

Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *Annual Meeting of the Association for Computational Linguistics*, 2018.

Sulaiman S Somani, Adam J Russak, Felix Richter, Shan P Zhao, Akhil Vaid, Fayzan F. Chaudhry, Jessica K De Freitas, Nidhi Naik, R. Miotto, Girish N. Nadkarni, Jagat Narula, Edgar Argulian, and Benjamin S. Glicksberg. Deep learning and the electrocardiogram: review of the current state-of-the-art. *Europace*, 23:1179 – 1191, 2021.

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: 1519–1528, 2021.

Patrick Wagner, Nils Strodthoff, R. Bousseljot, D. Kreiseler, F. Lunze, W. Samek, and T. Schaeffter. Ptb-xl, a large publicly available electrocardiography dataset. *Scientific Data*, 7, 2020.

Zhiguang Wang and Tim Oates. Encoding time series as images for visual inspection and classification using tiled convolutional neural networks. 2014.

Huiyi Wu, Kiran Haresh Kumar Patel, Xinyang Li, Bowen Zhang, Christoforos Galazis, Nikesh Bajaj, Arunashis Sau, Xili Shi, Lin Sun, Yanda Tao, et al. A fully-automated paper ecg digitisation algorithm using deep learning. *Scientific Reports*, 12 (1):20963, 2022.

Chen Xie, Lucas McCullum, Alistair Johnson, Tom Pollard, Brian Gow, and Benjamin Moody. Waveform database software package (wfdb) for python. *PhysioNet*.

Xinzhe Yuan, Dustin Tanksley, Pu Jiao, Liujun Li, Genda Chen, and Donald C. Wunsch. Encoding time-series ground motions as images for convolutional neural networks-based seismic damage evaluation. In *Frontiers in Built Environment*, 2021.

Deyun Zhang, Shijia Geng, Yang Zhou, Weilun Xu, Guodong Wei, Kai Wang, Jie Yu, Qiang Zhu, Yongkui Li, Yonghong Zhao, et al. Artificial intelligence system for detection and screening of cardiac abnormalities using electrocardiogram images. *arXiv preprint arXiv:2302.10301*, 2023.

Wei-Long Zheng, Jia-Yi Zhu, Yong Peng, and Bao-Liang Lu. Eeg-based emotion classification using deep belief networks. *2014 IEEE International Conference on Multimedia and Expo (ICME)*, pages 1–6, 2014.

Jiacheng Zhu, Jielin Qiu, Zhuolin Yang, Douglas Weber, Michael A. Rosenberg, Emerson Liu, Bo Li, and Ding Zhao. Geoecg: Data augmentation via wasserstein geodesic perturbation for robust electrocardiogram prediction. In *Machine Learning in Health Care*, 2022.

Appendix A. Model Parameters

Table 4: Model parameters in the experiments.

Parameters	Value
alpha	0.4
k_test	128
weight decay	0.05
queue size	57600

ViT-base	Value
train batch size	16
test batch size	16
ViT layer	4
init lr	1e-5

ViT-large	Value
train batch size	16
test batch size	16
ViT layer	10
init lr	5e-6

Appendix B. Encoding Methods

B.0.1. Markov Transition Field (MTF)

Markov Transition Field (MTF) is a method of transforming time series data, such as ECG signals, into visual representations. MTF works by calculating transition probabilities between adjacent data points in a time series, and then using these probabilities to generate a matrix of color-coded pixels. Each pixel in the matrix represents a unique transition probability, with darker colors indicating higher probabilities and lighter colors indicating lower probabilities. This matrix can be thought of as an image that encapsulates the key features of the original time series, making it easier for researchers and clinicians to analyze and interpret ECG signals. The development of the Markov Transition Field (MTF) draws inspiration from prior research on the interrelationship between time series and complex networks (Campanharo et al., 2011; Zheng et al., 2014; Wang and Oates, 2014). In essence, the MTF methodology involves constructing a Markov matrix based on quantile bins, which are derived through the discretization of the time series data. The dynamic transition probability of the time series is then encoded into a quasi-Gramian matrix, facilitating further analysis and interpretation of the underlying complex system.

In order to preserve time-domain information, the proposed method leverages Markov transfer probability to represent the dynamics of a given time series X. Specifically, the Q quantile bins are identified, and each data point x_i is assigned to its corresponding bin $q_j (j \in [1, Q])$. The resulting weighted adjacency matrix W, constructed using a first-order Markov chain model along the time axis, reflects the transition probabilities among the quantile bins. The frequency with which a data point in quantile bin q_j is followed by a point in bin q_i determines the value of the corresponding entry $w_{i,j}$ in W. Although W represents the Markov transition matrix after normalization by $\sum_j w_{ij} = 1$, it is insensitive to the distribution of X and the temporal dependencies between time steps t_i, resulting in a loss of information. To address this issue, the Markov Transition Field (MTF) M is defined as follows:

$$\begin{bmatrix} w_{ij|x_1 \in q_i, x_1 \in q_j} & w_{ij|x_1 \in q_i, x_2 \in q_j} & \cdots & w_{ij|x_1 \in q_i, x_n \in q_j} \\ w_{ij|x_2 \in q_i, x_1 \in q_j} & w_{ij|x_2 \in q_i, x_2 \in q_j} & \cdots & w_{ij|x_2 \in q_i, x_n \in q_j} \\ \vdots & \vdots & \ddots & \vdots \\ w_{ij|x_n \in q_i, x_1 \in q_j} & w_{ij|x_n \in q_i, x_2 \in q_j} & \cdots & w_{ij|x_n \in q_i, x_n \in q_j} \end{bmatrix} \tag{15}$$

It involves building a $Q \times Q$ Markov transition matrix W by dividing the time series data into Q quantile bins, where q_i and $q_j (q \in [1, Q])$ represent the quantile bins that contain the data at time stamps i and j along the temporal axis. The MTF matrix M encodes the transition probabilities of the time series by spreading out the transition probability values from matrix W along the magnitude axis to M while taking into consideration the temporal positions. At each pixel M_{ij}, the probability of transitioning from the quantile at time step i to the quantile at time step j is assigned. In this way, the MTF matrix M captures the multi-span transition probabilities of the time series. The entry $M_{i,j||i-j|=k}$ in M represents the transition probability between points with a time interval of k, where $M_{i,j||j-i} = 1$ represents the transition process along the time axis with a skip step. The main diagonal M_{ii} in M is a special case when $k = 0$ and captures the probability of transitioning from each quantile to itself, i.e., the self-transition probability, at time step i.

B.0.2. Gramian Angular Field (GAF)

Gramian Angular Field (GAF) (Wang and Oates, 2014) is another method for transforming ECG time series signals into visual representations. GAF generates a matrix of cosine and sine values based on the pairwise differences between the original data points in the time series. This matrix is then transformed into an image, where each pixel corresponds to a particular combination of cosine and sine values. Similar to MTF, the resulting image captures important features of the original ECG signal, such as patterns and trends, which can aid in the interpretation and analysis of the data. The advantage of GAF over MTF is that it preserves the phase information of the original time series, which can be important in some applications, such as detecting arrhythmias.

The Gramian Angular Field (GAF) (Wang and Oates, 2014) method represents time series data in a polar coordinate system instead of using the traditional Cartesian coordinates. In the Gramian matrix of GAF, each element corresponds to the cosine of the summation of angles. The rescaled time series $\tilde{X}$ of n real-valued observations are transformed to fall within the range of $[-1, 1]$ or $[0, 1]$ using the formula:

$$\tilde{x}^i_{-1} = \frac{(x_i - max(X) + (x_i - min(X))}{max(X) - min(X)} \tag{16}$$

$$or \quad \tilde{x}^i_0 = \frac{x_i - min(X)}{max(X) - min(X)} \tag{17}$$

Then, by encoding the value as the angular cosine and the time stamp as the radius, we represent the rescaled time series $\tilde{X}$ in polar coordinates as follows:

$$\phi = arccos(\tilde{x}_i), \quad -1 \leq \tilde{x}_i \leq 1, \quad \tilde{x}_i \in \tilde{X}, \quad r = \frac{t_i}{N}, \quad t_i \in N \tag{18}$$

Here, t_i is the time stamp, and N is a constant factor that regulates the span of the polar coordinate system. This encoding technique is a novel way to visualize time series data, where the values transform among different angular positions on the spanning circles as time passes, resembling water rippling. The encoding map is bijective, and it preserves absolute temporal relations, unlike Cartesian coordinates. The angular cosine function is monotonic for $\phi \in [0, \pi]$, producing a unique result in the polar coordinate system with a one-to-one inverse map.

Rescaled data in different intervals have different angular bounds. [0,1] corresponds to the cosine function in $[0, \pi/2]$, while cosine values in the interval $[-1$,1] fall into the angular bounds $[0, \pi]$. They can provide different information granularity in the Gramian Angular Field for classification tasks, and the Gramian Angular Difference Field (GADF) of [0,1] rescaled data has an accurate inverse map.

We utilize the angular perspective of the polar coordinate system to examine temporal correlations between different time intervals by calculating the trigonometric sum/difference between each point. Specifically, we define the Gramian Summation Angular Field (GASF) and Gramian Difference Angular Field (GADF) as follows:

$$GASF = [cos(\phi_i + \phi_j)] = \tilde{X}' \cdot \tilde{X} - \sqrt{I - \tilde{X}^2}' \cdot \sqrt{I - \tilde{X}^2} \tag{19}$$

$$GADF = [sin(\phi_i - \phi_j)] = \sqrt{I - \tilde{X}^2}' \cdot \tilde{X} - \tilde{X}' \cdot \sqrt{I - \tilde{X}^2} \tag{20}$$

Here, I is the unit row vector $[1, 1, ..., 1]$. After transforming the time series into the polar coordinate system, we treat each time step as a 1-D metric space. Defining the inner product as follows:

$$< x, y >_1 = x \cdot y - \sqrt{1 - x^2} \cdot \sqrt{1 - y^2} \tag{21}$$

$$< x, y >_2 = \sqrt{1 - x^2} \cdot y - x \cdot \sqrt{1 - y^2} \tag{22}$$

The two types of Gramian Angular Fields (GAFs) are actually quasi-Gramian matrices $[< \tilde{x_1}, \tilde{x_1} >]$.

The Gramian Angular Fields (GAFs) offer multiple benefits. First, they enable the retention of temporal relationships, as the position's movement from the top-left to the bottom-right corresponds to the increase in time. The GAFs incorporate temporal correlations since $G_{i,j||i-j|=k}$ symbolizes the relative correlation due to the superimposition/difference of directions concerning time interval k. The main diagonal $G_{i,i}$ is a special case for $k = 0$, containing the original angular/value information.

B.0.3. Recurrence Plot (RP)

Recurrence Plot (RP) (Eckmann et al., 1987) is a non-linear time series analysis technique that can also be applied to transform ECG time series signals into visual representations. RP generates a square matrix that reflects the similarity between all pairs of data points in the time series. The matrix is constructed by measuring the distance between each pair of data points and comparing them to a predefined threshold value. If the distance between two points is below the threshold, the corresponding matrix element is set to 1, otherwise, it is set to 0. This results in a binary matrix that can be visualized as an image, where dark pixels represent recurrent patterns in the time series. RP has been shown to be effective in capturing complex patterns in ECG signals, such as P-waves and QRS complexes, which are important for the accurate diagnosis of cardiovascular diseases.

Given a time series $(x_1, \ldots, x_n)$, we can extract trajectories from it as follows:

$$\boldsymbol{x}i = (x_i, xi + \tau, \ldots, x_{i+(m-1)\tau}), \quad \forall i \in 1, \ldots, n - (m-1)\tau \tag{23}$$

Here, m denotes the dimension of the trajectories, and τ is the time delay. Once we have extracted the trajectories, we can create a recurrence plot, denoted by R, which is essentially the pairwise distance between the trajectories. Formally, we define $R_{i,j}$ as:

$$R_{i,j} = \Theta(\varepsilon - |\boldsymbol{x}_i - \boldsymbol{x}_j|), \quad \forall i, j \in 1, \ldots, n - (m-1)\tau \tag{24}$$

Here, Θ is the Heaviside step function, and ε is the threshold. The recurrence plot helps us visualize the structure and patterns of the time series by preserving the temporal dependencies and revealing the relative correlations between the extracted trajectories.

Appendix C. Loss Objectives

There are three objectives during learning, including Image-Text Contrastive (ITC) Loss, Image-Text Matching (ITM) Loss, and Mask Language Modeling (MLM) Loss. An overview of each loss is provided in the subsequent sections. More details can be found in the Appendix due to the page limit.

Image-Text Contrastive Loss (ITC) The Image-Text Contrastive Loss (ITC) loss has been shown to be highly effective in improving vision and language understanding in a range of applications, including image captioning, visual question answering, and multimodal retrieval (Radford et al., 2021; Li et al., 2021). To compute the ITC loss, we follow the approach proposed by Li et al. (2021), which introduces a momentum encoder to generate features and creates soft labels from the momentum encoder to serve as training targets. The soft labels help account for the potential positive samples in the negative pairs and improve the quality of the learned representations. Our model learns a similarity function represented by $s = g_v(\boldsymbol{v}_{\text{cls}})^\top g_w(\boldsymbol{w}_{\text{cls}})$, which aims to increase the similarity scores for matching image-text pairs. Here, g_v and g_w refer to linear transformations that convert the `[CLS]` embeddings into lower-dimensional, normalized (256-d) representations. Following the MoCo approach (He et al., 2020), we use two queues to store the most recent M image-text representations obtained from the momentum unimodal encoders. The features obtained from the momentum encoders are normalized and denoted by $g'_v(\boldsymbol{v}'_{\text{cls}})$ and $g'_w(\boldsymbol{w}'_{\text{cls}})$. To calculate the similarity score between an image-text pair and a text-image pair, we define $s(I,T) = g_v(\boldsymbol{v}_{\text{cls}})^\top g'_w(\boldsymbol{w}'_{\text{cls}})$ and $s(T,I) = g_w(\boldsymbol{w}_{\text{cls}})^\top g'_v(\boldsymbol{v}'_{\text{cls}})$, respectively.

We use the softmax-normalized image-to-text and text-to-image similarity to calculate each image and text. This is represented by the equations below, where τ is a temperature parameter that can be learned:

$$p_m^{\text{i2t}}(I) = \frac{\exp(s(I,T_m)/\tau)}{\sum_{m=1}^{M}\exp(s(I,T_m)/\tau)}, \tag{25}$$

$$p_m^{\text{t2i}}(T) = \frac{\exp(s(T,I_m)/\tau)}{\sum_{m=1}^{M}\exp(s(T,I_m)/\tau)} \tag{26}$$

We represent the ground-truth one-hot similarity as $\boldsymbol{y}^{\text{i2t}}(I)$ and $\boldsymbol{y}^{\text{t2i}}(T)$, where negative pairs have a probability of 0, and the positive pair has a probability of 1. The image-text contrastive loss is defined as the cross-entropy H between $\boldsymbol{p}$ and $\boldsymbol{y}$, which is shown in the following equation:

$$\mathcal{L}\text{itc} = \frac{1}{2}\mathbb{E}(I,T) \sim D\left[\text{H}(\boldsymbol{y}^{\text{i2t}}(I), \boldsymbol{p}^{\text{i2t}}(I)) + \text{H}(\boldsymbol{y}^{\text{t2i}}(T), \boldsymbol{p}^{\text{t2i}}(T))\right] \tag{27}$$

Image-Text Matching Loss (ITM) The Image-Text Matching Loss (ITM) is responsible for activating the image-grounded text encoder, with the goal of learning a multimodal representation that captures the detailed alignment between visual and linguistic information. ITM is a binary classification task where the model predicts whether an image-text pair is positive (matched) or negative (unmatched) based on its multimodal feature. The ITM head, which is a linear layer, is used to make this prediction.

To obtain the joint representation of the image-text pair, we use the output embedding of the `[CLS]` token from the multimodal encoder, and then append a fully-connected (FC) layer followed by softmax to predict a two-class probability p^{itm}. The ITM loss is defined as:

$$\mathcal{L}\text{itm} = \mathbb{E}(I,T) \sim D\text{H}(\boldsymbol{y}^{\text{itm}}, \boldsymbol{p}^{\text{itm}}(I,T)) \tag{28}$$

where $\boldsymbol{y}^{\text{itm}}$ is a 2-dimensional one-hot vector representing the ground-truth label. To improve the selection of negative pairs, we employ a strategy called hard negative mining, as proposed by Li et al. (2021). This strategy involves selecting negative pairs that have a higher contrastive similarity within a batch, as they are more informative and can improve the learning process.

Mask Language Modeling Loss (MLM) The Mask Language Modeling Loss (MLM) is used to predict masked words using both the image and contextual text. In this loss, we randomly mask out input tokens with a probability of 15% and replace them with the special token `[MASK]`, with 10% random tokens, 10% unchanged, and 80% `[MASK]` replacements following the BERT approach. The predicted probability of a masked token is denoted by $\boldsymbol{p}^{\text{msk}}(I,\hat{T})$, where $\hat{T}$ represents the masked text. The cross-entropy loss is used to minimize the difference between the predicted and ground-truth distributions, which is expressed as follows:

$$\mathcal{L}\text{mlm} = \mathbb{E}(I,\hat{T}) \sim D\text{H}(\boldsymbol{y}^{\text{msk}}, \boldsymbol{p}^{\text{msk}}(I,\hat{T})) \tag{29}$$

$\boldsymbol{y}^{\text{msk}}$ represents a one-hot vocabulary distribution and the ground-truth token has a probability of 1.

Appendix D. More Related Works

Machine Learning in ECG With the development of machine learning and deep learning, many works have studied the application of using advanced models in ECG. Alfaras et al. (2019) proposed a fully automatic and fast ECG arrhythmia classifier based on a simple brain-inspired machine learning approach known as Echo State Networks. Mishra et al. (2020) converted ECG paper records into a 1-D signal and generated an accurate diagnosis of heart-related problems using deep learning. Peimankar and Puthusserypady (2020) combined CNN and long LSTM model to detect the onset, peak, and offset of different heartbeat waveforms such as the P-wave, QRS complex, T-wave, and No wave (NW). Aziz et al. (2021) exploited two-event related moving-averages (TERMA) and fractional-Fourier-transform (FrFT) algorithms. Somani et al. (2021) proposed a review focusing on orienting the clinician towards fundamental tenets of deep learning, state-of-the-art prior to its use for ECG analysis, and current applications of deep learning on ECGs. Kim et al. (2022) proposed a ML model for real-time classification of atrial fibrillation (AF) between Paroxysmal atrial fibrillation (PAF) and Non-paroxysmal atrial fibrillation (Non-PAF). Adedinsewo et al. (2022) evaluated how well the AI-ECG model output obtained using digitized paper ECGs agreed with the predictions from the native digital ECGs for the detection of low ejection fraction. Ayano et al. (2022) summarized the achievements in ECG signal interpretation using interpretable machine learning techniques. Qiu et al. (2023a) proposed an approach for cardiovascular disease diagnosis and automatic ECG diagnosis report generation. Zhu et al. (2022); Qiu et al. (2023b, 2022) proposed a physiologically-inspired data augmentation method to improve performance and increase the robustness of heart disease detection based on ECG signals.

Transform Time Series Signals into Images Encoding time series data as different types of images have been explored by many studies in different areas (Wang and Oates, 2014). Hatami et al. (2017) used Recurrence Plots (RP) to transform time series into 2D texture images. Kavasidis et al. (2017) converted brain signals into images. Martínez-Arellano et al. (2019) proposed an approach for tool wear classification based on signal imaging. Barra et al. (2020) encoded financial time series into images to predict the future trend of the U.S. market. Qin et al. (2020) encoded time series of sensor data as images to retain necessary features for human activity recognition. Bi et al. (2021) transformed time series into images to improve the accuracy of tourism demand forecasting. Yuan et al. (2021) developed a new image encoding technique based on time-series segmentation (TS) to transform acceleration (A), velocity (V), and displacement (D) ground motion records into a three-channel AVD image of the ground motion event. Sayed et al. (2023) transformed multivariate time-series data into images for better encoding and extracting relevant features for non-intrusive occupancy detection.