Cardiac Evidence Mining for Eating Monitoring using Collocative Electrocardiogram Imagining

This paper was downloaded from TechRxiv (<https://www.techrxiv.org>).

LICENSE

CC BY-NC-SA 4.0

SUBMISSION DATE / POSTED DATE

09-01-2022 / 19-01-2022

CITATION

Zhang, Xulu; Wei, Xiaoyong; Yang, Zhenqun; Jiang, Dongmei; Liao, Ga; Li, Qing; et al. (2022): Cardiac Evidence Mining for Eating Monitoring using Collocative Electrocardiogram Imagining. TechRxiv. Preprint. <https://doi.org/10.36227/techrxiv.18093275.v1>

DOI

[10.36227/techrxiv.18093275.v1](https://doi.org/10.36227/techrxiv.18093275.v1)

This work has been submitted to the IEEE for possible publication. Copyright may be transferred without notice, after which this version may no longer be accessible.

Cardiac Evidence Mining for Eating Monitoring using Collocative Electrocardiogram Imaging

Xu-Lu Zhang, Zhen-Qun Yang, Dong-Mei Jiang, Ga Liao, Qing Li, *Fellow, IEEE*, Ramesh Jain, *Fellow, IEEE*, Xiao-Yong Wei, *Senior Member, IEEE*

Abstract—Eating monitoring has remained an open challenge in medical research for years due to the lack of non-invasive sensors for continuous monitoring and the reliable methods for automatic behavior detection. In this paper, we present a pilot study using the wearable 24-hour ECG for sensing and tailoring the sophisticated deep learning for ad-hoc and interpretable detection. This is accomplished using a collocative learning framework in which 1) we construct collocative tensors as pseudo-images from 1D ECG signals to improve the feasibility of 2D image-based deep models; 2) we formulate the cardiac logic of analyzing the ECG data in a comparative way as periodic attention regulators so as to guide the deep inference to collect evidence in a human comprehensible manner; and 3) we improve the interpretability of the framework by enabling the backtracking of evidence with a set of methods designed for Class Activation Mapping (CAM) decoding and decision tree/forest generation. The effectiveness of the proposed framework has been validated on the largest ECG dataset of eating behavior with superior performance over conventional models, and its capacity of cardiac evidence mining has also been verified through the consistency of the evidence it backtracked and that of the previous medical studies.

Index Terms — Collocative Electrocardiogram Imaging; Eating Monitoring; Cardiac Evidence Mining.

I. INTRODUCTION

EATING monitoring is of great importance to eating pathology (e.g., obesity, binge eating, eating disorders) and essential to chronic disease management (e.g., diabetes, hypertension, and a wide range of other cardiovascular diseases). Users (patients) are usually required to record the time, duration, location, and food whenever they are taking any food. It is a tedious work which requires intensive cooperation from the users and is subject to recall bias. A lot of efforts towards eating monitoring have been made recently (e.g., GPS for the location, food image recognition for dishes and ingredients [1]). However, the automatic eating monitoring (AEM), which detects the time and duration of eating in a non-invasive way, has not been addressed well.

Studies show that the eating activity is closely linked to heart movement as food digestion will affect how the heart

pumps blood [2]–[4]. However, these studies are still far from creating practical methods for AEM, because they have been limited in the scope of qualitative medical research. The 24-hour Electrocardiogram (ECG) monitoring is an emerging technique to track the continuous heart activity for cardiovascular health problems, and thus theoretically can be used for AEM. Nevertheless, our study shows it only gives a moderate performance by using the features found in previous qualitative studies (e.g., heart rates or heart rate variability) [2].

To build a better detector based on the 24-hour ECG, one may naturally think of it as a signal processing problem which can be solved using the advanced deep learning methods. Despite its success in other signal processing applications, this handy solution does not make much sense in a cardiology point of view, because the cardiology analysis is based more on the *comparative* features between cardiac waves rather than the statistical features used in signal processing. Furthermore, in cardiac analysis, features or models need to be *interpretable*, because either the waves themselves or the inter-wave relation (e.g., intervals) have their cardiac seedbed, and it is more important for a model to locate the evidence than to make the prediction. This makes the nontransparent deep models less feasible to this problem. Moreover, there is an even more fundamental issue which lies in the 1D and *periodic* nature of the ECG data, on which the majority of the sophisticated deep models (e.g., VGG [5], ResNet [6], Inception-v3 [7]) cannot play to their strength, because these models are built on 2D image-based data and the periodic characteristics have not been considered in the design.

In this paper, we present a pilot study of using 24-hour ECG for non-invasive and continuous eating monitoring. A deep learning framework is proposed to guide the learning towards human comprehensible inference and thus enable the cardiac evidence backtracking. The framework has addressed the aforementioned issues (i.e., comparative analysis, interpretability, periodic nature, and 1D-to-2D gap). More specifically, we build a new collocative learning [8] scheme to transfer the 1D ECG signals into 2D pseudo-images which encapsulates the comparative features among ECG segments, so that we can bring the sophisticated deep models into full play. In addition, a coaching attention mechanism is designed to adapt these models to the periodic nature of ECG signals. More importantly, we show that within the framework, we are able to backtrack the cardiac evidence of the predictions in the form of deep learning favored saliency maps and human comprehensible representations. The framework is shown in Fig. 1. It has advantages over existing methods as follows

- **Comparative Relation:** We transform the 1D ECG signals into collocative tensors (based on 2D views) in which the comparative relations are encapsulated with

Xiao-Yong Wei (e-mail: cswei@scu.edu.cn) and Xu-Lu Zhang are with the College of Computer Science, Sichuan University, Chengdu 610065, China
Zhen-Qun Yang is with the Department of Biomedical Engineering, Chinese University of Hong Kong, Kowloon, Hong Kong (e-mail: jes-sicayang@cuhk.edu.hk).

Dong-Mei Jiang is with the School of Computer Science, Northwestern Polytechnical University, Xi'an, China

Ga Liao is with the State Key Laboratory of Oral Diseases, National Clinical Research Center for Oral Diseases, West China Hospital of Stomatology of Sichuan University, Chengdu 610041, China.

Qing Li and Xiao-Yong Wei are with the Department of Computing, Hong Kong Polytechnic University, Kowloon, Hong Kong

Ramesh Jain is with the Institute for Future Health, University of California, Irvine, CA, USA, and Donald Bren School of Information and Computer Sciences, University of California, Irvine, CA, USA.

Fig. 1. The collocative framework for modeling the comparative relations and cardiac evidence backtracking in Electrocardiograms. It consists of three parts: 1) the transformation of 1D ECG signals into 2D views to be adapted to the conventional deep models; 2) the periodic attention coach which guides the deep inference with human logic of comparative analysis; 3) the evidence backtracking based on saliency maps and the generation of human comprehensible representations of the evidence and inference logic.

the inter-segment correlation. The tensors are pseudo-images which are compatible to the majority of deep models. It gives us the opportunity of analyzing the ECG data in a image liked way. Moreover, we will show in Section III-A that the multi-view setting is able to create a more comprehensive representation for deep learning.

- **Period Sensing:** With the multi-view tensors, we will see in Section III-B that the morphological patterns of waves will be converted into “backslashes” in the pseudo-images, on the basis of which we have designed a periodic attention template that is able to adapt its period to the distance between or length of the “backslashes”. This has transformed the problem of learning the period of signals into a distance/length sensing problem which is easier to be addressed with image-based deep models.
- **Interpretable Evidence Backtracking:** With the periodic attention template, we can guide the deep models to select only “backslashes” that repeat with a certain period so as to imitate the human logic of evidence building. This is reflected in both the tensor- and wave-level saliency maps. By decoding the saliency maps, we are able to backtrack the evidence (i.e., the comparative features of wave pairs) and finally represent the evidence and inference logic in a human comprehensible way (i.e., decision trees/forests), which can help the technicians to identify cardiac seedbed of the eating behavior.

The contribution of this paper includes: we have conducted a pilot study of ECG-based eating behavior logging/monitoring, and have brought image-based deep learning methods into this area for comparative relation mining and interpretable evidence backtracking. In addition, we will also release the first and largest ECG dataset for eating behavior analysis to

promote the research in this area.

II. RELATED WORK

A. Wearable Eating Behavior Monitoring

Eating is essential for the health of human beings. The monitoring of eating is thus of great importance to either health management of the general population or eating pathology for patients. Through its long history of thousands of years, the monitoring has been mainly based on self-reporting, which is subjective and with recall bias [9]. This brings forward automatic methods especially those using wearable devices. A wide range of efforts has been made in the last two decades [10]–[18]. Various sensors have been applied, such as acoustic sensors using microphones [10], visual sensors using cameras [17], wrist or arm motion sensor using accelerometers [11], physiological sensors using EMG [12], piezoelectric sensors [13], and multi-sensor by combining these sensors [14].

The choice of sensors is a choice of balance between the comfort and reliability. For example, microphones are widely used for monitoring through the sound of chewing and swallowing. The outer-ear [15] or neck [16] microphones might be comfort to wear, but less reliable due to the disturbance of the environmental noise. The inner-ear microphones are reliable but less user friendly. In addition, in most applications (especially when users are patients), more important is that whether the sensors can provide pathological evidence for in-depth analysis of the users’ condition. The EMG or piezoelectric sensors are good in this regard, because they are connected to the inner processes (e.g, muscle movement and tension) of the body. However, these sensors are less acceptable to users as they have to be installed around the neck.

In this paper, we study the possibility of using the 24-hour ECG sensors. As a non-invasive method, it is comparatively

comfort, and most importantly, it has been widely used in collecting both the physiological and psychological evidence [19], because it is associated to the heart which is one of the most important organs of our body and determines both the physical and mental health.

B. Deep Learning for ECG Analysis

In terms of learning methods employed for eating monitoring, although neural network-based models have also been explored by a few researchers [20], conventional models, such as HMM [21], SVM [22], random forests [23], are still dominating in this area, because it is straightforward to consider the monitoring as a 1D signal processing (detection) problem. Early work in [20] outperforms SVM with a 3-layer neural network, but it remains as a conventional shallow network which is not leveraging the power of deep learning.

Beyond the task of eating behavior monitoring, deep learning has already been adopted in a number of studies for ECG analysis [24]–[31]. A majority of them are for arrhythmia classification [25], [27] together with a few for emotion recognition [30] or atrial fibrillation detection [26]. Main-stream methods are either to feed the 1D ECG signals to the 1D CNNs (ResNet or VGG) directly [25]–[27] or to adopt 2D CNNs after transforming the 1D inputs into 2D spectrograms or simply reshaping the signals into 2D [24], [28], [29], [31]. None of these methods has modeled the periodic nature of the signal. Furthermore, none of these methods has modeled the high order relation (inter-wave comparison) specifically, and thus is limited to provide explicit evidence for cardiac analysis. In this paper, we address this problem with collocative learning which is capable of period sensing and evidence backtracking. This is implemented through a periodic attention coaching mechanism and a CAM-based decoding process for backtracking and human comprehensible representation generation.

C. Attention and CAM Methods

Attention mechanisms have been employed in deep learning for weighting the feature maps so that the highly important regions can stand out from the feed-forward inference [32]–[34]. The attention usually appears as a weighting mask which has been learned through back-propagation adaptively, meaning the mask formation is data-driven, network dependent, and without human guidance. The coached attention gates (Section III-B2) we proposed is different from standard attention mechanisms in a way that we define a parameterized template mask with embedded periodic patterns and allow the back-propagation to learn the parameters only rather than the mask directly. It is a combination of prior knowledge injection and adaptive learning.

CAM-based methods are often used to generate saliency maps which indicate which parts of the inputs have played a more significant role in the inference than others [35]–[37]. However, the saliency maps are usually with an arbitrary shape for providing indicative rather than explicit evidence (e.g., exact localization of waves or wave pairs). In this paper, we have built a decoding scheme (see Section III-C) on top of the Grad-CAM [36], which is able to convert the indicative evidence into explicit evidence for generating human comprehensible representations.

In terms of collocative learning, this paper is also related to [8] in which the prototype framework of collocative learning has been proposed. In this paper, we have followed the thread but have implemented the learning in a different way (see Section III). Most importantly, this framework in this paper has been designed specifically for the ECG-based eating behavior monitoring which is a completely different task from the Immunofixation Electrophoresis analysis in [8].

III. COMPARATIVE RELATION MODELING AND CARDIAC EVIDENCE BACKTRACKING

A. Multi-view ECG Tensor Construction

1) *Inter-Segment Relation Encoding*: It is straightforward to convert a 1D signal S into an “image” by cutting a signal into n segments and then encapsulating the inter-segment correlations/distances into a $n \times n$ matrix $\mathbf{R}$. However, it will lose the original 1D features despite the inter-segment relation has been encoded. To address this issue, we can encode the original features into the diagonal of the matrix $\mathbf{R}$. It is a near-lossless replacement as the diagonal elements are all ones (correlation) or zeros (distance). It can be formulated as

$$\mathbf{R} = \underbrace{\mathbf{R} \odot (\mathbf{I} - \mathbf{E})}_{\text{diagonal mask-out}} + \underbrace{(\bar{\mathbf{1}}\bar{\mathbf{x}}^\top) \odot \mathbf{E}}_{\text{segment feature encoding}} \quad (1)$$

$$\bar{\mathbf{x}} = f(S), S = \{s_i | 1 \leq i \leq n, i \in \mathbb{Z}^+\}$$

where the $\odot$ is the Hadamard product, $\mathbf{E}$ is a $n \times n$ identity matrix, $\bar{\mathbf{1}}$ is a $n \times 1$ all-ones vector, and $\top$ denotes transpose operator. The $\bar{\mathbf{x}}$ is a $n \times 1$ feature vector where the i^{th} element is the original feature value of the i^{th} segment of the ECG signal S which can be extracted by any convectional signal processing function $f(\cdot)$. Eq. (1) equals to a process that “masks out” the diagonal of $\mathbf{R}$ and then replaces with the original segment features.

2) *Composing Relation Matrices into A Tensor*: The resulting $\mathbf{R}$ is indeed a function of the metric used to calculate the correlations/distances (denote as $\phi(s_i, s_j)$ hereafter) and the feature extraction function $f(s_i)$. It is more representative than the original data because of the binary and comparative relation (of two segments) extracted by $\phi(s_i, s_j)$ and the unary relation (of a single segment) extracted by $f(s_i)$. By varying $\phi(\cdot)$ and $f(\cdot)$, we can generate a set of $\mathbf{R}$ s that represent our understanding about the original ECG data from different point of views. The multi-view tensor $\mathcal{T}$ can then be defined as an order-3 tensor composed by the concatenation of these $\mathbf{R}$ s as

$$\mathcal{T} = \{\mathbf{R}_{f_1, \phi_1}, \mathbf{R}_{f_2, \phi_2}, \dots, \mathbf{R}_{f_m, \phi_m}\} = \bigcup_{i=1}^m \mathbf{R}_{f_i, \phi_i} \quad (2)$$

where m denotes the number of views. The f_i can be selected from a wide range of sophisticated signal processing functions, and the ϕ_i can be implemented by metrics such as cosine similarity, Euclidean distance, Manhattan distance, and Mahalanobis distance.

B. Periodic Attention Regulation

With the multi-view tensor as the input to deep networks, we can build an attention regulator to guide the inference focus more on the comparative features with the periodic nature of ECG signal taken into consideration.

Fig. 2. The comparative relations among segments have already been encoded as “backslashes” into the collocative tensors. In the figure, several backslashes for the QRS complex are shown. It can be used to model the comparative relations of Q, R, and S waves. The duration has also been encoded as the length of backslashes. The period is encoded into the distances among backslashes.

1) *Periodic and Comparative Features Encoded in $\mathcal{T}$* : The first challenge is that there is no existing image-based attention mechanism which has specifically addressed the periodic nature of the original signal. By reviewing our input tensor $\mathcal{T}$, we can see that the signal period (denote as T hereafter) is already encoded, in a way that the correlation/distance will also repeat while the signal repeats. In the tensor $\mathcal{T}$, we will see in Fig. 2 that if a comparative pattern of the 1D signal is repeating consistently, it will form in $\mathcal{T}$ multiple “backslashes” with distances among them indicating the periods.

Let us review how a human technician detects such patterns for a more intuitive explanation through an example. QRS complex, which consists of 3 subwaves of Q, R, and S, is one of the most frequently used “makers” for the cardiovascular examination, with which a technician can identify from its morphology whether the depolarization (activation) of the ventricles functions properly. More specifically, a QRS usually has a unique appearance with a sharp peak at the R and two short valleys at the Q and S. As shown in Fig. 2, this appearance is indeed a trend change of consecutive segments which can be captured by their correlation/distance stored in the input tensor $\mathcal{T}$ where its values from the onset to the end of a QRS form a “backslash”. The backslash repeats itself following the period T resulting in parallel backslashes. It is a distinguishable pattern that can be easily learned by most of the CNN models designed for image-like inputs. Furthermore, the duration of a QRS in fact indicates how strong and fast the ventricles contract to pump blood to other parts of the body. Generally speaking, a normal QRS is short

Fig. 3. The coached periodic attention gates (CAGs). The network of CAGs consists of three layers. The first layer is a single node fixed value 1. The second layer is with 4 parameter nodes of $(\alpha, \beta, \gamma, T)$. The third layer is a parameterized (by the parameters of the second layer) and diagonalized cosine attention mask Ω which will guide the learning to focus on the diagonal “backslash” patterns and sense the period of the signals at the same time.

(or narrow) while a QRS within a ventricular dysrhythmia is wide (or broad). This will be reflected by the length of the QRS backslashes in $\mathcal{T}$. Moreover, we can find from Fig. 2 that the period T can be measured straightforwardly from the distances of the “backslashes”. However, CNN models are usually not specifically designed to learn such distances and their variations. We can address this issue by adding the attention regulation on the CNN inference.

2) *Attention Regulation with Coached Attention Gates*: We have learned that there are three characteristics of the tensor encoding that we need to consider in our attention mechanism design, namely the *back-slashed patterns*, the *distances of backslashes*, and the *periods*. We propose to model these characteristics with the Coached Attention Gates (CAGs). The idea is to use a diagonalized attention mask to coach the networks to only search evidence at the diagonal direction (for backslash modeling), then an attention gate (i.e., a CAG unite) can be formed by repeating the mask with an adaptive period learned from the back-propagation (for distance sensing), and finally we can apply the principal to different layers/blocks of the networks and generate CAG unites to adapt to the period variation at different degrees (for period variation).

CAG Intuition: Let us denote a CAG unite as Ω that is a matrix with the dimensions $(m \times m)$ which are the same as those of the spatial dimensions of a target feature map $\mathcal{F}$ (of shape $m \times m \times Z$, Z denotes the number of channels). The attention regulation can be written as

$$\mathcal{F}_k^* = \mathcal{F}_k \odot \Omega + \mathcal{F}_k, 1 \leq k \leq Z \quad (3)$$

where the k is an iterator to traversing channels of $\mathcal{F}$.

The CAG unite can be implemented with a trigonometric function (e.g., *sine*, *cosine*) along the diagonal, because it is easy to parameterize the function to control its amplitude and period. An illustration of CAG generated using a parameterized *cosine* is shown as the Ω in Fig. 3, from which the coaching process becomes intuitive. Firstly, the diagonalization of

the attention “strips” has confined the way of the networks to collect evidence only from backslash patterns. Secondly, the repetition of the strips has regulated the inter-backslash relation to be periodic. In other words, the more consistent the repetition periods of the strips and backslashes are, the larger “resonance” (i.e., agreement) between the CAG and the feature map reaches, which thus leads to the more robust evidence for inference. Furthermore, the agreement is also an indication of better consistency between the human logic (formulated through the CAG) and the deep network inference. Otherwise, the deep network may collect evidence from patterns of arbitrary shapes if there is no such CAG regulation. Nonetheless, in spite of the rationale of the intuition, we should see that the period T for the *cosine* is not easy to be defined because of the period variation. We will leverage the power of deep learning to “sense” it and also make it adaptive to the variation. Let us formulate CAG to see how this has been implemented.

CAG Formulation: The values of a CAG unite Ω can be obtained from a function $\omega(i, j | \alpha, \beta, \gamma, T) \in [0, 1] : \mathbb{Z}^+ \times \mathbb{Z}^+ \rightarrow \mathbb{R}$ that assigns each location (i, j) an attention value $\Omega_{ij} = \omega(i, j)$ based on a set of learned parameters $(\alpha, \beta, \gamma, T)$ controlling the shape (and offsets) of the *cosine* wave. Putting all together, it is given as

$$\Omega_{i,j} = \omega(i, j | \alpha, \beta, \gamma, T) = \underbrace{\alpha}_{\text{scalar}} \cdot \cos \left(2\pi T \underbrace{\frac{|i-j|}{m}}_{\text{normalizer}} + \underbrace{\beta}_{\text{offset}} \right) + \underbrace{\gamma}_{\text{adjuster}} \quad (4)$$

where α is a scalar on the *cosine* to control the amplitude of the attention, $\frac{|i-j|}{m}$ is a normalizer to convert the spatial locations into the range of the input field of the *cosine*, β controls the offset of the cosine, and γ helps the attention values to fit into a proper range so that it can be used in Eq. (3). Note that from an attention mask point of view, the α determines the weight range between these highly trusted evidence (backslashes) and the less trusted ones, while the β determines how the mask shifts along the anti-diagonal so that the signal input can start at any given point of a cardiac cycle. This setting releases the requirements of detecting the onset of P waves and segmenting the signal into standard cardiac waves, and thus makes the process more feasible, because cardiac wave/peak detection is an open question and ongoing study by itself [38], [39].

CAG Learning: The parameter set $(\alpha, \beta, \gamma, T)$ is not easy to learn in a conventional setting of signal process or cardiac analysis. Fortunately, we can leverage the power of deep learning to figure them out. The CAGs will be organized as an attention network as shown in Fig. 3. The first layer is with only one node of value 1 as the common root of the network. The second layer is a parameter layer consisting of L CAG blocks where L equals to the number of blocks/layers of the host CNN model. Each block is with 4 nodes for α , β , γ , and T respectively. The third layer is a gate generator layer for creating H CAG masks (Ω s) with parameters of the second layer. The resulting CAGs will be planted into each convolutional layer/block of the host CNN model to implement the attention regulation in Eq. (3). In each CAG block, mask nodes in the third layer are fully connected to the

parameter nodes in the second layer, so that the parameters $(\alpha, \beta, \gamma, \text{ and } T)$ can be then learned during the back-propagation. However, we have to modify the gradient updating functions for the second layer since the *cosine* used for connecting nodes between layers is a non-standard DNN schema. With the chain rules, we update the gradients of the (i, j) node of the Ω regarding current prediction Y^c for the class c as

$$\begin{aligned} \frac{\partial Y^c}{\partial \alpha} &= \sum_{i,j} \frac{\partial Y^c}{\partial \Omega_{ij}} \cos \left(2\pi T \frac{|i-j|}{m} + \beta \right), \\ \frac{\partial Y^c}{\partial T} &= - \sum_{i,j} \frac{\partial Y^c}{\partial \Omega_{ij}} \alpha \sin \left(2\pi T \frac{|i-j|}{m} + \beta \right) \frac{2\pi |i-j|}{m}, \\ \frac{\partial Y^c}{\partial \beta} &= - \sum_{i,j} \frac{\partial Y^c}{\partial \Omega_{ij}} \alpha \sin \left(2\pi T \frac{|i-j|}{m} + \beta \right) \end{aligned} \quad (5)$$

where $\frac{\partial Y^c}{\partial \Omega_{ij}}$ and $\frac{\partial Y^c}{\partial \Omega_{ij}}$ can be learned from the standard back-propagation. Note that the period parameter T learned in each block varies from each other. It generates the maximum “resonance” to support the inference when the T s at the early layers approach the cardiac period of the original ECG signal. This helps the inference to “sense” the period at different scales so that the period variation issue can be addressed.

C. Evidence Backtracking and Generation of Human Comprehensible Representations

In this section, we will “decode” the collocative saliency maps and use it to generate human comprehensible evidence representations to assist the cardiac analysis.

1) *Collocative Saliency Map Generation:* We can simply generate a saliency map $\mathcal{S}_l^{\mathcal{F}^*}$, $1 \leq l \leq L$ using Grad-CAM [36] for each of the L layers/blocks of the host CNN model based on the feature map $\mathcal{F}^*$, and then integrate these maps into the final map using average pooling. It is written as

$$\mathcal{S} = \frac{1}{L} \bigoplus_{l=1}^L \mathcal{S}_{\{l\}}^{\mathcal{F}^*}. \quad (6)$$

The resulting $\mathcal{S}$ can be resized to match the spatial dimensions of the collocative tensor $\mathcal{T}$, so that the significance of the comparative feature of any segment pair (s_i, s_j) in the inference can be evaluated quantitatively.

2) *Decoding The Maps Into Wave Scale:* The segments that represent a machine perceptible way of data reading are with a much finer scale than that of the cardiac waves (i.e., human defined segments) used by human technicians. We need to decode the segment-based saliency map $\mathcal{S}$ into wave-based rating to help the human interpretation of the results. With the collocative configuration of $\mathcal{T}$, the map can be decoded into either the rating of unary waves or that of the binary wave-to-wave comparison.

Assume we have a set of waves $W = \{w_p | 1 \leq p \leq |W|, p \in \mathbb{Z}^+\}$ and another set of segments $S = \{s_i | 1 \leq i \leq |S|, i \in \mathbb{Z}^+\}$, we can decode the segment-based saliency map $\mathcal{S}$ into a rating of unary waves by a voting schema in which the weight $\mathcal{S}_{ij}$ (the significance of comparing s_i and s_j) will

be assigned to the waves w_p and w_q respectively if $s_i \in w_p$ and $s_j \in w_q$. This can be implemented as

$$\begin{aligned} \vec{s}_W &= \underbrace{\frac{1}{|S|} \mathcal{M}^\top}_{\text{membership}} \underbrace{\frac{1}{2\|\vec{1}\|} (\mathcal{S}\vec{1} + \mathcal{S}^\top \vec{1})}_{\text{segment-based voting}}, \\ \mathcal{M}_{ip} &= \begin{cases} 1, & s_i \in w_p \\ 0, & s_i \notin w_p \end{cases}, \end{aligned} \quad (7)$$

where the $\mathcal{M}_{ip}$ is a $|S| \times |W|$ binary matrix indicating the membership between segments in S and waves in W , and the $\vec{s}_W$ is the resulting vector indicating the significance of waves.

Similarly, we can implement the decoding for the rating of wave-to-wave comparison by assigning the weight $\mathcal{S}_{ij}$ to the corresponding wave pair (w_p, w_q) if and only if $s_i \in w_p$ and $s_j \in w_q$ hold. Let us denote the rating of wave pairs as a matrix $\mathcal{W}$, and write the membership matrix $\mathcal{M}$ as a set of its column vectors (i.e., $\{\mathcal{M}\} = \{\vec{m}_p\}$ where the i^{th} element of $\vec{m}_p$ indicates if $s_i \in w_p$). We have

$$\mathcal{W}_{pq} = \vec{1}^\top \left(\underbrace{\left[\vec{m}_p (\vec{m}_q)^\top \right]}_{s_i \in w_p \wedge s_j \in w_q} \odot \mathcal{S} \right) \vec{1}, \quad (8)$$

where $\mathcal{W}_{pq}$ implies the significance of comparing two waves w_p and w_q to the inference.

3) *Human Comprehensible Representations*: With the ratings of the waves and pairs, we are able to inspect what kind of evidence has been referred for the decision making in the CNN model. While this improves the transparency and interpretability of the deep learning, the saliency maps are evidence generated more from a machine learning perceptive, which is not necessarily friendly enough to medical technicians. In this section, we convert the saliency maps into human comprehensible representations of decision trees/forests¹.

Attributes: To this regard, first we need to establish the candidate attributes for tree/forest generation. Before we start, hereafter we will use the terms “wave” and “wave genre” differently, because the term “waves” used in cardiology are indeed wave genres while from a data point of view, the waves (e.g., those in the previous section) are instances of wave genres. Our candidate attributes are thus established on wave genres to be cardiac meaningful. Therefore, for the wave-level evidence, the candidates are straightforwardly all wave genres, while for the comparative evidence, the candidates are wave genre pairs. Furthermore, for a signal sample, which is usually composed with several wave instances for a genre, we need to summarize the instances for the genre using statistics (e.g., mean, standard deviation)².

Construction: To construct the trees/forests, we can use the saliency maps as a guide for attribute selection. More specifically, we will select an appropriate subset of attributes from the candidate attributes, which can significantly reduce

the size of the decision trees/forests into human comprehensive levels (i.e., with much less number of attributes and shorter heights)³. More specifically, the saliency information in the maps, which indicates which segments/pairs play more important roles than others, can be considered as a ranking information for selecting attributes from the candidates in an orderly manner. Therefore, we can control the scale of the evidence representations by controlling the maximum number of attributes t_{max} together with the maximum height h_{max} of the trees/forests. Furthermore, the inference logic will be formed in a friendly way for medical technicians through the tree branching. Denote the candidate attribute set as $\{a_i\}$ where a_i can be either a wave genre or a pair of wave genres and all the a_i s are ranked in descending order according to their saliency values in the maps. This process can be formulated as

$$\begin{aligned} \{a_t\}^* | t_{max}, h_{max} &= \arg \max_{\{a_t\} \subseteq \{a_i\}} \rho \left(Tr(\{a_t\}, h) \right), \\ 1 \leq t \leq t_{max}, 1 \leq h \leq h_{max} \end{aligned} \quad (9)$$

where the $\{a_t\}$ is a set consisting the first t attributes in $\{a_i\}$, $\{a_t\}^*$ denotes the optimal set of attributes under the conditions of t_{max} and h_{max} , $Tr(\cdot)$ is a tree/forest builder which takes a subset of attributes $\{a_i\}$ and the height of the tree/forest h as the inputs, and the $\rho(\cdot)$ is a performance evaluator (e.g., Accuracy or F1-score for classification tasks). More details about the implementation can be found in Section IV-F, in which we will conduct experiments to generate and evaluate the trees/forests. More importantly, we can see that the resulting evidence representations can provide more and finer-scale cardiac evidence comparing with that in the previous medical studies. Some of these findings, which have not been reported in the previous studies of eating behavior, not only validate the effectiveness of the proposed method, but also serve as new evidence for cardiology analysis.

IV. EXPERIMENTS

A. Setup

1) *Dataset*: To further study the proposed method, we have employed a desensitized dataset collected by the West China Hospital of Sichuan University. It includes 150,037 samples where each of them is an ECG segment with a duration of 10 seconds and has been labeled by medical technicians with either “eating/digesting” (85,508 samples) or “non-eating” (64,529 samples). The “eating/digesting” samples are selected and labeled by technicians for their regular meals according to the design of the corresponding medical study. Each label is a result based on the agreement of more than 2 (out of 3) technicians. The experiments in this section have been conducted on this Regular Meal dataset⁴. The desensitization has been

³The reduction has to be done here not only for representation generation. Computationally, selecting the optimal subset of attributes is a NP-hard combinational problem which is impractical to solve by brute-force search.

⁴The whole dataset in fact also comes with 29,557 samples for snacks and has been collected from the same of group of volunteers but not fully labeled, because the existing studies on eating are all based on regular meals. Furthermore, the labeled subset is with a less restrict setting (i.e., each sample has only been labeled by 1 technician). However, we will release the snack dataset for reference.

¹The decision trees/forests are selected because they have been considered (among our medical technicians) as the best understandable representations of the inference logic for the classification tasks.

²Among a wide range of features available for summarization, the mean and standard are selected, because it is a common agreement among our medical technicians that they are the most straightforward features.

TABLE I
PERFORMANCE COMPARISON FOR BASE MODEL SELECTION. THE BEST RESULTS ARE IN BOLD FONT.

Model	Accuracy (%)	F1-score (%)	Recall (%)	Precision (%)	TPR (%)	TNR (%)	Model Complexity	Efficiency (s)	Accuracy/Complexity	Accuracy/Efficiency
ResNet18	71.85	71.22	71.09	71.52	77.71	64.47	15.5	1.93	0.05	0.37
ResNet34	71.51	70.93	70.82	71.15	76.88	64.75	29	2.55	0.02	0.28
ResNet50	71.61	70.73	70.55	71.43	79.79	61.32	64.2	3.39	0.01	0.21
VGG16	72.29	71.63	71.49	71.99	78.50	64.47	146	2.14	0.01	0.34
DenseNet121	72.56	72.03	71.92	72.22	77.51	66.33	22.7	5.70	0.03	0.13

done through anonymization and shuffling. To the best of our knowledge, this is the first and largest ECG dataset for eating behavior analysis with high quality labeling. Both the dataset and source code can be found at <https://github.com/XXX> (will open to public upon acceptance).

2) *Evaluation*: We use 10-fold cross validation where in each round, the dataset will be randomly divided into a training and a testing set consisting of 90% and 10% samples of the whole dataset, respectively. The standard metrics of accuracy, F1-score, Recall, and Precision have been employed for evaluating the performance of the classification (e.g., eating or non-eating). True Positive Rate (TPR), and True Negative Rate (TNR) have also been included for reference.

B. Selection of Base Model

We feed the samples into a set of popular CNN models directly (without using any collocative learning processes) to study how these models fit to the problem. The models include ResNet18 [6], ResNet34 [6], ResNet50 [6], VGG16 [5], and DenseNet121 [40]. In addition to the standard metrics, we have included Model Complexity (the number of parameters), and Efficiency (time used in seconds for predicting the labels of samples in the testing set), together with two results of Accuracy/Complexity and Accuracy/Efficiency for evaluating how a model is balanced between the performance and cost.

The results are shown in Table I. The DenseNet121 has obtained the best performance over others. However, the performance gaps among models are minor, indicated by the mean F1-scores at 71.31 with a narrow standard deviation of 0.53. Surprisingly, a better performance is not guaranteed by increasing the complexity of the model. This is particularly obvious if we compare the three ResNet models. The best performance has been obtained by using the simplest model of ResNet18. This holds true in the comparison between the VGG16 and DenseNet121 as well. By contrast, we can see the ResNet18 is the best balanced model which gives a satisfactory performance (with F1-score approaching the mean) and best efficiency (9.81% ~ 66.14% faster than others). Therefore, we select ResNet18 as the “Base Model” to continue the study.

C. Collocative Tensor Construction

To find the best configuration for the collocative tensor construction, we calculate the collocative relation of two ECG segments with standard metrics of Euclidean distance, Manhattan distance, Cosine similarity, and Mahalanobis distance, respectively, to generate a set of views. In addition, we calculate the Maximum, Average, and Minimum values of two segments, respectively, for more views. The multi-view tensor

TABLE II
PERFORMANCE COMPARISON FOR COLLOCATIVE TENSOR CONSTRUCTION WITH DIFFERENT METRICS. THE BEST RESULTS ARE IN BOLD FONT.

Model	Accuracy (%)	F1-score (%)	Recall (%)	Precision (%)	TPR (%)	TNR (%)
Base Model	71.85	71.22	71.09	71.52	77.71	64.47
Euclidean	72.74	72.45	72.51	72.41	74.53	70.49
Manhattan	72.38	72.11	72.19	72.05	73.77	70.62
Cosine	70.82	70.01	69.86	70.55	78.33	61.38
Mahalanobis	70.38	69.85	69.78	69.98	75.10	64.45
Max	70.84	69.75	69.58	70.79	80.60	58.57
Avg	72.84	72.15	71.99	72.59	79.41	64.57
Min	72.69	71.91	71.72	72.52	80.17	63.28
Multi-View	73.57	72.56	72.32	73.76	83.26	61.39

is then constructed by concatenating these single-view tensors following Eq. (2). The results are shown in Table II.

The performance is improved from the Base Model by using most of the relation metrics, and 4 metrics (out of 8), namely the Euclidean, Manhattan, Avg, and Multi-view), even outperform the DenseNet121 (which obtained the best performance in the base model selection). However, we only consider the performance gain as a “by-product” of introducing collocative learning. As we discuss in Section III-C, the key advantage of the proposed methods is to bring possibility for evidence backtracking on the comparative features, so that human understandable knowledge can be extracted for more effective pathological studies. This will be investigated specifically in Section IV-F.

Comparing between multi- and single-view tensor constructions, multi-view obtains the best performance on all metrics except the TNR, resulting from the complementary information encapsulated in its member (single-) views. This validates its effectiveness of integrating single-views for a more comprehensive representation.

D. Periodic Attention Regulation

In this section, we further evaluate the effectiveness of CAGs for period regulation. The results are shown in Table III. It is not surprising that by adding the CAGs as the periodic regulator, the performance of the proposed method has been further improved by 3.52% (3.51%) in Accuracy (F1-score) when compared to that of the Base Model (the multi-view learner). In addition, its superiority over the Base Model has been validated on all metrics, including the TNR on which the multi-view learner has a slightly degraded performance. Let us zoom in for more details.

1) *Period Modeling*: In Fig. 4, we have shown several typical examples of the ECG signals and the corresponding saliency maps generated by the Base Model (1D convolutional model) and the Collocative Learning (2D convolutional models with and without CAGs), respectively. It is clear to see

Fig. 4. Examples of saliency maps generated by the Base Model (using the ECG signals as 1D inputs directly) and collocative learning (using the 2D-based collocative tensors as inputs). Maps by the collocative learning are generated at wave- and tensor-levels. Results of the collocative models with and without the periodic CAG regulations have been demonstrated respectively. It is easy to see that the collocative with CAGs has captured the periodic nature of the signals at either wave or tensor level.

TABLE III
PERFORMANCE COMPARISON BETWEEN THE COLLOCATIVE MODELS WITH AND WITHOUT PERIODIC CAG REGULATION. THE BEST RESULTS ARE IN BOLD FONT.

Model	Accuracy (%)	F1-score (%)	Recall (%)	Precision (%)	TPR (%)	TNR (%)
Base Model	71.85	71.22	71.09	71.52	77.71	64.47
w/o CAGs	73.57	72.56	72.32	73.76	83.26	61.39
Impr. (%)	2.39	1.88	1.73	3.13	7.14	-4.78
w/ CAGs	74.38	73.72	73.53	74.21	80.92	66.13
Impr. (%)	3.52	3.51	3.43	3.76	4.13	2.57

that the collocative learning with CAGs has captured the periodic characteristics of the original signals with superior performance over the Base Model and the one without CAGs.

At wave-level, the Base Model focuses on much fewer segments of non-eating signals than those of the eating signals. This is good for building discriminative inference models, which explains why the Base Model can still obtain an acceptable performance (with F1-score of 71.22%) in the classification task. However, the selection of segments does not appear to have a fixed pattern in terms of signal periods. It will limit its application in evidence backtracking for medical studies. By contrast, the attention patterns of the collocative learning (either with or without CAGs) are well established, in the sense that collocative learning without CAGs puts its focus more on the ST intervals of the signals, and the one with CAGs focuses more on the PR intervals for non-eating signals and the ST intervals for eating signals. More importantly, we can find that the attention of the collocative learning with CAGs has formed a periodic pattern much more obvious than that of the one without CAGs.

This becomes more intuitive when one compares at the 2D tensor saliency maps of the two models. The periodic and diagonal patterns of the model with CAGs are distributed more evenly and with more uniformed shapes in the maps than those without CAGs. This validates the effectiveness of the periodic and diagonal regulator of CAGs. Furthermore, the 2D saliency maps provide much more comparative features than the 1D

signals. We will see that this is useful for generating human comprehensible evidence in Section IV-F.

E. Performance Comparison with the State-of-the-art

We compare it with 4 groups of 11 state-of-the-art (SOTA) methods as follows.

- Traditional machine learning: Wave_SVM [41] and Comp_SVM [41] are selected because they are indeed more popular (than deep models) in ECG analysis and with recognized performance.
- Sophisticated CNN for images: ResNet18 [6], ResNet34 [6], ResNet50 [6], VGG16 [5], and DenseNet121 [40] are selected for their popularity.
- Recurrent models: LSTM [42] is selected to represent the deep models for sequential data and it has also been widely adopted for ECG processing [43]–[45].
- Ad-hoc deep models: iResNet [46] is a representative for 1D convolutional networks which is a modification of ResNet for processing the ECG data specifically; Spect_CNN [31] has employed 2D CNN for ECG processing in frequency domain (on spectrum); CNN+LSTM [47] is a deep model for ECG processing which has been designed to take the advantage of CNN for representation learning and that of LSTM for sequential data processing.

The results are shown in Table IV. Not surprisingly, our collocative learning framework (w/ CAGs) has outperformed all 11 SOTA methods with performance gain 1.27% ~ 7.83% (0.82% ~ 10.05%) in Accuracy (F1-score). Therefore, its performance is validated.

F. Generation of Human Comprehensible Evidence

We use the method introduced in Section III-C2 to build the attributes according to the rankings indicated by the saliency maps generated in Section III-C3. The decision trees/forests can then be built with the attributes. The results with the maximum tree height at 8 is shown in Figure 5. It is surprising that the performance with only 10 attributes used is already

TABLE IV
PERFORMANCE COMPARISON WITH THE STATE-OF-THE-ART.

Group	Model	Accuracy (%)	F1-score (%)	Recall (%)	Precision (%)	TPR (%)	TNR (%)
Traditional Machine Learning	Wave_SVM	68.98	66.99	67.05	69.62	83.96	50.14
	Comp_SVM	71.01	69.69	69.53	71.24	82.49	56.57
Sophisticated CNNs	ResNet18	71.85	71.22	71.09	71.52	77.71	64.47
	ResNet34	71.51	70.93	70.82	71.15	76.88	64.75
	ResNet50	71.61	70.73	70.55	71.43	79.79	61.32
	VGG16	72.29	71.63	71.49	71.99	78.50	64.47
	DenseNet121	72.56	72.03	71.92	72.22	77.51	66.33
Recurrent Models	LSTM	72.42	71.73	71.57	72.16	79.02	64.12
Ad-hoc Deep Models	iResNet	72.75	71.77	71.56	72.79	81.98	61.10
	Spect_CNN	73.45	73.12	73.15	73.10	75.80	70.49
	CNN+LSTM	72.13	71.19	70.99	72.06	80.91	61.07
Ours	w/o CAGs	73.57	72.56	72.32	73.76	83.26	61.39
	w/ CAGs	74.38	73.72	73.53	74.21	80.92	66.13

Fig. 5. Performance of the decision trees and forests (LightGBM) generated with the maximum tree height of 8.

Fig. 6. Examples of decision trees generated with the attributes learned by the proposed method. The tree with yellow nodes is generated with the wave-level (unary) attributes and the one with blue nodes is generated with the wave-level (comparative) attributes.

comparable to that of the deep models. Examples of the resulting trees are given in Figure 6. The ST sections, which are highly ranked attributes, are frequently refereed by the decision trees/forests. Given that ST is a sub-wave of QT, The tree/forest model's dependency on ST is consistent with

the findings in previous medical research, which concluded that the eating/digesting mainly affects the QT section of the ECG signals. Moreover, QT identified in previous medical researches is a coarse-scale reference as it covers 4 out of 6 cardiac waves of a period. By contrast, we can see from Fig. 6 that the proposed method has located the evidence in a much finer scale and more practical forms (i.e., with finer sub-waves as attributes and quantized branches). It is much intuitive for medical technicians/researchers to analyze the interactions between the ECG and the eating behavior than using only coarse-scale statistics.

V. CONCLUSION

We have presented a pilot study of AEM using 24-hour ECG. A collocative learning framework is proposed for this problem, which has addressed the issue of adapting the sophisticated 2D image-based deep models to 1D ECG signal analysis. More importantly, the framework is capable of modeling the human logic of cardiac analysis based on comparative features, and therefore, the decision trees/forests can be generated as human comprehensive representations of the inference evidence. The effectiveness of the framework is validated on the largest ECG dataset for eating behavior, where the superior performance over SOTA has been observed. The capability of cardiac evidence mining has also been verified through the consistency of the evidence backtracked by the framework and that of the previous medical studies.

REFERENCES

- [1] H. Jiang, J. Starkman, M. Liu, and M.-C. Huang, "Food nutrition visualization on google glass: Design tradeoff and field evaluation," *IEEE Consumer Electronics Magazine*, vol. 7, no. 3, pp. 21–31, 2018.
- [2] K. Hnatkova, D. Kowalski, J. J. Keirns, E. M. van Gelderen, and M. Malik, "Qtic changes after meal intake: sex differences and correlates," *Journal of Electrocardiology*, vol. 47, no. 6, pp. 856–862, 2014.
- [3] E. M. Scott, J. P. Greenwood, G. Vacca, J. B. Stoker, S. G. Gilbey, and D. A. Mary, "Carbohydrate ingestion, with transient endogenous insulinaemia, produces both sympathetic activation and vasodilatation in normal humans," *Clinical Science*, vol. 102, no. 5, pp. 523–529, 2002.
- [4] J. Taubel, A. H. Wong, A. Naseem, G. Ferber, and A. J. Camm, "Shortening of the qt interval after food can be used to demonstrate assay sensitivity in thorough qt studies," *The Journal of Clinical Pharmacology*, vol. 52, no. 10, pp. 1558–1565, 2012.
- [5] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.
- [6] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 770–778.

- [7] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 2818–2826.
- [8] X.-Y. Wei, Z.-Q. Yang, X.-L. Zhang, G. Liao, A.-L. Sheng, S. K. Zhou, Y. Wu, and L. Du, "Deep collocation learning for immunofixation electrophoresis image analysis," *IEEE Transactions on Medical Imaging*, 2021.
- [9] P. P. Basiotis, S. O. Welsh, F. J. Cronin, J. L. Kelsay, and W. Mertz, "Number of days of food intake records required to estimate individual and group nutrient intakes with defined confidence," *The Journal of Nutrition*, vol. 117, no. 9, pp. 1638–1641, 1987.
- [10] O. Amft, "A wearable earpad sensor for chewing monitoring," in *Sensors*. IEEE, 2010, pp. 222–227.
- [11] Y. Dong, A. Hoover, J. Scisco, and E. Muth, "A new method for measuring meal intake in humans via automated wrist motion tracking," *Applied Psychophysiology and Biofeedback*, vol. 37, no. 3, pp. 205–215, 2012.
- [12] M. Farooq, J. M. Fontana, and E. Sazonov, "A novel approach for food intake detection using electroglottography," *Physiological Measurement*, vol. 35, no. 5, p. 739, 2014.
- [13] M. Farooq and E. Sazonov, "Comparative testing of piezoelectric and printed strain sensors in characterization of chewing," in *2015 37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*. IEEE, 2015, pp. 7538–7541.
- [14] J. Liu, E. Johns, L. Atallah, C. Pettitt, B. Lo, G. Frost, and G.-Z. Yang, "An intelligent food-intake monitoring system using wearable sensors," in *2012 Ninth International Conference on Wearable and Implantable Body Sensor Networks*. IEEE, 2012, pp. 154–160.
- [15] S. Päßler, M. Wolff, and W.-J. Fischer, "Food intake monitoring: an acoustical approach to automated food intake activity detection and classification of consumed food," *Physiological Measurement*, vol. 33, no. 6, p. 1073, 2012.
- [16] E. Sazonov, S. Schuckers, P. Lopez-Meyer, O. Makeyev, N. Sazonova, E. L. Melanson, and M. Neuman, "Non-invasive monitoring of chewing and swallowing for objective quantification of ingestive behavior," *Physiological Measurement*, vol. 29, no. 5, p. 525, 2008.
- [17] M. Sun, L. E. Burke, Z.-H. Mao, Y. Chen, H.-C. Chen, Y. Bai, Y. Li, C. Li, and W. Jia, "ebutton: a wearable computer for health monitoring and personal assistance," in *Proceedings of the 51st Annual Design Automation Conference*, 2014, pp. 1–6.
- [18] T. Vu, F. Lin, N. Alshurafa, and W. Xu, "Wearable food intake monitoring technologies: A comprehensive review," *Computers*, vol. 6, no. 1, p. 4, 2017.
- [19] M. A. Serhani, H. T. El Kassabi, H. Ismail, and A. Nujum Navaz, "Ecg monitoring systems: Review, architecture, processes, and key challenges," *Sensors*, vol. 20, no. 6, p. 1796, 2020.
- [20] M. Farooq, J. M. Fontana, A. F. Boateng, M. A. McCrory, and E. Sazonov, "A comparative study of food intake detection using artificial neural network and support vector machine," in *2013 12th International Conference on Machine Learning and Applications*, vol. 1. IEEE, 2013, pp. 153–153.
- [21] Y. Bi, M. Lv, C. Song, W. Xu, N. Guan, and W. Yi, "Autodietary: A wearable acoustic sensor system for food intake recognition in daily life," *IEEE Sensors Journal*, vol. 16, no. 3, pp. 806–816, 2015.
- [22] P. Lopez-Meyer, S. Schuckers, O. Makeyev, and E. Sazonov, "Detection of periods of food intake using support vector machines," in *2010 Annual International Conference of the IEEE Engineering in Medicine and Biology*. IEEE, 2010, pp. 1004–1007.
- [23] J. M. Fontana, M. Farooq, and E. Sazonov, "Estimation of feature importance for food intake detection based on random forests classification," in *2013 35th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*. IEEE, 2013, pp. 6756–6759.
- [24] M. M. Al Rahhal, Y. Bazi, M. Al Zuair, E. Othman, and B. BenJdira, "Convolutional neural networks for electrocardiogram classification," *Journal of Medical and Biological Engineering*, vol. 38, no. 6, pp. 1014–1025, 2018.
- [25] F. Andreotti, O. Carr, M. A. Pimentel, A. Mahdi, and M. De Vos, "Comparing feature-based classifiers and convolutional neural networks to detect arrhythmia from short segments of ecg," in *2017 Computing in Cardiology (CinC)*. IEEE, 2017, pp. 1–4.
- [26] X. Fan, Q. Yao, Y. Cai, F. Miao, F. Sun, and Y. Li, "Multiscaled fusion of deep convolutional neural networks for screening atrial fibrillation from single lead short ecg recordings," *IEEE Journal of Biomedical and Health Informatics*, vol. 22, no. 6, pp. 1744–1753, 2018.
- [27] A. Y. Hannun, P. Rajpurkar, M. Haghpasahi, G. H. Tison, C. Bourn, M. P. Turakhia, and A. Y. Ng, "Cardiologist-level arrhythmia detection and classification in ambulatory electrocardiograms using a deep neural network," *Nature Medicine*, vol. 25, no. 1, pp. 65–69, 2019.
- [28] R. He, K. Wang, N. Zhao, Y. Liu, Y. Yuan, Q. Li, and H. Zhang, "Automatic detection of atrial fibrillation based on continuous wavelet transform and 2d convolutional neural networks," *Frontiers in Physiology*, vol. 9, p. 1206, 2018.
- [29] M. Naz, J. H. Shah, M. A. Khan, M. Sharif, M. Raza, and R. Damaševičius, "From ecg signals to images: a transformation based approach for deep learning," *PeerJ Computer Science*, vol. 7, p. e386, 2021.
- [30] P. Sarkar and A. Etemad, "Self-supervised learning for ecg-based emotion recognition," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 3217–3221.
- [31] A. Ullah, S. M. Anwar, M. Bilal, and R. M. Mehmood, "Classification of arrhythmia by using deep learning with 2-d ecg spectral image representation," *Remote Sensing*, vol. 12, no. 10, p. 1685, 2020.
- [32] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 7132–7141.
- [33] X. Wang, R. Girshick, A. Gupta, and K. He, "Non-local neural networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 7794–7803.
- [34] S. Woo, J. Park, J.-Y. Lee, and I. So Kweon, "Cbam: Convolutional block attention module," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 3–19.
- [35] A. Chattopadhyay, A. Sarkar, P. Howlader, and V. N. Balasubramanian, "Grad-cam++: Generalized gradient-based visual explanations for deep convolutional networks," in *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*. IEEE, 2018, pp. 839–847.
- [36] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-cam: Visual explanations from deep networks via gradient-based localization," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 618–626.
- [37] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba, "Learning deep features for discriminative localization," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 2921–2929.
- [38] J. Laitala, M. Jiang, E. Syrjälä, E. K. Naeini, A. Airola, A. M. Rahmani, N. D. Dutt, and P. Liljeberg, "Robust ecg r-peak detection using lstm," in *Proceedings of the 35th Annual ACM Symposium on Applied Computing*, 2020, pp. 1104–1111.
- [39] S. Vijayarangan, R. Vignesh, B. Murugesan, S. Preejith, J. Joseph, and M. Sivaprakasam, "Rpnet: A deep learning approach for robust r peak detection in noisy ecg," in *2020 42nd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*. IEEE, 2020, pp. 345–348.
- [40] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 4700–4708.
- [41] C. Cortes and V. Vapnik, "Support-vector networks," *Machine Learning*, vol. 20, no. 3, pp. 273–297, 1995.
- [42] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, 1997.
- [43] B. Hou, J. Yang, P. Wang, and R. Yan, "Lstm-based auto-encoder model for ecg arrhythmias classification," *IEEE Transactions on Instrumentation and Measurement*, vol. 69, no. 4, pp. 1232–1240, 2019.
- [44] S. Saadatnejad, M. Oveis, and M. Hashemi, "Lstm-based ecg classification for continuous monitoring on personal wearable devices," *IEEE Journal of Biomedical and Health Informatics*, vol. 24, no. 2, pp. 515–523, 2019.
- [45] Ö. Yildirim, "A novel wavelet sequence based on deep bidirectional lstm network model for ecg signal classification," *Computers in Biology and Medicine*, vol. 96, pp. 189–202, 2018.
- [46] S. Yang, H. Xiang, Q. Kong, and C. Wang, "Multi-label classification of electrocardiogram with modified residual networks," in *2020 Computing in Cardiology*. IEEE, 2020, pp. 1–4.
- [47] C. Chen, Z. Hua, R. Zhang, G. Liu, and W. Wen, "Automated arrhythmia classification based on a combination network of cnn and lstm," *Biomedical Signal Processing and Control*, vol. 57, p. 101819, 2020.