

# CAFO: Feature-Centric Explanation on Time Series Classification

JaeHo Kim  
 kjh3690@unist.ac.kr  
 Ulsan National Institute of Science and  
 Technology (UNIST)  
 Artificial Intelligence Graduate School (AIGS)  
 Ulsan, South Korea

Junghye Lee  
 junghye@snu.ac.kr  
 Seoul National University (SNU)  
 Technology Management, Economics and  
 Policy Program &  
 Graduate School of Engineering Practice &  
 Institute of Engineering Research  
 Seoul, South Korea

Seok-Ju Hahn  
 seokjuhahn@unist.ac.kr  
 Ulsan National Institute of Science and  
 Technology (UNIST)  
 Department of Industrial Engineering (IE)  
 Ulsan, South Korea

Yoontae Hwang  
 yoontae@unist.ac.kr  
 Ulsan National Institute of Science and  
 Technology (UNIST)  
 Department of Industrial Engineering (IE)  
 Ulsan, South Korea

Seulki Lee<sup>†</sup>  
 seulki.lee@unist.ac.kr  
 Ulsan National Institute of Science and  
 Technology (UNIST)  
 Computer Science and Engineering (CSE) &  
 Artificial Intelligence Graduate School (AIGS)  
 Ulsan, South Korea

## ABSTRACT

In multivariate time series (MTS) classification, finding the important features (e.g., sensors) for model performance is crucial yet challenging due to the complex, high-dimensional nature of MTS data, intricate temporal dynamics, and the necessity for domain-specific interpretations. Current explanation methods for MTS mostly focus on time-centric explanations, apt for pinpointing important time periods but less effective in identifying key features. This limitation underscores the pressing need for a feature-centric approach, a vital yet often overlooked perspective that complements time-centric analysis. To bridge this gap, our study introduces a novel feature-centric explanation and evaluation framework for MTS, named **CAFO** (Channel Attention and Feature Orthogonalization). CAFO employs a convolution-based approach with channel attention mechanisms, incorporating a depth-wise separable channel attention module (DepCA) and a QR decomposition-based loss for promoting feature-wise orthogonality. We demonstrate that this orthogonalization enhances the separability of attention distributions, thereby refining and stabilizing the ranking of feature importance. This improvement in feature-wise ranking enhances our understanding of feature explainability in MTS. Furthermore, we develop metrics to evaluate global and class-specific feature importance. Our framework's efficacy is validated through extensive empirical analyses on two major public benchmarks and real-world datasets, both synthetic and self-collected, specifically designed to highlight class-wise discriminative features. The results confirm CAFO's robustness and informative capacity in assessing feature importance in MTS classification tasks. This study not only advances the understanding of feature-centric explanations in MTS but also sets a foundation for future explorations in feature-centric explanations. The codes are available at <https://github.com/eai-lab/CAFO>.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for third-party components of this work must be honored. For all other uses, contact the owner/author(s).

KDD '24, August 25–29, 2024, Barcelona, Spain

© 2024 Copyright held by the owner/author(s).

ACM ISBN 979-8-4007-0490-1/24/08

<https://doi.org/10.1145/3637528.3671724>

## CCS CONCEPTS

- Computing methodologies → Learning to rank; Temporal reasoning.

## KEYWORDS

Time Series Classification, Feature-Centric Explanation, Explainable AI

## 1 INTRODUCTION

With the advancement of Internet of Things (IoT) technologies, time series classification (TSC) tasks have proliferated in recent years. A notable characteristic of TSC data derived from these sources is that they are usually 1) multivariate; that is, they contain multiple measurements or sensors and 2) characterized by patterns that are complex and intertwined, which poses challenges for semantic interpretation [3, 34] by humans, a stark contrast to more intuitively graspable domains of image and text data. These multivariate time series (MTS) data have found practical applications ranging from the classification of human activities to the detection of industrial faults [10, 22], demonstrating its broad applicability.

As MTS data find broader applications, an essential need emerges in the phase of model development. Engineers and domain experts seek not just to use these models but to understand how they process data. This understanding is vital; it can drastically reduce computational and manufacturing costs and foster confidence in the model's deployment, ensuring it leverages features recognized as important [47]. In this context, the role of explainable AI (XAI) is not just beneficial but indispensable. Yet, a concerning observation arises: XAI research in MTS has predominantly concentrated on generating time-step-specific or instance-specific explanations [7, 19, 32, 40, 41], focusing narrowly on segments of time critical to the model's decision-making in a given instance. Such local explanations, while invaluable in contexts like healthcare, reveal a significant gap for a more comprehensive, feature-centric overview that can provide a broader understanding of the data. In MTS, a ‘feature’ is commonly identified as a separate channel or measurement variable, which is independent of the time axis. This

<sup>†</sup> S. Lee is the corresponding author

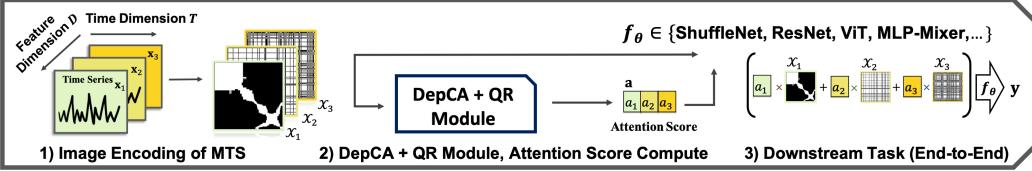
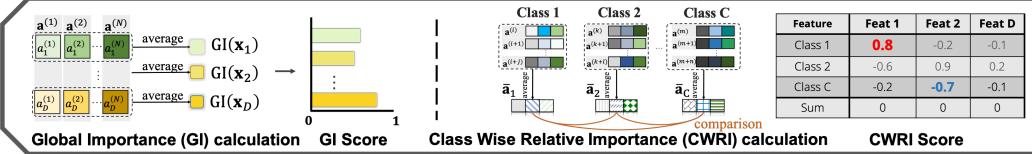
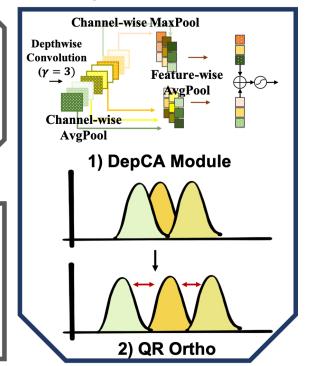
**A. End-to-End Training****C. Feature Importance Calculation****B. DepCA + QR Module**

Figure 1: Overview of CAFO: (A) End-to-end training. Raw time series are converted into images using image encoding methods, followed by the extraction of channel-wise attention scores using the DepCA+QR Module. These attention scores are element-wise multiplied to image features for end-to-end model training. (B) DepCA assesses feature contributions, while QR-Ortho Loss minimizes feature redundancy through orthogonality regularization. (C) Feature Importance Calculation. The calculated attention scores are utilized to explain MTS data via Global Importance (GI) and Class-Wise Relative Importance (CWRI) metrics.

need is particularly acute in both industry and academia, where a global understanding of TSC tasks, at both the class level and across the model as a whole, is crucial. However, previous MTS XAI works [2, 17] have addressed this topic in a limited manner, leaving room for more extensive exploration and discussion.

**Herein lies the main theme of our paper: we address the imperative need for a feature-centric explanation** - a perspective that is not just complementary but essential to the time-centric explanation in the MTS classification task.

To illustrate the practical impact of our approach, consider the example of a smart shoe manufacturer employing a deep learning model for classifying user activities based on sensor data (accelerometers, force sensors, etc.). Unlike previous methods that used fourteen sensors (or features) for 95% accuracy [22], our feature-centric analysis achieves a near-comparable 94% accuracy with just three key sensors (See Fig. 4 where we sequentially dropped important and unimportant features and measured the model performance). This insightful identification of key sensors would not have been easily attainable through previous time-step or local instance-based explanation methods. Conversely, relying on the three least important sensors, as identified by our model, drops accuracy to 71%, highlighting the critical nature of sensor selection. Moreover, this revelation goes beyond mere accuracy; it offers manufacturers and engineers invaluable insights. Manufacturers can reduce costs by focusing on essential sensors, potentially eliminating redundant ones. Engineers gain insights for optimizing model performance. These scenarios, spanning industrial and academic fields, introduce the broader application of feature-centric explanation in MTS classification tasks. We present detailed use cases motivating feature-centric explanation throughout this paper and in Appendix A.

While a number of studies have explored time-step based explanations in depth, feature-based explanations have typically been addressed as a secondary focus or in a limited scope [2, 17]. To our knowledge, there has been a lack of discussion regarding feature-centric explanations in TSC in deep learning, especially with regard

to multivariate data (i.e., MTS)<sup>1</sup>, presented as a comprehensive research paper. The lack of explanation strategy, appropriate MTS datasets, characterized by clearly defined feature importance, alongside matching evaluation protocols, presents a considerable challenge [32] in the realm of MTS XAI. Our paper bridges this gap by presenting the following contributions:

- (1) **Methodologies.** Introduction of Channel Attention and Feature Orthogonalization (CAFO): (1) DepCA, a novel convolution-based framework that utilizes channel attention, which measures feature importance and (2) QR-Ortho, a QR decomposition-based regularizer that ensures feature separability for improved feature-centric explanation.
- (2) **Datasets.** Compilation of both synthetic and real world datasets collected with known class-discriminative features.
- (3) **Metrics.** Development of a comprehensive set of metrics aimed at quantifying global and class-specific feature importance, complete with a corresponding evaluation protocol.
- (4) **Experiments.** Extensive number of empirical evaluations confirming the practical value of the proposed work.

## 2 PRELIMINARIES AND RELATED WORKS

We first formalize the notation used in our work. To better understand, it is beneficial to have a big overview of CAFO depicted in Fig. 1. Subsequently, the rest of the section illustrates prior works that are helpful in understanding our method.

### 2.1 Preliminaries

Given  $N$  number of MTS samples, the  $i$ -th MTS instance  $\mathbf{X}^{(i)} = [\mathbf{x}_1^{(i)}, \dots, \mathbf{x}_D^{(i)}] \in \mathbb{R}^{T \times D}$  encompasses  $T$  times steps and  $D$  features. In this context, a univariate time series is defined as a single  $j$ -th feature sequence  $\mathbf{x}_j^{(i)} = [x_1, \dots, x_T]^T$ , and an aggregation of such univariate sequences constitutes an MTS. Traditional time-step-based explanations offer insights along the temporal ( $T$ ) dimension, but our research pivots towards elucidating the feature ( $D$ ) dimension.

<sup>1</sup>Our focus is on multivariate time series as we provide feature-centric explanations.

Consequently, our primary interest lies in discerning the significance of each feature, establishing a hierarchy of feature importance that is both global and class-specific. Specific to our problem setting, rather than using the raw MTS as-is, we transform each feature  $\mathbf{x}_j$  into a single-channel image of size  $\{0, 1\}^{T \times T}$  using image encoding methods. As a whole, the raw MTS input  $\mathbf{X}^{(i)} \in \mathbb{R}^{T \times D}$  is converted into an image of size  $\mathcal{X}^{(i)} \in \mathbb{R}^{D \times T \times T}$ . Employing a channel attention (CA) module, we compute a set of attention scores along the  $D$  dimension  $\text{CA}(\mathcal{X}^{(i)}) = \mathbf{a}^{(i)} = [a_1, \dots, a_D]^\top$  within the range  $[0, 1]^D$ , where each  $a_j$  represents the attention allocated to the  $j$ -th feature channel  $X_j$  (equivalent to  $\mathbf{x}_j$ ).

## 2.2 Image Encoding of MTS

Image encoding of MTS involves transforming time series data into image formats, such as a Recurrence Plot (RP) [9] or Gramian Angular Fields (GAF) [44]. Encoding time series into images offers several benefits in analyzing feature-wise importance. First, image encoding operates independently of specific standardization methods [9], which is often crucial due to the heterogeneity of features, e.g., the varying scales of accelerometers and gyroscopes. This can even significantly impact the end performance of a modeling [52]. Image encoding, however, primarily relies on point-wise relations (e.g., inner products with threshold) to represent time series, thereby liberating the feature representation from the implicit biases of any particular scaling method. This aspect ensures a more equitable comparison and ranking of features. Second, image encoding enhances the representation of temporal dependency within features. By converting the original feature  $\mathbf{x}_j \in \mathbb{R}^T$  into a  $\mathcal{X}_j \in \mathbb{R}^{T \times T}$  image, recurrent patterns become more explicit and discernible [9], which allows for the use of well-curated vision models, e.g., ViT [8]. Our empirical results (Appendix C) suggest these models more effectively discern feature importance in our study, potentially owing to image encoding or inherent model capabilities. We use Recurrence Plot to capture the recurrence patterns in MTS, and explore alternative encoding techniques such as GAF in Appendix B.

## 2.3 Channel Attention (CA) Modules

CA is primarily used in the image classification domain to improve model performance by emphasizing relevant feature channels. The pioneering SENet [18], BAM [30], CBAM [46], and SIMAM [48] harness CA by collating channel specific statistics (e.g. global average), passing them through parametrized functions to obtain channel or spatial attention. In contrast to the use of CA in the image domain, which integrates CA at multiple points within the latent space, **our approach applies CA singularly and directly to the input representation, to obtain the attention scores for each feature**. We note that in the time series domain, the joint usage of image encoding and CA has been previously explored in temporal [24], frequency [21], and wavelet [12]-based literature, often to augment model performance, and occasionally to offer interpretative insights via raw attention visualizations. Our research pioneers the use of CA scores to systematically evaluate feature importance on a global and class-specific scale in MTS data.

## 2.4 Multivariate Time Series Explanation

**Post-hoc explanation in multivariate time series (MTS)** elucidate model decisions by deriving explanations from their output, making them generally agnostic to the underlying model. In MTS explanation, several post-hoc methods employed in the image domain have been repurposed for MTS by viewing the raw time series as a  $T \times D$  image. A recent study by Turbé et al. [41] has undertaken a comprehensive assessment of various post-hoc interpretability methods—Integrated Gradients [36], GradientSHAP [26], and Shapley value sampling [6]—in the context of TSC, highlighting substantial discrepancies in time-centric explanation across methods, also noted by Schlegel et al [32]. Our research extends these observations, confirming these inconsistencies in feature-centric explanations of MTS, and first identifying the impact of train/validation distribution on feature importance inconsistency. The study [41] also highlights the limitations on the use of synthetic data in prior time-centric explanation research [19], emphasizing the need for real-world datasets with clear discriminative features for validating MTS explanation methods. In the pursuit of enhancing these methods, past works have applied these post-hoc methods on LSTM [15], TCN [23], and Transformer[43] models for time-based explanations. Orthogonal to these approaches, DynaMask [7] is a post-hoc method, providing an explanation based on optimizing perturbation masks for MTS. However, its requirement for numerous optimization steps per instance presents a challenge, limiting its efficiency in global and class-specific importance calculation.

**Model-based explanation for time series** rely on specific neural architecture such as recurrent neural networks (RNNs), as these models inherently handle sequential inputs. Nevertheless, recent works show that they suffer from saliency vanishing [14] and may have limitations in explaining time series data [20]. For example, TimeSHAP [2] is a recurrent explainer extending KernelSHAP [26] to the temporal domain by grouping sequential data into coalitions. Shapley-based methods are known to be computationally-intensive [26], while TimeSHAP provides efficient pruning methods to overcome this. However, pruning relies on the assumption that recent events have a predominant influence on model outcome might not apply universally, such as in continuous event recording like human activity monitoring. Another recurrent-based approach, FIT [40], assigns significance to events using counterfactuals within a generative model. Unfortunately, training a generator adds an extra cost, and the explanation depends on the generator's performance. LAXCAT [17] is another model-based explanation method utilizing both temporal and variable attention scores using 1D convolution methods. Our CAFO method, however, distinguishes itself by employing 2D convolutions and channel attention (CA), offering a unique structural approach to derive attention scores.

## 3 CAFO: CHANNEL ATTENTION AND FEATURE ORTHOGONALIZATION

Fig. 1 provides an overview of CAFO for extracting feature-centric importances from MTS data. Starting with image encoding (specifically recurrence plot (RP) [9]), the raw MTS is transformed into image-like data, where our DepCA module (Sec. 3.1) is used to compute the channel attention score  $\mathbf{a} \in \mathbb{R}^D$ . These scores are then element-wise multiplied with their respective channels, which are

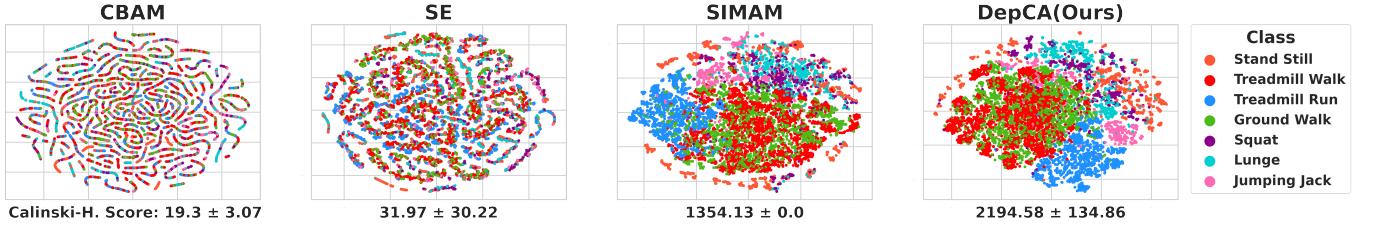


Figure 2: Visualization of the channel attention (CA) values using t-SNE [42] for CBAM [46], SE [18], SIMAM [48], and the proposed DepCA module on the GILON dataset [22]. The Calinski-Harabasz score [5] at the bottom indicates their clustering performance (the higher, the better). As observed, DepCA effectively captures sample and class-specific information even though the CA scores are computed in the early layer of the network, in contrast to existing methods [18, 46] that compute the CA scores in latent channel spaces (middle layers of the network)

further processed for **end-to-end training with downstream backbone models** such as a ResNet [13]. Throughout the training, we employ a novel QR-Ortho loss (Sec. 3.2) to ensure the orthogonality along the feature dimension of the attention, effectively reducing feature redundancy through orthogonalization. Upon completion of the training, we harness these attention scores to compute the Global Importance (GI) and Class-Wise Relative Importance (CWRI) metrics, with further details regarding the metric described in Sec. 4.

### 3.1 Depthwise Channel Attention (DepCA)

In our work, a raw time series  $X \in \mathbb{R}^{T \times D}$  is encoded into an image-like format  $X \in \mathbb{R}^{D \times T \times T}$ , with each channel representing a distinct feature. The CA (Channel Attention) module evaluates these channels, employing attention scores  $\mathbf{a} = [a_1, \dots, a_D]^T$  for global and class-specific metric computation. It's essential that attention scores,  $\mathbf{a}$ , precisely captures each feature's unique information. To achieve this, we introduce the **Depthwise Channel Attention (DepCA)** module, which surpasses traditional CA techniques in extracting comprehensive information from each channel. Where previous CA methods [46] rely on simple statistics like global maxima—apt for latent channel spaces for efficiency—our methodology, with its depthwise convolution, allows the model to learn informative statistics from each feature representation. By applying depthwise convolutions, we treat each feature independently, capturing distinct details without inter-feature interference. It efficiently extracts clear, differentiated information, as shown by our t-SNE visualizations in Fig. 2.

The DepCA module begins by applying a set of depthwise convolutional filters to the input  $X$ . We use  $\gamma$  number of filters per each feature channel (with  $\gamma = 3$  in all experiments). This yields the feature descriptor as  $\text{Conv}_\gamma(X) = F_{\text{out}} \in \mathbb{R}^{(\gamma \times D) \times H \times W}$ , where  $D$  is the number of channels,  $H$  and  $W$  denotes the height and width of the output channels, respectively. Following this, DepCA performs two pooling operations on  $F_{\text{out}}$ : an average and max pooling, executed channel-wise (CW) [46] to produce  $F_{\text{avg}}, F_{\text{max}} \in \mathbb{R}^{\gamma \times D}$ . These pooled features are then averaged across the channels coming from the same original feature (i.e., feature-wise; FW). As we aim to provide a feature-centric explanation, we performed FW pooling to get an attention score for each feature. Finally, the two output features are combined through element-wise summation to obtain the aggregated feature representation and then passed to the sigmoid function  $\sigma(x) = 1/(1 + \exp(-x))$ , to constrain the CA score between zero and one. Putting it all together, the CA score of the image  $X$ , denoted as  $\text{CA}(X) = [a_1, \dots, a_D]^T \in [0, 1]^D$  where

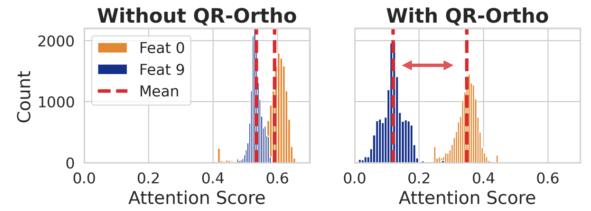


Figure 3: Orthogonal regularization on the feature-dimension of the attentions enhances separability. Using QR-Ortho loss, we demonstrate an enhanced distinction between previously overlapping attentions in the Gilon dataset [22], consistent across five-fold CV.

$a_j$  is the attention score of  $X_j$ , is computed as:

$$\mathbf{F}_{\text{out}} = \text{Conv}_\gamma(X)$$

$$\mathbf{F}_{\text{avg}} = \text{CWAvgPool}(\mathbf{F}_{\text{out}}), \quad \mathbf{F}_{\text{max}} = \text{CWMaxPool}(\mathbf{F}_{\text{out}}) \quad (1)$$

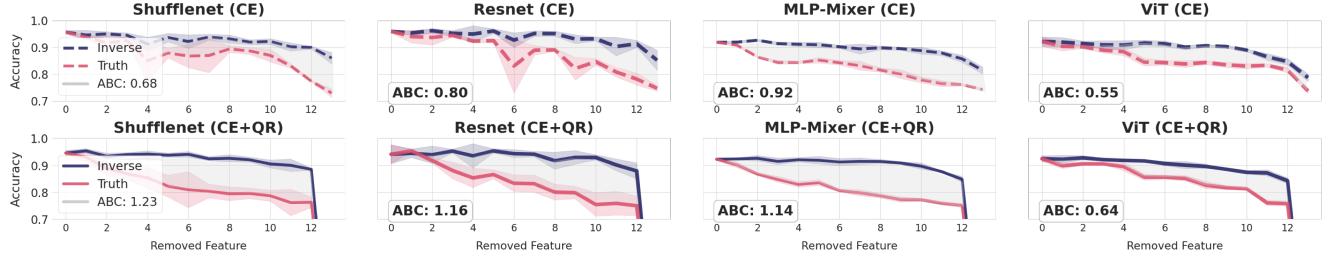
$$\text{CA}(X) \triangleq \sigma(\text{FWAvgPool}(\mathbf{F}_{\text{avg}}) + \text{FWAvgPool}(\mathbf{F}_{\text{max}})) \equiv \mathbf{a}$$

The CA score,  $\mathbf{a}$ , is element-wise multiplied with the image representation  $X$ , expressed as  $\mathbf{a} \odot X$ . Here, **attention values determine the retention or suppression of features**; values near 1 retain features, while those close to 0 suppress them. Consequently, these attention scores are crucial in constructing feature importance metrics and determining the relevance of different features.

### 3.2 Enhancing Feature Separability: QR-Ortho

During the development of DepCA, we observed an overlapping distribution of CA (channel attention) scores between each feature (see Fig. 3). This overlap complicates the derivation of precise feature importance rankings, which are essential for computing both global and class-specific metrics. To address this issue, we enforce an orthogonal regularization on the feature average of the CA scores to obtain distinguished CA distribution, leading to enhanced and distinct feature rankings of MTS data. Fig. 3 illustrates the clear separation of previously overlapping CA distributions when orthogonality is enforced. This enhanced separation contributes to three key outcomes: (1) distinct CA distributions for each feature, (2) improved ranking measures for global and class-wise feature importance, and as a result (3) overall better explainability of the MTS data. In our evaluation, we observe a substantial improvement in the explainability of MTS data with feature-wise orthogonality, empirically verifying its critical role in TSC model explanations.

To this end, we propose **QR-Ortho Loss** that enforces feature-wise orthogonality along the feature dimension of the CA scores through QR decomposition [11] that factorizes a given matrix  $\tilde{\mathbf{A}}$  into



**Figure 4: RemOve And Retrain (ROAR) with Gilon [22].** The feature ranks of the Gilon task were first identified by our CAFO using the whole 14 feature set, with potential rank variations across figures. To assess the importance of each feature, we systematically removed them from the train and test datasets, ensuring consistency in distribution. This process involved the progressive subtraction of more important (red as ‘Truth’) and less important (blue as ‘Inverse’) features. After each removal, the model was retrained, and its accuracy was evaluated. The X-axis represents the number of features removed (with zero indicating no removal), while the Y-axis shows the model’s accuracy. A notable decline in accuracy is observed with the removal of key features, in contrast to a minimal impact when less important features are omitted. The area between the curve (ABC) metric quantifies the gap between the two curves, where a higher ABC indicates superior feature-wise ranking. The first row exhibits the model’s performance using cross-entropy (CE) alone, while the second row shows integration of QR-Ortho (our approach) with CE. A marked improvement in ABC scores across all models is evident, underscoring QR-Ortho’s efficacy in identifying pivotal features.

an orthonormal matrix  $\mathbf{Q}$  and a residual upper triangular matrix  $\mathbf{R}$ , i.e.,  $\bar{\mathbf{A}} = \mathbf{QR}$ . Here, the class prototype matrix  $\bar{\mathbf{A}}$  is constructed by stacking  $\bar{\mathbf{a}}_c$  in a row-wise manner, where row-wise means arranging the class-prototype vectors into a matrix format where each class prototype occupies a single row. Given  $C$  classes in a TSC task, MTS data can be represented as  $\mathbf{K} = \{(X^{(1)}, y^{(1)}), \dots, (X^{(N)}, y^{(N)})\}$ , where  $y^{(i)} \in \{1, \dots, C\}$  is the corresponding class label. Then, the class prototype of class  $c$  denoted as  $\bar{\mathbf{a}}_c$ , is defined as the average CA scores of samples belonging to class  $c$ , given by:

$$\bar{\mathbf{a}}_c = \frac{1}{|\mathbf{K}_c|} \sum_{i \in \mathbf{K}_c} \mathbf{a}^{(i)} \quad (2)$$

where  $\mathbf{K}_c = \{(\mathbf{X}, y) \mid y \in c\}$  denotes the MTS instances in class  $c$ .

As such, we denote  $\bar{\mathbf{A}}_{:,j}$  ( $j=1, \dots, D$ ) as the column (feature) vector of  $\bar{\mathbf{A}}$  used to perform QR decomposition as  $\bar{\mathbf{A}} = \mathbf{QR}$ , given by:

$$\underbrace{\begin{bmatrix} | & | & & | \\ q_1 & q_2 & \cdots & q_D \\ | & | & & | \end{bmatrix}}_{\mathbf{Q}} \underbrace{\begin{bmatrix} \langle q_1, \bar{\mathbf{A}}_{:,1} \rangle & \langle q_1, \bar{\mathbf{A}}_{:,2} \rangle & \cdots & \langle q_1, \bar{\mathbf{A}}_{:,D} \rangle \\ 0 & \langle q_2, \bar{\mathbf{A}}_{:,2} \rangle & \cdots & \langle q_2, \bar{\mathbf{A}}_{:,D} \rangle \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \langle q_D, \bar{\mathbf{A}}_{:,D} \rangle \end{bmatrix}}_{\mathbf{R}} \quad (3)$$

The  $\mathbf{Q}$  matrix embodies the orthogonal basis of the feature dimension of  $\bar{\mathbf{A}}$ , and the upper diagonal elements of the  $\mathbf{R}$  matrix, i.e.,  $R_{ij}$  ( $i \leq j$ ) =  $\langle q_i, \bar{\mathbf{A}}_{:,j} \rangle$ , signify the dot products between class feature representation  $\bar{\mathbf{A}}_{:,j}$  and the orthonormal basis  $q_i$ . The decomposition process ensures direct orthogonality, as the orthonormal columns of  $\mathbf{Q}$  inherently exhibit orthogonal properties. Also, by leveraging the widely-used Gram-Schmidt [4] or Householder [33] algorithms, it maintains numerical stability while guaranteeing a unique set of orthogonal vectors.

Thus, by penalizing the upper off-diagonals of  $\mathbf{R}$ , i.e.,  $R_{ij}$  ( $i < j$ ), feature-wise orthogonality of the channel attentions can be effectively regularized. From this, we define QR-Ortho loss as:

$$\mathcal{L}_{QR} = \sum_{i < j} |\mathbf{R}_{ij}| \quad (4)$$

which is to be minimized in addition to the cross-entropy (CE) loss  $\mathcal{L}_{CE}$  with the hyperparameter  $\lambda$  that controls the strength of

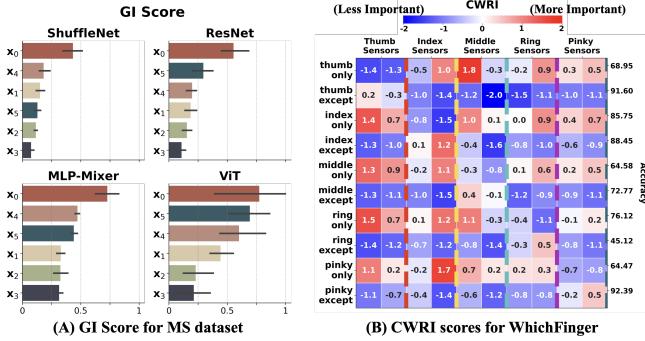
orthogonality as  $\mathcal{L} = \mathcal{L}_{CE} + \lambda \mathcal{L}_{QR}$ . The loss jointly optimizes the DepCA module and the downstream model end-to-end. In practice, the class prototype matrix  $\bar{\mathbf{A}}$  is formed in a mini-batch, with QR-Ortho Loss details in Appendix I.

## 4 FEATURE EXPLANATION MEASURES

We present two feature importance measures: (1) **Global Importance (GI)** and (2) **Class-Wise Relative Importance (CWRI)**, which provide reliable feature-wise explanations of MTS data (Fig. 5). While we elucidate GI and CWRI in terms of the attention scores, the computation of both measures does not favor or rely on attention scores. Rather, both measures can be effectively applied in conjunction with any attribution method capable of producing instance-wise attributions, as in prior studies [7, 40]. This is achieved by averaging the scores across the time dimension, aligning with Eq. (5). This comprehensiveness ensures that our measures are broadly applicable across various attribution methodologies. **Global Importance (GI).** The GI score quantifies the significance of each feature in relation to classification performance over the entire data and thus simplifies the interpretation and comparison between features. A feature with a high GI score is globally essential for accurate classifications, while a low GI score suggests negligible influence on the overall model performance. We denote the GI score of the  $j$ -th feature  $\mathbf{x}_j$  as  $GI(\mathbf{x}_j)$ , and it is calculated by averaging the  $j$ -th channel attention (CA) scores  $a_j \in [0, 1]$  of all data samples over all classes, as shown in Eq. (5).

$$GI(\mathbf{x}_j) = \frac{1}{N} \sum_{i=1}^N a_j^{(i)} \quad (5)$$

**GI Evaluation.** The evaluation on the GI score focuses on two aspects: (1) the removal of high GI ranked feature should have a higher drop in model performance compared to low GI ranked feature, and (2) the order of GI ranks should be consistent within models. For model performance, we employ the renowned RemOve And Retrain (ROAR) method [16], which sequentially eliminates the most important (truth) and least important (inverse) features before retraining the model to maintain consistent train and test distributions (See Fig. 4). Based on ROAR, we report the *Drop-in*



**Figure 5:** (A) The GI (Global Importance) score for the MS Dataset [28] is provided.  $x_0$  to  $x_5$  denotes the feature index. (B) An example of WhichFinger’s CWRI (Class-Wise Relative Importance) score: columns represent sensors (features), rows denote classes, and cell values convey CWRI scores. Red indicates the higher relative importance of the feature for the class, whereas blue denotes features of lesser importance in the context of the specific class.

*Accuracy (DA)*, which is the drop in model performance after excluding 20% of the important features. Second, to complement the manual selection of K% of features to be removed, we introduce the *Weighted Drop in Accuracy (WDA)* metric, which measures the decrease in accuracy when important features are removed sequentially, giving greater weight to scenarios where a smaller fraction of features is dropped. In real-world scenarios, practitioners and developers are often more interested in discerning the most or least important features rather than knowing mediocre features. Such scenarios arise when we have to extensively reduce the number of features to fit in a small memory budget or remove features that do not contribute to model performance. This weighting scheme ensures that the impact of removing high ranked GI feature is more pronounced (see Appendix C.1). Third, as an enhancement to the ROAR, we introduce the *Area Between Curves (ABC)* metric (Eq. (6)) based on the trapezoid rule [49], quantifying the area enclosed by the inverse ( $f(x)$ ) and truth curve ( $g(x)$ ) between the interval  $[a, b]$ ; a larger ABC value denotes a more precise feature ranking assessment, highlighted as gray areas in Fig. 4.

$$\begin{aligned} \text{ABC} &\triangleq \int_a^b (f(x) - g(x)) dx \\ &\simeq \frac{1}{2} \sum_{i=1}^{n-1} (f(x_i) - g(x_i) + f(x_{i+1}) - g(x_{i+1})) \Delta x_i \end{aligned} \quad (6)$$

To assess the consistency in GI rankings produced for different model runs, we utilize the Spearman correlation ( $\rho_S$ ) [29] and Kendall correlation ( $\rho_K$ ) [1] based on a 5-fold cross-validation (CV). **Class-Wise Relative Importance (CWRI).** While GI offers a global view of feature importance, CWRI provides detailed, class-specific insights into the role of each feature for every class  $c \in \{1, \dots, C\}$ . This approach is particularly valuable because a feature with a high GI score may not necessarily be of high importance for each individual class. CWRI, therefore, provides a class-centric perspective of feature importance in MTS data. We define the CWRI score for class  $c$ , denoted as  $\text{CWRI}(c) \in \mathbb{R}^D$ , as outlined in Eq. (7). For instance, the rows of Fig. 5-B shows the CWRI score for each class in the WhichFinger dataset. This class-specific score is derived

by evaluating the deviation in the class prototype of class  $c$  in comparison to other classes  $k \neq c$ .

$$\text{CWRI}(c) \triangleq \bar{a}_c - \frac{1}{C-1} \sum_{k \neq c} \bar{a}_k, \bar{a}_c: \text{see Eq. (2)} \quad (7)$$

A key aspect of CWRI is its relative computation across different classes, ensuring that the sum of CWRI scores for any given feature is zero. This method naturally produces both positive and negative CWRI values for the same feature across various classes. To illustrate, in a situation with three classes, the CWRI for feature  $x_j$  might be +1.7 for Class 1, -1.4 for Class 2, and -0.3 for Class 3. Here, a positive CWRI of feature  $x_j$  for Class 1 indicates higher relative importance of this feature in classifying Class 1 compared to Classes 2 and 3. The negative scores in Classes 2 and 3 indicate that these features were relatively unimportant. This approach in relatively calculating the difference is advantageous over simple class average scoring, which can be ambiguous and less informative, particularly when classes exhibit similar scores.

**CWRI Evaluation.** We utilize the CWRI scores to categorize each class and feature into relatively important (positive scores,  $x_j \geq 0$ ; red cells in Fig. 5-B) and relatively unimportant sets (negative scores,  $x_j < 0$ ; blue cells). Our evaluation compares these categorized feature sets against the established ground truth to determine the accuracy of feature importance identification. Given the lack of public datasets with known class discriminative feature importance, we created both real-world and synthetic datasets specifically for this purpose (detailed in Sec. 5). The comparison between our identified important/unimportant sets and the ground truths employs binary classification metrics like the F1 score, Jaccard index, and accuracy (we use the term *interpretative accuracy (IACC)* for clarity over standard model accuracy). The methodology for establishing these ground truths is elaborated in Appendix E.

## 5 DATASET AND BASELINE

**GI Datasets.** For GI measure, we curated two large public datasets, chosen for their substantial size and relevance (Appendix F). The Gilon Activity (7-class) comprises 14 features collected from smart insoles utilized by 72 users [22]. The Microsoft (MS) Activity (10-class) contains 6 features from armbands worn by 92 users [28]. The details of all the datasets can be found in Appendix F.

**CWRI Datasets.** Evaluating the CWRI measure on existing public datasets is challenging due to: (1) lacking in-depth comprehension of the features that influence the performance for each class [41], and (2) unmet requirements for ample classes ( $C \geq 3$ ), features ( $D \geq 3$ ), and samples ( $N \geq 10,000$ ) for generalization. Thus, we introduce synthetic and real-world MTS datasets, as follows.

**Dataset 1: SquidGame** ( $C = 3; D = 10; N = 54,000$ ). We designed the SquidGame dataset (Fig. 6-A), a 3-class synthetic MTS data, comprising of 30 features, which are divided into sets  $G_{\text{circle}} = \{1, \dots, 10\}$ ,  $G_{\text{triangle}} = \{11, \dots, 20\}$ , and  $G_{\text{square}} = \{21, \dots, 30\}$ . For Class 1, distinct time signals, such as sine waves, are produced within the circular mask in the feature set  $G_{\text{circle}}$ . Meanwhile, Gaussian noise fills the remaining areas (grey region) outside the circular mask in  $G_{\text{circle}}$ . This process mirrors the approach taken by Ismail et al. [19]. Similarly, Classes 2 and 3 have unique time signals within their respective feature sets,  $G_{\text{triangle}}$  and  $G_{\text{square}}$ . For each MTS

Models	Methods	GILON						MS					
		ABC( $\uparrow$ )	DA( $\uparrow$ )	WDA( $\uparrow$ )	$\rho_S(\uparrow)$	$\rho_K(\uparrow)$	ACC( $\uparrow$ )	ABC( $\uparrow$ )	DA( $\uparrow$ )	WDA( $\uparrow$ )	$\rho_S(\uparrow)$	$\rho_K(\uparrow)$	ACC( $\uparrow$ )
ShuffleNet	GS	0.152	-0.451	0.195	0.105	0.081		-0.159	7.014	0.166	0.051	<b>0.066</b>	
	SVS	0.819	0.092	0.331	0.100	0.068		0.225	<b>11.062</b>	0.329	-0.068	-0.066	
	Saliency	<b>1.192</b>	<b>4.015</b>	<b>0.546</b>	<b>0.371</b>	<b>0.257</b>	<b>0.958</b>	0.039	-4.861	0.142	0.102	0.146	<b>0.846</b>
		FA	0.714	0.313	0.331	0.148	0.107	<b>0.491</b>	4.934	0.287	-0.091	-0.040	
	IG	0.210	-0.118	0.193	0.045	0.033		-0.015	-2.084	0.241	<b>0.074</b>	<b>0.066</b>	
	CE	0.684	3.325	0.478	0.207	0.142		<b>0.355</b>	7.781	<b>0.475</b>	0.062	0.013	
	CE+QR(Ours)	<b>1.227</b>	<b>8.157</b>	<b>0.760</b>	<b>0.352</b>	<b>0.270</b>	<b>0.945</b>	0.337	<b>21.03</b>	<b>0.450</b>	<b>0.640</b>	<b>0.546</b>	<b>0.810</b>
ResNet	GS	1.284	<b>3.935</b>	<b>0.671</b>	<b>0.453</b>	<b>0.318</b>		0.059	<b>8.438</b>	<b>0.356</b>	0.051	0.013	
	SVS	0.577	0.835	0.281	0.134	0.116		0.160	4.520	0.344	<b>0.360</b>	<b>0.280</b>	
	Saliency	<b>1.338</b>	2.652	<b>0.663</b>	0.364	0.287		0.154	<b>8.327</b>	0.345	0.251	0.200	<b>0.752</b>
		FA	0.847	0.118	0.440	-0.014	-0.024	<b>0.350</b>	7.037	<b>0.373</b>	0.108	0.066	
	IG	<b>1.289</b>	2.137	0.620	<b>0.453</b>	<b>0.318</b>		-0.005	7.935	0.244	0.051	0.013	
	CE	0.801	1.553	0.420	0.370	0.270		<b>0.175</b>	-0.667	0.007	-0.062	-0.066	
	CE+QR(Ours)	1.163	<b>6.393</b>	0.615	<b>0.581</b>	<b>0.441</b>	<b>0.940</b>	0.117	1.266	0.254	<b>0.440</b>	<b>0.333</b>	<b>0.787</b>
MLP-Mixer	GS	-0.307	0.494	0.112	0.230	0.156		0.294	2.572	0.218	-0.040	0.013	
	SVS	-0.708	3.330	0.063	0.164	0.112		<b>0.328</b>	3.983	0.226	-0.137	-0.120	
	Saliency	0.986	3.446	0.490	<b>0.525</b>	<b>0.411</b>		0.321	<b>6.642</b>	<b>0.263</b>	<b>0.040</b>	<b>0.040</b>	<b>0.736</b>
		FA	0.457	1.041	0.269	-0.065	-0.046	0.920	0.300	3.209	0.226	-0.045	-0.013
	IG	-0.301	0.787	0.115	0.235	0.169		<b>0.326</b>	2.788	<b>0.263</b>	-0.040	0.013	
	CE	<b>0.920</b>	<b>8.117</b>	<b>0.531</b>	<b>0.330</b>	<b>0.239</b>		0.125	0.730	0.165	-0.142	-0.093	
	CE+QR(Ours)	<b>1.144</b>	<b>8.105</b>	<b>0.697</b>	0.165	0.129	<b>0.922</b>	0.276	<b>6.093</b>	<b>0.289</b>	<b>0.908</b>	<b>0.813</b>	<b>0.726</b>
ViT	GS	-0.197	1.244	0.116	0.218	0.147		<b>0.169</b>	5.219	0.361	<b>0.360</b>	<b>0.280</b>	
	SVS	-0.607	<b>2.830</b>	0.083	0.131	0.081		<b>0.182</b>	<b>5.227</b>	0.318	-0.040	-0.066	
	Saliency	<b>0.587</b>	1.216	0.376	0.120	0.098		0.122	5.219	0.329	0.051	0.040	<b>0.703</b>
		FA	0.244	0.367	0.223	-0.054	-0.024	0.922	<b>0.182</b>	<b>5.227</b>	0.318	0.141	0.120
	IG	-0.176	1.244	0.119	0.200	0.138		0.169	5.219	0.361	<b>0.177</b>	<b>0.173</b>	
	CE	0.553	<b>3.512</b>	<b>0.387</b>	<b>0.407</b>	<b>0.292</b>		-0.054	-1.109	0.151	-0.097	-0.066	
	CE+QR(Ours)	<b>0.636</b>	2.046	<b>0.456</b>	<b>0.502</b>	<b>0.389</b>	<b>0.924</b>	0.128	<b>10.962</b>	<b>0.530</b>	0.120	0.093	<b>0.706</b>

Table 1: Performance Evaluation of GI Metrics on Gilon and MS datasets. Each performance metric is explained in Sec. 4. Optimal performance is indicated by values in bold red, while the second-highest performance is marked in bold black. See Appendix C for full comparison (including standard deviation from five-fold cross validation) of explainers based on raw MTS data such as LSTM, and TCN.

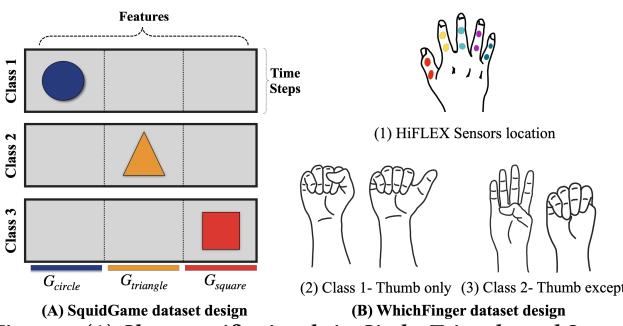


Figure 6: (A) Class-specific signals in Circle, Triangle, and Square masks; grey regions are filled with Gaussian noise. (B-1) The smart glove has 10 sensors, two per finger. (B-2, 3) We depict the data acquisition process for Class 1 and 2 for the WhichFinger. Specifically, Class 1 involves folding and unfolding movements of the thumb. In contrast, Class 2 is the complement of Class 1, focusing on the folding and unfolding of the remaining four fingers (See Appendix G).

instance, the location and size of the masks are randomly generated within the three feature sets to increase complexity.

**Dataset 2: WhichFinger** ( $C = 10; D = 10; N = 18,010$ ). We gathered a real-world MTS dataset using a smart glove [25] from 19 users, called WhichFinger (Fig. 6-B), to validate the CWRI measure. The smart glove incorporated with two sensors for each finger measures the resistance change in response to the tensile force exerted by each finger. We capture ten unique finger movements, achieved by either flexing and extending a single finger or a group of four fingers. Owing to the interlinked nature of hand muscles, we observe realistic correlations among features, making the task both intricate and non-trivial, providing a valuable MTS dataset for XAI applications. Detailed descriptions of the task design and data collection methodologies are provided in Appendix G.

**Implementation and Baselines.** We compare CAFO to several post-hoc explanation methods i.e., Gradient Shap (GS) [26], Shapley Value Sampling (SVS) [6], Saliency [35], Feature Ablation (FA) [37], Integrated Gradients (IG) [36], and DynaMask (DM) [7]. We utilized several deep architectures including vision-based deep models: ShuffleNet [51], ResNet [13], MLP-Mixer [39], and Vision Transformer (ViT) [8], and sequence based deep models like LSTM [15] and TCN. We adopt FIT [40], an explainer designed for recurrent models. We also employ LAXCAT [17], a 1-D CNN based MTS explainer. A detailed description is in Appendix H.

## 6 RESULTS

### 6.1 Evaluation of Global Importance

We evaluated CAFO’s performance in the GI measure using Gilon and MS datasets. The results for vision-based models like ShuffleNet, ResNet, MLP-Mixer, and ViT are in Tab. 1. Models using raw MTS format (e.g., LSTM, TCN, LAXCAT) and post-hoc methods relying on raw MTS (e.g., FIT, DynaMask) are detailed in Appendix C. Generally, vision-based models excel in most scenarios.

CAFO consistently demonstrates superior performance across key metrics (ABC, DA, WDA), highlighted in bold red in the Gilon dataset: notably, ShuffleNet (1.227), MLP-Mixer (1.144), and ViT (0.636) in the ABC metric. The use of QR-Ortho with cross-entropy (CE) significantly improved GI metric performance in most cases: 10 of 12 in Gilon and 9 of 12 in MS datasets. Notably, several baselines showed negative ABC scores, indicating a mismatch between critical feature identification and the drop in model accuracy, as seen in ROAR’s inverse and truth lines (Fig. 4). This suggests some baseline explainers inadequately measure GI rankings. Our findings show minimal difference in model accuracy between with and without QR-Ortho integration. The regularization parameter  $\lambda$  in

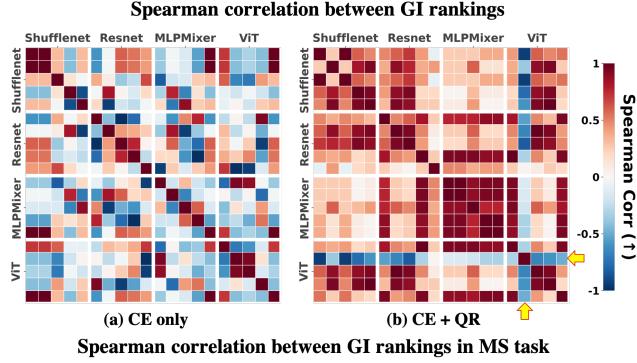


Figure 7: (a) Pairwise Spearman correlation ( $\rho_S$ ) of GI ranks from four models with cross-entropy (CE) loss in five-fold CV, revealing lower correlations and inconsistent GI ranks. (b) Enhanced GI rank consistency observed with QR-Ortho integration, demonstrated by higher  $\rho_S$  values in both inter and intra-model comparisons.

$\mathcal{L} = \mathcal{L}_{\text{CE}} + \lambda \mathcal{L}_{\text{QR}}$  was selected based on dataset characteristics, not through exhaustive tuning. We believe a more tailored selection of  $\lambda$  for each model may offer additional performance enhancements.

## 6.2 Consistency in GI Ranks

**6.2.1 Within Models.** Establishing consistency in model explanations is imperative for fostering user trust, as highlighted by Ribeiro et al. [31]. Conversely, models that yield divergent feature rankings across multiple runs undermine confidence in user’s decision-making processes. Although prior research [41] has underscored the variability in time-step importance produced with post-hoc explanation methods, our study is the initial effort to identify and quantitatively evaluate such variability in the context of feature-based importance. Our evaluation involves executing each model through 5 iterations of CV, with each fold serving once as a validation set and the remaining as training. This process yields 5 distinct feature rankings for the same left-out test set, from which we compute the pairwise Spearman’s  $\rho_S$  and Kendall’s  $\rho_K$  coefficients to gauge rank consistency, with the findings presented in Tab. 1. Notably, CAFO demonstrates the highest consistency in 11 out of 16 instances. Across the board, our results reveal there exists a huge variability in feature rankings, even under constant model architectures and explanatory methods. Alarmingly, certain explanatory methods yield negative correlations, indicating inverted ranking orders across different runs. These findings raise the need for more robust explanatory frameworks that can deliver dependable and stable feature rankings.

**6.2.2 Between Models.** Our analysis extends to assessing feature rank consistency across different models. The Spearman correlation coefficients, visualized as a heatmap in Fig. 7, reveal that the use of QR-Ortho significantly improves feature ranking consistency across models compared to CE alone, demonstrating CAFO’s robustness in providing consistent feature rankings, independent of model architecture. While different models naturally prioritize varying features for optimal performance, a degree of ranking consistency is a robustness indicator, fostering model trust. Additionally, analyzing ranking discrepancies offers deep insights. For example, as indicated by the yellow arrow in Fig. 7, we observe an anomaly

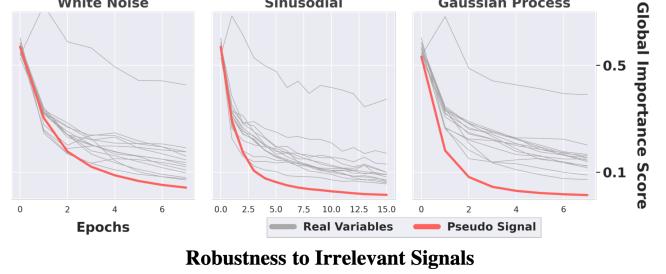


Figure 8: The GI metric at each training epoch is visualized, with bold red lines representing the irrelevant signal, while the thin grey lines correspond to the actual variables in the Gilon task. Over the course of training, the pseudo signals (i.e., bold red lines) consistently converge to the lowest GI values.

where a single run from the ViT model presents an entirely reversed feature ranking relative to all other runs. Such an outlier warrants further investigation by developers to ascertain the presence of potential errors or anomalies in the model training or data processing. These insights can prove invaluable in enhancing model reliability.

## 6.3 Robustness to Irrelevant Signals

In real world MTS problems, the overabundance of data often results in the accumulation of measurements from superfluous sensors [45]. Eliminating such irrelevant feature is, therefore, a critical task for practitioners. To assess the efficacy of CAFO in filtering out insignificant variables during model training, we generate pseudo signals [27] from time series processes: White noise, Sinusoidal, and Gaussian Process (detailed in Appendix J). The GI measure from each training epoch is visualized in Fig. 8. Initially, the pseudo-variable’s GI value is near 0.5, but it converges to the lowest GI ranking as training advances. This demonstrates CAFO’s robustness against non-significant variables and its potential to identify and discard non-significant features.

## 6.4 Class-Wise Relative Importance

Assessing feature relevance for specific classes is critical in applications like predictive maintenance in Heating, Ventilating, and Air Conditioning (HVAC) systems, where sensor importance varies by fault (class) type [38, 50] (Appendix A). Our CWRI methodology offers valuable information for sensor prioritization for each class.

In this study, we evaluate the ability of CAFO to identify critical class-wise features using the CWRI metric on two datasets: SquidGame and WhichFinger. As discussed in Sec. 5, these datasets come with ground truth labels indicating the relevance of features on a class-wise basis. As detailed in Tab. 2, CAFO surpasses other explanatory models in accurately identifying class-specific features in 13 out of 24 scenarios. We also note that integrating QR-Ortho Loss consistently enhances the discernment of class-wise relevant features across the table, compared to the standalone use of CE (cross-entropy) loss. Moreover, we observe a general performance degradation of all explainers in the WhichFinger dataset compared to SquidGame dataset, which may be attributed to the increased complexity inherent in real-world data. Such observation underscores the need for the development of real-world data oriented for XAI, especially in the time series domain.

Models	Methods	SquidGame			WhichFinger		
		F1(↑)	Jaccard(↑)	IACC(↑)	F1(↑)	Jaccard(↑)	IACC(↑)
ShuffleNet	GS	0.689	0.531	0.649	0.590	0.424	0.596
	SVS	<b>0.811</b>	<b>0.699</b>	<b>0.853</b>	<b>0.635</b>	<b>0.466</b>	<b>0.636</b>
	Saliency	0.533	0.371	0.489	0.601	0.436	<b>0.644</b>
	FA	0.765	0.624	0.818	0.552	0.383	0.588
	IG	0.690	0.531	0.649	0.582	0.414	0.588
	CE	0.561	0.394	0.518	0.429	0.273	0.530
ResNet	CE+QR(Ours)	<b>0.983</b>	<b>0.967</b>	<b>0.978</b>	<b>0.679</b>	<b>0.514</b>	0.606
	GS	0.747	0.598	0.693	0.630	0.462	0.636
	SVS	<b>0.758</b>	<b>0.621</b>	<b>0.800</b>	0.626	0.456	0.634
	Saliency	0.489	0.329	0.467	0.588	0.419	0.604
	FA	0.739	0.603	0.789	<b>0.670</b>	<b>0.503</b>	<b>0.696</b>
	IG	0.746	0.597	0.693	0.632	0.463	0.638
MLP-Mixer	CE	0.524	0.362	0.580	0.357	0.232	0.530
	CE+QR(Ours)	<b>0.987</b>	<b>0.974</b>	<b>0.982</b>	<b>0.724</b>	<b>0.570</b>	<b>0.698</b>
	GS	0.929	0.877	0.918	<b>0.859</b>	<b>0.753</b>	<b>0.862</b>
	SVS	<b>0.994</b>	<b>0.988</b>	<b>0.996</b>	0.778	0.641	0.784
	Saliency	0.514	0.349	0.453	0.726	0.571	0.744
	FA	<b>0.987</b>	<b>0.975</b>	<b>0.991</b>	0.732	0.586	0.752
ViT	IG	0.929	0.877	0.918	<b>0.861</b>	<b>0.756</b>	<b>0.864</b>
	CE	0.596	0.427	0.736	0.496	0.333	0.600
	CE+QR(Ours)	0.949	0.904	0.933	0.709	0.551	0.660
	GS	0.861	0.763	0.842	0.685	0.527	0.690
	SVS	<b>0.883</b>	<b>0.803</b>	<b>0.902</b>	<b>0.811</b>	<b>0.686</b>	<b>0.812</b>
	Saliency	0.526	0.359	0.484	<b>0.722</b>	<b>0.570</b>	<b>0.750</b>
	FA	0.779	0.640	0.809	0.674	0.512	0.684
	IG	0.861	0.763	0.842	0.702	0.544	0.710
	CE	0.700	0.546	0.796	0.520	0.352	0.536
	CE+QR(Ours)	<b>0.925</b>	<b>0.863</b>	<b>0.898</b>	0.531	0.363	0.572

**Table 2: Performance Evaluation of CWRI Metrics.** Here, the features identified as important by the model against the established ground truth importance is evaluated using binary metrics, including F1 score, Jaccard, and Accuracy (distinguished from model accuracy).

## 6.5 Additional Experiments

Due to the space constraint, additional experimental results are presented in the appendix, with key highlights summarized below.

**6.5.1 Other Image Encoding Methods.** We provide several main experiment results with the Gramian Angular Field image encoding method in Appendix B.

**6.5.2 Effect of  $\lambda$ .** We evaluated the effect of  $\lambda$  - a key hyperparameter in our model which modulates the QR-Ortho Loss (Eq. (4))-(ranging from 0 to 1) on two tasks: SquidGame and WhichFinger. Results indicate that increasing  $\lambda$  improves CWRI-related metrics, but excessively high values can reduce model accuracy. Detailed findings are in Appendix K.

**6.5.3 Alignment with Domain Knowledge.** Using Gilon and MS datasets, we demonstrated CAFO’s alignment with established domain insights. For the Gilon task, accelerometer features were crucial for speed differentiation, and in the MS task, similar activities yielded similar attention scores. Visual evidence of these correlations is provided in Appendix L.

## 6.6 Limitations and Discussions

We discuss the following limitations of CAFO. As our evaluation strategy for the GI method inherits the ROAR method [16], the retraining and re-evaluation cost is computationally intensive. Consequently, there is a need for alternative explanation methodologies that either do not rely on model accuracy or employ more computationally efficient evaluation techniques for the GI method. Additionally, our CAFO leverages image encoding to represent a time series into an image-like representation. While this approach has its merits, it also restricts the type of models used. As such, our research agenda includes the development of evaluation methods that are not only less demanding in terms of computational resources but also architecture-agnostic.

## 7 CONCLUSION

In this paper, we introduce CAFO, a feature-centric explanation framework for MTS classification. An in-depth discussion regarding the feature-centric explanation for MTS has been missing in much of the previous literature despite its huge importance, due to the lack of evaluation protocols, pertinent benchmarks, and methodologies. Addressing these problems, our contribution is threefold: First, we present CAFO, a channel attention-based feature explainer which combines a novel depth-wise channel attention module, DePCA, with QR-Ortho regularization for feature explanation in time series. Second, we curate a collection of both real-world and synthetic datasets, each annotated with known discriminative feature importance. Third, we introduce a set of feature importance metrics designed to quantify both global and class-specific importance, complete with corresponding evaluation schemes. We believe that our work will serve as a new groundwork for understanding feature importance within MTS classification.

## ACKNOWLEDGMENTS

This work was supported by the National Research Foundation of Korea(NRF) grant funded by the Korea government(MSIT) (RS-2023-00277383), and Institute of Information & communications Technology Planning & Evaluation(IITP) grant funded by the Korea government(MSIT) (No.RS-2020-II201336, Artificial Intelligence graduate school support(UNIST)). The authors extend their gratitude to the Gilon Corporation for inspiring this work. Special thanks are due to Prof. Sunghoon Lim, Gyeongho Kim, Sujin Jeon, and Jae Gyeong Choi for providing the smart glove essential for the WhichFinger data collection and for their valuable insights into our research. We are also grateful to MyoungHoon Lee, Prof. Suhyeon Kim, Wonho Sohn, and Hyewon Kang for their initial discussions that shaped our paper. Appreciation is further extended to Yeonjoo Kim, Solang Kim, Bosung Kim, Isu Jeong, Jaewook Lee, and Changhyeon Lee for their thorough review of our manuscript. Lastly, we thank the numerous anonymous reviewers whose constructive feedback significantly enhanced our work.

## REFERENCES

- [1] Hervé Abdi. 2007. The Kendall rank correlation coefficient. *Encyclopedia of Measurement and Statistics*. Sage, Thousand Oaks, CA (2007), 508–510.
- [2] João Bento, Pedro Saleiro, André F Cruz, Mário AT Figueiredo, and Pedro Bizarro. 2021. Timeshap: Explaining recurrent models through sequence perturbations. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, 2565–2573.
- [3] Jonas Beuchert, Friedrich Solowjow, Sebastian Trimpe, and Thomas Seel. 2020. Overcoming bandwidth limitations in wireless sensor networks by exploitation of cyclic signal patterns: An event-triggered learning approach. *Sensors* 20, 1 (2020), 260.
- [4] Åke Björck. 1994. Numerics of gram-schmidt orthogonalization. *Linear Algebra and Its Applications* 197 (1994), 297–316.
- [5] Tadeusz Caliński and Jerzy Harabasz. 1974. A dendrite method for cluster analysis. *Communications in Statistics-theory and Methods* 3, 1 (1974), 1–27.
- [6] Javier Castro, Daniel Gómez, and Juan Tejada. 2009. Polynomial calculation of the Shapley value based on sampling. *Computers & Operations Research* 36, 5 (2009), 1726–1730.
- [7] Jonathan Crabbé and Mihaela Van Der Schaar. 2021. Explaining time series predictions with dynamic masks. In *International Conference on Machine Learning*. PMLR, 2166–2177.
- [8] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929* (2020).

- [9] Jean-Pierre Eckmann, S Ollifson Kamphorst, David Ruelle, et al. 1995. Recurrence plots of dynamical systems. *World Scientific Series on Nonlinear Science Series A* 16 (1995), 441–446.
- [10] Pavel Filonov, Andrey Lavrentyev, and Artem Vorontsov. 2016. Multivariate industrial time series with cyber-attack simulation: Fault detection using an lstm-based predictive data model. *arXiv preprint arXiv:1612.06676* (2016).
- [11] Colin R Goodall. 1993. 13 Computation using the QR decomposition. (1993).
- [12] Jianing He, Xiaolong Gong, and Linpeng Huang. 2021. Wavelet-temporal neural network for multivariate time series prediction. In *2021 International Joint Conference on Neural Networks (IJCNN)*. IEEE, 1–8.
- [13] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 770–778.
- [14] Sepp Hochreiter. 1998. The vanishing gradient problem during learning recurrent neural nets and problem solutions. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems* 6, 02 (1998), 107–116.
- [15] Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation* 9, 8 (1997), 1735–1780.
- [16] Sara Hooker, Dumitru Erhan, Pieter-Jan Kindermans, and Been Kim. 2019. A benchmark for interpretability methods in deep neural networks. *Advances in neural information processing systems* 32 (2019).
- [17] Tsung-Yu Hsieh, Suhang Wang, Yiwei Sun, and Vasant Honavar. 2021. Explainable multivariate time series classification: a deep neural network which learns to attend to important variables as well as time intervals. In *Proceedings of the 14th ACM international conference on web search and data mining*. 607–615.
- [18] Jie Hu, Li Shen, and Gang Sun. 2018. Squeeze-and-excitation networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 7132–7141.
- [19] Aya Abdelsalam Ismail, Mohamed Gunady, Hector Corrada Bravo, and Soheil Feizi. 2020. Benchmarking deep learning interpretability in time series predictions. *Advances in neural information processing systems* 33 (2020), 6441–6452.
- [20] Aya Abdelsalam Ismail, Mohamed Gunady, Luiz Pessoa, Hector Corrada Bravo, and Soheil Feizi. 2019. Input-cell attention reduces vanishing saliency of recurrent neural networks. *Advances in Neural Information Processing Systems* 32 (2019).
- [21] Maowei Jiang, Pengyu Zeng, Kai Wang, Huan Liu, Wenbo Chen, and Haoran Liu. 2022. FECAM: Frequency Enhanced Channel Attention Mechanism for Time Series Forecasting. *arXiv preprint arXiv:2212.01209* (2022).
- [22] Jaeho Kim, Hyewon Kang, Jaewan Yang, Haneul Jung, Seulki Lee, and Junghye Lee. 2023. Multi-task Deep Learning for Human Activity, Speed, and Body Weight Estimation using Commercial Smart Insoles. *IEEE Internet of Things Journal* (2023).
- [23] Colin Lea, Michael D Flynn, Rene Vidal, Austin Reiter, and Gregory D Hager. 2017. Temporal convolutional networks for action segmentation and detection. In *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 156–165.
- [24] Jiyoung Lee, Hyungrok Do, Mingu Kwak, Hyungu Kahng, and Seoung Bum Kim. 2021. Hierarchical segment-channel attention network for explainable multi-channel signal classification. *Information Sciences* 567 (2021), 312–331.
- [25] Minhyuk Lee and Joonbum Bae. 2020. Deep learning based real-time recognition of dynamic finger gestures using a data glove. *IEEE Access* 8 (2020), 219923–219933.
- [26] Scott M Lundberg and Su-In Lee. 2017. A unified approach to interpreting model predictions. *Advances in neural information processing systems* 30 (2017).
- [27] JR Maat, A Malali, and P Protopapas. 2017. Timesynth: A multipurpose library for synthetic time series in python.
- [28] Dan Morris, T Scott Saponas, Andrew Guillory, and Ilya Kelner. 2014. RecoFit: using a wearable sensor to find, recognize, and count repetitive exercises. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. 3225–3234.
- [29] Leann Myers and Maria J Sirois. 2004. Spearman correlation coefficients, differences between. *Encyclopedia of statistical sciences* 12 (2004).
- [30] Jongchan Park, Sanghyun Woo, Joon-Young Lee, and In So Kweon. 2018. Bam: Bottleneck attention module. *arXiv preprint arXiv:1807.06514* (2018).
- [31] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. "Why should i trust you?" Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*. 1135–1144.
- [32] Udo Schlegel, Hiba Arnout, Mennatallah El-Assady, Daniela Oelke, and Daniel A Keim. 2019. Towards a rigorous evaluation of XAI methods on time series. In *2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW)*. IEEE, 4197–4201.
- [33] Robert Schreiber and Charles Van Loan. 1989. A storage-efficient WY representation for products of Householder transformations. *SIAM J. Sci. Statist. Comput.* 10, 1 (1989), 53–57.
- [34] Shoaib Ahmed Siddiqui, Dominique Mercier, Mohsin Munir, Andreas Dengel, and Sheraz Ahmed. 2019. Tsviz: Demystification of deep learning models for time-series analysis. *IEEE Access* 7 (2019), 67027–67040.
- [35] Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. 2013. Deep inside convolutional networks: Visualising image classification models and saliency maps. *arXiv preprint arXiv:1312.6034* (2013).
- [36] Mukund Sundararajan, Ankur Taly, and Qiqi Yan. 2017. Axiomatic attribution for deep networks. In *International conference on machine learning*. PMLR, 3319–3328.
- [37] Harini Suresh, Nathan Hunt, Alistair Johnson, Leo Anthony Celi, Peter Szolovits, and Marzyeh Ghassemi. 2017. Clinical intervention prediction and understanding using deep networks. *arXiv preprint arXiv:1705.08498* (2017).
- [38] Saman Taheri, Amirhossein Ahmadi, Behnam Mohammadi-Ivatloo, and Somayeh Asadi. 2021. Fault detection diagnostic for HVAC systems via deep learning algorithms. *Energy and Buildings* 250 (2021), 111275.
- [39] Ilya O Tolstikhin, Neil Houlsby, Alexander Kolesnikov, Lucas Beyer, Xiaohua Zhai, Thomas Unterthiner, Jessica Yung, Andreas Steiner, Daniel Keysers, Jakob Uszkoreit, et al. 2021. Mlp-mixer: An all-mlp architecture for vision. *Advances in neural information processing systems* 34 (2021), 24261–24272.
- [40] Sana Tonekaboni, Shalmali Joshi, Kieran Campbell, David K Duvenaud, and Anna Goldenberg. 2020. What went wrong and when? Instance-wise feature importance for time-series black-box models. *Advances in Neural Information Processing Systems* 33 (2020), 799–809.
- [41] Hugues Turbé, Mina Bjelogrlic, Christian Lovis, and Gianmarco Mengaldo. 2023. Evaluation of post-hoc interpretability methods in time-series classification. *Nature Machine Intelligence* 5, 3 (2023), 250–260.
- [42] Laurens Van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-SNE. *Journal of machine learning research* 9, 11 (2008).
- [43] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems* 30 (2017).
- [44] Zhiguang Wang, Tim Oates, et al. 2015. Encoding time series as images for visual inspection and classification using tiled convolutional neural networks. In *Workshops at the twenty-ninth AAAI conference on artificial intelligence*, Vol. 1. AAAI Menlo Park, CA, USA.
- [45] Samuel R West, Ying Guo, X Rosalind Wang, and Joshua Wall. 2011. Automated fault detection and diagnosis of HVAC subsystems using statistical machine learning. (2011).
- [46] Sanghyun Woo, Jongchan Park, Joon-Young Lee, and In So Kweon. 2018. Cbam: Convolutional block attention module. In *Proceedings of the European conference on computer vision (ECCV)*. 3–19.
- [47] Feiyu Xu, Hans Uszkoreit, Yangzhou Du, Wei Fan, Dongyan Zhao, and Jun Zhu. 2019. Explainable AI: A brief survey on history, research areas, approaches and challenges. In *Natural Language Processing and Chinese Computing: 8th CCF International Conference, NLPCC 2019, Dunhuang, China, October 9–14, 2019, Proceedings, Part II*. Springer, 563–574.
- [48] Lingxiao Yang, Ru-Yuan Zhang, Lida Li, and Xiaohua Xie. 2021. Simam: A simple, parameter-free attention module for convolutional neural networks. In *International conference on machine learning*. PMLR, 11863–11874.
- [49] Shi-Tao Yeh et al. 2002. Using trapezoidal rule for the area under a curve calculation. *Proceedings of the 27th Annual SAS® User Group International (SUGI'02)* (2002), 1–5.
- [50] Liang Zhang and Jin Wen. 2019. A systematic feature selection procedure for short-term data-driven building energy forecasting model development. *Energy and Buildings* 183 (2019), 428–442.
- [51] Xiangyu Zhang, Xinyu Zhou, Mengxiao Lin, and Jian Sun. 2018. Shufflenet: An extremely efficient convolutional neural network for mobile devices. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 6848–6856.
- [52] Zhibin Zhao, Tianfu Li, Jingyao Wu, Chuang Sun, Shabin Wang, Ruqiang Yan, and Xuefeng Chen. 2020. Deep learning algorithms for rotating machinery intelligent diagnosis: An open source benchmark study. *ISA transactions* 107 (2020), 224–255.

## A MOTIVATION REGARDING FEATURE-BASED IMPORTANCE

This section provides motivating examples to further illustrate the benefits of the proposed GI and CWRI measures in the MTS context and how they can be utilized in the real-world scenarios, especially for both engineers and domain experts.

### A.1 Case Studies

**Case Study 1: Optimization of Fault Detection and Diagnosis (FDD) in Heating, Ventilation, and Air Conditioning (HVAC) Systems** is an important issue in energy preservation as it can prevent excessive energy consumption in buildings [3, 5]. Preemptive maintenance in HVAC systems aims to predict potential equipment failure, using data from Internet of Things (IoT) sensors embedded within the system [1]. Despite the integration of numerous sensors and controllers to monitor a variety of system faults, fault detection in HVAC systems remains a complex task due to the non-linearity of fault patterns and strong feature correlations. The situation is further complicated by the hundreds of sensors indicating different types of defects across varying operational modes, such as air conditioning and heating [6, 8]. Thus, identifying the sensors most contributory to a specific fault can aid maintenance specialists in accurately diagnosing the malfunction. Prioritizing sensors capable of diagnosing a broad spectrum of faults can enhance system efficiency. This comprehension of the global and class-specific importance of features provides invaluable insights into sensor design and placement, beneficial for manufacturers. Armed with these insights, domain experts can elucidate the role of each feature in fault detection, aiding engineers in the development of more effective fault detection models.

**Case Study 2. The Smart insole** contains numerous sensors, such as an accelerometer, force-sensitive resistor (FSR), and gyroscope, to capture various gait characteristics. The data acquired from these sensors serve to monitor a range of human activities. Notably, in the process of recognizing these activities, some sensors may prove multi-functional across several classes, while others may be specifically critical for a distinct task. For example, an accelerometer may be instrumental in detecting varying speed ranges of movement, whereas an FSR sensor could effectively discern the user's posture. Yet, how can we ascertain that the black-box model is appropriately employing these features? In this scenario, the Global Importance (GI) and Class-Wise Relevant Importance (CWRI), calculated from CAFO, offer invaluable insights into the model's sensor prioritization. While GI denotes the sensor that contributes most significantly to enhanced classification accuracy, a high GI score for an accelerometer, for instance, does not necessarily confer high importance to each class (e.g., Squat, Lunge). CWRI, on the other hand, provides class-specific insights into feature prioritization.

Understanding the role of each sensor affords substantial benefits for both engineers and practical applications. Engineers, for instance, can harness this information to implement feature selection and extraction strategies for predictive models. Moreover, manufacturers can leverage this knowledge to design more efficient and cost-effective smart insoles, by eliminating redundant sensors in the production phase, thus lowering both production and data transfer costs for cloud processing.

## B IMAGE ENCODING METHODS

### B.1 Recurrence Plot

A univariate time series  $\mathbf{x} = [x_1, \dots, x_T]^\top$  is converted to a two-dimensional recurrence plot (RP) image by composing  $L \triangleq T - (m - 1)\tau$  number of new vectors  $\mathbf{v}_1, \dots, \mathbf{v}_L$ , where  $\mathbf{v}_k$  is a vector consisting of raw time series as in Eq. (8). Here  $\tau$  is the time delay, and  $m$  is the embedding dimension selected as a hyperparameter.

$$\mathbf{v}_k = [x_k, x_{k+\tau}, x_{k+2\tau}, \dots, x_{k+(m-1)\tau}]^\top \quad (8)$$

These vectors are finally transformed into an RP image by measuring the pairwise distance between all vectors  $\mathbf{v}_1, \dots, \mathbf{v}_L$ , leading to an  $L \times L$  image. Each element of the RP image is defined in Eq. (9). The threshold distance  $\varepsilon$ , Heaviside step function  $\mathbb{H}$ , and a norm function  $\|\cdot\|$  should be properly selected.

$$\text{RP}_{i,j} = \mathbb{H}(\varepsilon - \|\mathbf{v}_j - \mathbf{v}_i\|), \forall i, j \in [L] \quad (9)$$

In all experiments, we used  $\tau = 1$ ,  $m = 1$ . The threshold  $\varepsilon$  was set to 10% of the maximum distance.

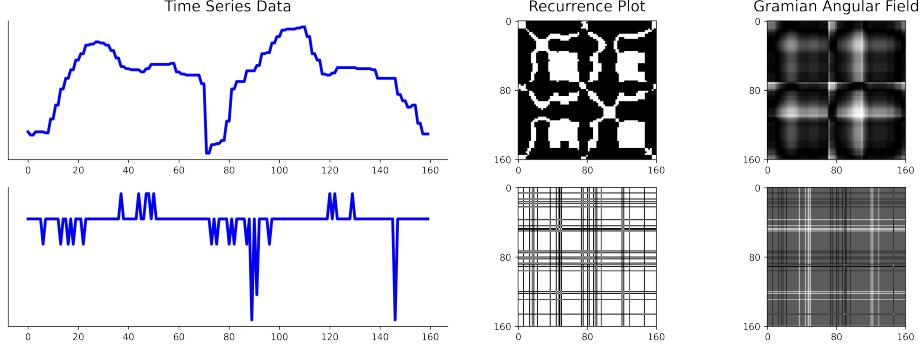
### B.2 Gramian Angular Field

The Gramian Angular Field (GAF) is an image encoding method used for time series data. It has two subtypes: the Gramian Angular Summation Field (GASF) and the Gramian Angular Difference Field (GADF). In our supplementary experiment, we utilized the GASF subtype of GAF and referred to them as GAF unless otherwise indicated.

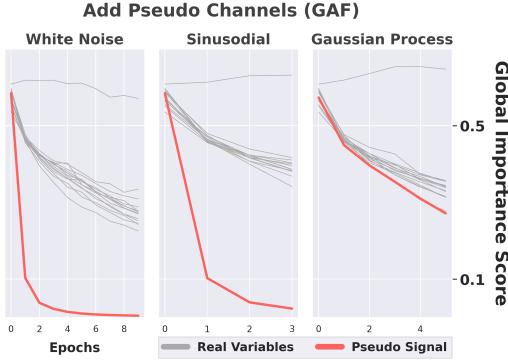
The data is first scaled between the range  $[-1, 1]$  to represent the time series in a polar coordinate system using  $\phi_i = \arccos(x_i)$  for  $i = 1, \dots, T$ . A Gram matrix is calculated between all pairs of  $\phi_{i=1, \dots, T}$  as in Eq. (10) and is used as the GAF.

$$\text{GAF} = \begin{pmatrix} \cos(\phi_1 + \phi_1) & \cos(\phi_1 + \phi_2) & \cdots & \cos(\phi_1 + \phi_T) \\ \cos(\phi_2 + \phi_1) & \cos(\phi_2 + \phi_2) & \cdots & \cos(\phi_2 + \phi_T) \\ \vdots & \vdots & \ddots & \vdots \\ \cos(\phi_T + \phi_1) & \cos(\phi_T + \phi_2) & \cdots & \cos(\phi_T + \phi_T) \end{pmatrix} \quad (10)$$

An illustrative figure of how time series is converted to an RP and GAF is shown in Fig. 9.



**Figure 9: Different Image Encoding Methods** The time series presented in the left column is encoded with Recurrence Plot (middle) and Gramian Angular Field (right).



**Figure 10: Robustness to Pseudo Signals (GAF)** Pseudo signals were incorporated and trained on the Gilon activity task, as illustrated in ???. In this instance, we employed the GAF image encoding method rather than RP. The findings demonstrate that encoding techniques other than RP exhibit robustness against pseudo signals.

Models	Methods	SquidGame			
		F1( $\uparrow$ )	Jaccard( $\uparrow$ )	IACC( $\uparrow$ )	Jaccard( $\uparrow$ )
ShuffleNet	CE	0.609 $\pm$ 0.07	0.441 $\pm$ 0.07	0.527 $\pm$ 0.09	0.441 $\pm$ 0.07
	CE+QR	0.815 $\pm$ 0.06	0.692 $\pm$ 0.09	0.869 $\pm$ 0.05	0.692 $\pm$ 0.09
	GAINS	<b>0.21</b>	<b>0.25</b>	<b>0.34</b>	<b>0.25</b>
ResNet	CE	0.673 $\pm$ 0.06	0.510 $\pm$ 0.06	0.598 $\pm$ 0.04	0.510 $\pm$ 0.06
	CE+QR	0.817 $\pm$ 0.05	0.693 $\pm$ 0.07	0.871 $\pm$ 0.03	0.693 $\pm$ 0.07
	GAINS	<b>0.14</b>	<b>0.18</b>	<b>0.27</b>	<b>0.18</b>
MLP-Mixer	CE	0.475 $\pm$ 0.06	0.314 $\pm$ 0.06	0.622 $\pm$ 0.10	0.314 $\pm$ 0.06
	CE+QR	0.703 $\pm$ 0.10	0.551 $\pm$ 0.12	0.778 $\pm$ 0.08	0.551 $\pm$ 0.12
	GAINS	<b>0.23</b>	<b>0.24</b>	<b>0.16</b>	<b>0.24</b>
ViT	CE	0.751 $\pm$ 0.08	0.607 $\pm$ 0.10	0.678 $\pm$ 0.10	0.607 $\pm$ 0.10
	CE+QR	0.811 $\pm$ 0.04	0.684 $\pm$ 0.06	0.860 $\pm$ 0.03	0.684 $\pm$ 0.06
	GAINS	<b>0.06</b>	<b>0.08</b>	<b>0.18</b>	<b>0.08</b>

**Table 3: Performance Evaluation of CWRI Metrics using GAF encoding on the SquidGame task.** We observe that using QR-Ortho Loss improves the identification of important features. Here, we set  $\lambda = 1$  for all models.

Panel A. Performance Comparison of CAFO with CE and QR-Ortho loss

Models	Methods	GILON										MS				
		ABC( $\uparrow$ )	DA( $\uparrow$ )	WDA( $\uparrow$ )	$\rho_S(\uparrow)$	$\rho_K(\uparrow)$	ACC( $\uparrow$ )	ABC( $\uparrow$ )	DA( $\uparrow$ )	WDA( $\uparrow$ )	$\rho_S(\uparrow)$	$\rho_K(\uparrow)$	ACC( $\uparrow$ )			
ShuffleNet	GS	0.152	-0.451	0.195	0.105 $\pm$ 0.27	0.081 $\pm$ 0.22		-0.159	7.014	0.166	0.051 $\pm$ 0.67	<b>0.066<math>\pm</math>0.54</b>				
	SVS	0.819	0.092	0.331	0.100 $\pm$ 0.28	0.068 $\pm$ 0.21		0.225	<b>11.062</b>	0.329	-0.068 $\pm$ 0.44	-0.066 $\pm$ 0.38				
	Saliency	<b>1.192</b>	<b>4.015</b>	<b>0.546</b>	<b>0.371<math>\pm</math>0.27</b>	<b>0.257<math>\pm</math>0.24</b>		0.039	-4.861	0.142	0.102 $\pm$ 0.60	0.146 $\pm$ 0.51				
	FA	0.714	0.313	0.331	0.148 $\pm$ 0.32	0.107 $\pm$ 0.24		<b>0.958</b>	<b>0.491</b>	4.934	0.287	-0.091 $\pm$ 0.46	-0.040 $\pm$ 0.36			<b>0.846</b>
	IG	0.210	-0.118	0.193	0.045 $\pm$ 0.22	0.033 $\pm$ 0.16		-0.015	-2.084	0.241	<b>0.074<math>\pm</math>0.69</b>	<b>0.066<math>\pm</math>0.55</b>				
	CE	0.684	3.325	0.478	0.207 $\pm$ 0.23	0.142 $\pm$ 0.18		0.355	7.781	<b>0.475</b>	0.062 $\pm$ 0.50	0.013 $\pm$ 0.46				
ResNet	CE+QR(Ours)	<b>1.227</b>	<b>8.157</b>	<b>0.760</b>	<b>0.352<math>\pm</math>0.26</b>	<b>0.270<math>\pm</math>0.22</b>		<b>0.945</b>	<b>0.337</b>	<b>21.032</b>	<b>0.450</b>	<b>0.640<math>\pm</math>0.26</b>	<b>0.546<math>\pm</math>0.28</b>			<b>0.810</b>
	GS	1.284	<b>3.935</b>	<b>0.671</b>	<b>0.453<math>\pm</math>0.23</b>	<b>0.318<math>\pm</math>0.21</b>		0.059	<b>8.438</b>	<b>0.356</b>	0.051 $\pm$ 0.53	0.013 $\pm$ 0.45				
	SVS	0.577	0.835	0.281	0.134 $\pm$ 0.20	0.116 $\pm$ 0.15		0.160	4.520	0.344	<b>0.360<math>\pm</math>0.26</b>	<b>0.280<math>\pm</math>0.23</b>				
	Saliency	<b>1.338</b>	2.652	<b>0.663</b>	0.364 $\pm$ 0.18	0.287 $\pm$ 0.13		0.154	<b>8.327</b>	0.345	0.251 $\pm$ 0.39	0.200 $\pm$ 0.34				<b>0.757</b>
	FA	0.847	0.118	0.440	-0.014 $\pm$ 0.19	-0.024 $\pm$ 0.15		<b>0.960</b>	<b>0.350</b>	7.037	<b>0.373</b>	0.108 $\pm$ 0.43	0.066 $\pm$ 0.32			
	IG	<b>1.289</b>	2.137	0.620	<b>0.453<math>\pm</math>0.23</b>	<b>0.318<math>\pm</math>0.21</b>		-0.005	7.935	0.244	0.051 $\pm$ 0.53	0.013 $\pm$ 0.45				
MLP-Mixer	CE	0.801	1.553	0.420	0.370 $\pm$ 0.25	0.270 $\pm$ 0.21		<b>0.175</b>	-0.667	0.007	-0.062 $\pm$ 0.53	-0.066 $\pm$ 0.44				
	CE+QR(Ours)	1.163	<b>6.393</b>	0.615	<b>0.581<math>\pm</math>0.13</b>	<b>0.441<math>\pm</math>0.11</b>		<b>0.940</b>	0.117	1.266	0.254	<b>0.440<math>\pm</math>0.28</b>	<b>0.333<math>\pm</math>0.27</b>			<b>0.787</b>
	GS	-0.307	0.494	0.112	0.230 $\pm$ 0.25	0.156 $\pm$ 0.18		0.294	2.572	0.218	-0.040 $\pm$ 0.53	0.013 $\pm$ 0.43				
	SVS	-0.708	3.330	0.063	0.164 $\pm$ 0.32	0.112 $\pm$ 0.23		<b>0.328</b>	3.983	0.226	-0.137 $\pm$ 0.44	-0.120 $\pm$ 0.36				
	Saliency	0.986	3.446	0.490	<b>0.525<math>\pm</math>0.26</b>	<b>0.411<math>\pm</math>0.24</b>		<b>0.920</b>	0.321	<b>6.642</b>	<b>0.263</b>	<b>0.040<math>\pm</math>0.56</b>	<b>0.040<math>\pm</math>0.45</b>			<b>0.736</b>
	FA	0.457	1.041	0.269	-0.065 $\pm$ 0.33	-0.046 $\pm$ 0.23		0.300	3.209	0.226	-0.045 $\pm$ 0.68	-0.013 $\pm$ 0.62				
ViT	IG	-0.301	0.787	0.115	0.235 $\pm$ 0.26	0.169 $\pm$ 0.19		<b>0.326</b>	2.788	<b>0.263</b>	-0.040 $\pm$ 0.53	0.013 $\pm$ 0.43				
	CE	<b>0.920</b>	<b>8.117</b>	<b>0.531</b>	<b>0.330<math>\pm</math>0.24</b>	<b>0.239<math>\pm</math>0.20</b>		0.125	0.730	0.165	-0.142 $\pm$ 0.46	-0.093 $\pm$ 0.28				
	CE+QR(Ours)	<b>1.144</b>	<b>8.105</b>	<b>0.697</b>	0.165 $\pm$ 0.37	0.129 $\pm$ 0.29		<b>0.922</b>	0.276	<b>6.093</b>	<b>0.289</b>	<b>0.908<math>\pm</math>0.08</b>	<b>0.813<math>\pm</math>0.14</b>			<b>0.726</b>
	GS	-0.197	1.244	0.116	0.218 $\pm$ 0.39	0.147 $\pm$ 0.30		<b>0.169</b>	5.219	0.361	<b>0.360<math>\pm</math>0.34</b>	<b>0.280<math>\pm</math>0.35</b>				
	SVS	-0.607	<b>2.830</b>	0.083	0.131 $\pm$ 0.25	0.081 $\pm$ 0.18		<b>0.182</b>	<b>5.227</b>	0.318	-0.04 $\pm$ 0.46	-0.066 $\pm$ 0.35				
	Saliency	<b>0.587</b>	1.216	0.376	0.120 $\pm$ 0.24	0.098 $\pm$ 0.20		0.122	5.219	0.329	0.051 $\pm$ 0.67	0.040 $\pm$ 0.58				<b>0.703</b>
ViT	FA	0.244	0.367	0.223	-0.054 $\pm$ 0.43	-0.024 $\pm$ 0.30		<b>0.922</b>	<b>0.182</b>	<b>5.227</b>	0.318	0.142 $\pm$ 0.59	0.120 $\pm$ 0.51			
	IG	-0.176	1.244	0.119	0.200 $\pm$ 0.38	0.138 $\pm$ 0.29		0.169	5.219	0.361	<b>0.177<math>\pm</math>0.51</b>	<b>0.173<math>\pm</math>0.44</b>				
	CE	0.553	<b>3.512</b>	<b>0.387</b>	<b>0.407<math>\pm</math>0.13</b>	<b>0.292<math>\pm</math>0.11</b>		-0.054	-1.109	0.151	-0.097 $\pm$ 0.52	-0.066 $\pm$ 0.41				
	CE+QR(Ours)	<b>0.636</b>	2.046	<b>0.456</b>	<b>0.502<math>\pm</math>0.17</b>	<b>0.389<math>\pm</math>0.14</b>		<b>0.924</b>	0.128	<b>10.962</b>	<b>0.530</b>	0.120 $\pm$ 0.57	0.093 $\pm$ 0.48			<b>0.706</b>

Panel B. Performance Evaluation with Baseline Methods

Models	Methods	GILON										MS				
		ABC( $\uparrow$ )	DA( $\uparrow$ )	WDA( $\uparrow$ )	$\rho_S(\uparrow)$	$\rho_K(\uparrow)$	ACC( $\uparrow$ )	ABC( $\uparrow$ )	DA( $\uparrow$ )	WDA( $\uparrow$ )	$\rho_S(\uparrow)$	$\rho_K(\uparrow)$	ACC( $\uparrow$ )			
Baseline (LSTM)	GS	0.242	<b>1.404</b>	0.141	0.145 $\pm$ 0.39	0.130 $\pm$ 0.29		-0.230	-3.356	-0.027	-0.029 $\pm$ 0.63	-0.040 $\pm$ 0.54				
	SVS	-0.053	-0.278	0.245	0.427 $\pm$ 0.20	0.332 $\pm$ 0.15		-0.192	-3.356	-0.036	<b>0.703<math>\pm</math>0.20</b>	<b>0.547<math>\pm</math>0.27</b>				
	Saliency	-0.680	0.283	0.010	0.543 $\pm$ 0.17	<b>0.411<math>\pm</math>0.15</b>		-0.031	<b>4.517</b>	0.071	0.234 $\pm$ 0.33	0.200 $\pm$ 0.23				
	FA	0.098	-1.857	0.115	0.330 $\pm$ 0.16	0.235 $\pm$ 0.12		0.952	-0.143	-3.356	-0.036	<b>0.703<math>\pm</math>0.14</b>	<b>0.547<math>\pm</math>0.17</b>			<b>0.807</b>
	IG	0.241	<b>1.404</b>	0.135	0.200 $\pm$ 0.33	0.174 $\pm$ 0.25		-0.230	-3.356	-0.027	0.029 $\pm$ 0.48	0.013 $\pm$ 0.43				
	FIT	0.285	0.506	0.252	0.429 $\pm$ 0.27	0.301 $\pm$ 0.23		0.089	1.655	0.070	0.657 $\pm$ 0.20	0.493 $\pm$ 0.28				
Baseline (TCN)	DM	<b>0.354</b>	-0.843	<b>0.308</b>	<b>0.495<math>\pm</math>0.19</b>	0.358 $\pm$ 0.16		<b>0.150</b>	-2.127	<b>0.075</b>	0.063 $\pm$ 0.36	0.040 $\pm$ 0.28				
	GS	0.555	2.209	0.226	0.399 $\pm$ 0.22	0.275 $\pm$ 0.17		-0.048	2.513	0.058	0.189 $\pm$ 0.37	0.173 $\pm$ 0.31				
	SVS	0.113	-0.858	0.078	0.060 $\pm$ 0.24	0.046 $\pm$ 0.19		-0.335	1.653	-0.008	0.257 $\pm$ 0.37	0.253 $\pm$ 0.34				
	Saliency	0.528	<b>3.093</b>	0.248	<b>0.618<math>\pm</math>0.14</b>	<b>0.468<math>\pm</math>0.11</b>		0.007	3.054	0.105	0.234 $\pm$ 0.29	0.200 $\pm$ 0.30				0.815
	FA	0.230	-2.185	0.027	0.165 $\pm$ 0.15	0.134 $\pm$ 0.11		0.928	-0.065	1.927	0.017	<b>0.474<math>\pm</math>0.27</b>	<b>0.360<math>\pm</math>0.26</b>			
	IG	<b>0.666</b>	2.298	<b>0.240</b>	0.346 $\pm$ 0.18	0.235 $\pm$ 0.14		-0.048	0.764	0.030	0.189 $\pm$ 0.37	0.173 $\pm$ 0.31				
Baseline (None)	DM	-0.280	-0.270	-0.024	0.077 $\pm$ 0.21	0.046 $\pm$ 0.14		<b>0.186</b>	<b>4.270</b>	<b>0.140</b>	0.000 $\pm$ 0.56	-0.013 $\pm$ 0.49				
	Laxcat	-0.458	-4.217	-0.230	-0.016 $\pm$ 0.34	-0.007 $\pm$ 0.25		0.711	-0.225	1.003	0.139	0.074 $\pm$ 0.46	0.093 $\pm$ 0.40			0.759

Table 4: Performance Evaluation of GI Metrics: This table presents the evaluation of GI metrics such as ABC, DA, WDA,  $\rho_S$ ,  $\rho_K$  from a five-fold cross-validation (CV) process. As ABC, DA, and WDA are metrics that are derived from the averaged outcomes of the five-fold CV, it is not possible to calculate a standard deviation for these metrics. For clarity, the top-performing result for each model is indicated in red bold, while the second-highest performance is denoted in black bold. The evaluation is divided into two sections: Panel A focuses on the outcomes of vision-based deep learning models, whereas Panel B details the performance of LSTM and TCN models, assessed using various explainer methods including Gradient Shap (GS), Shapley Value Sampling (SVS), Saliency, Feature Ablation (FA), Integrated Gradients (IG), FIT, and DynaMask (DM). Here, Laxcat is featured both as an explainer method and as a model in its own.

## C GI FULL RESULTS

**CAFO Models.** ShuffleNet [51], ResNet [13], MLP-Mixer [39], and ViT (Vision Transformer) [8].

**Baseline Models.** LSTM (long short-term memory) [15] and TCN (temporal convolutional network) [23].

**Baseline Explainers.** Gradient Shap (GS) [26], Shapley Value Sampling (SVS) [6], Saliency [35], Feature Ablation (FA) [37], Integrated Gradients (IG) [36], DynaMask (DM) [7], FIT [40], and LAXCAT [17].

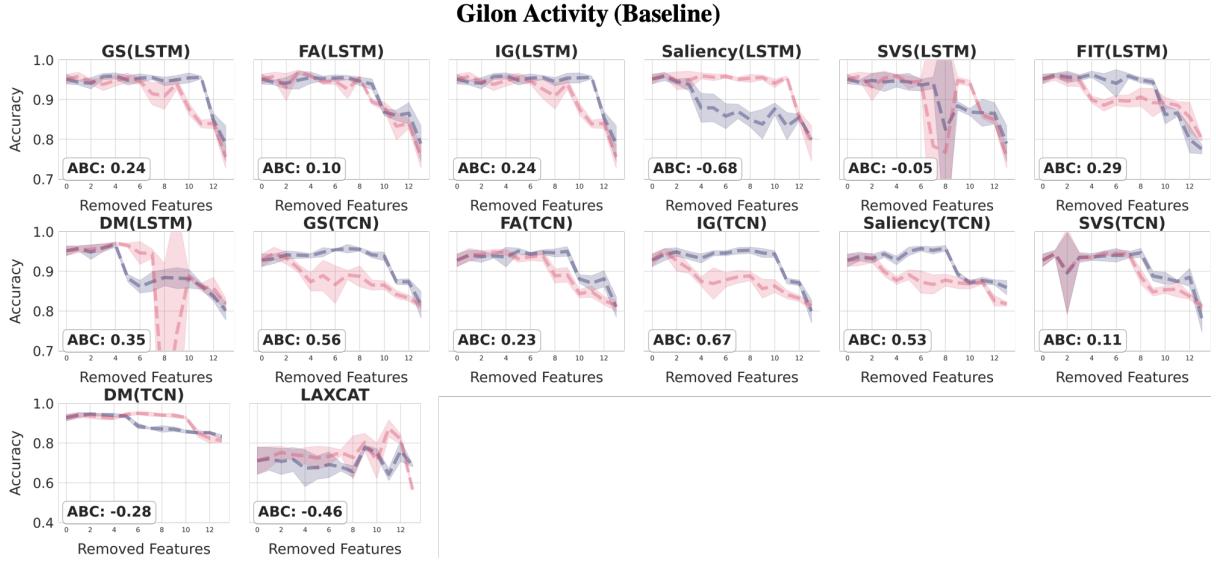


Figure 11: The ROAR plot on the Gilon Activity task for all baseline models.

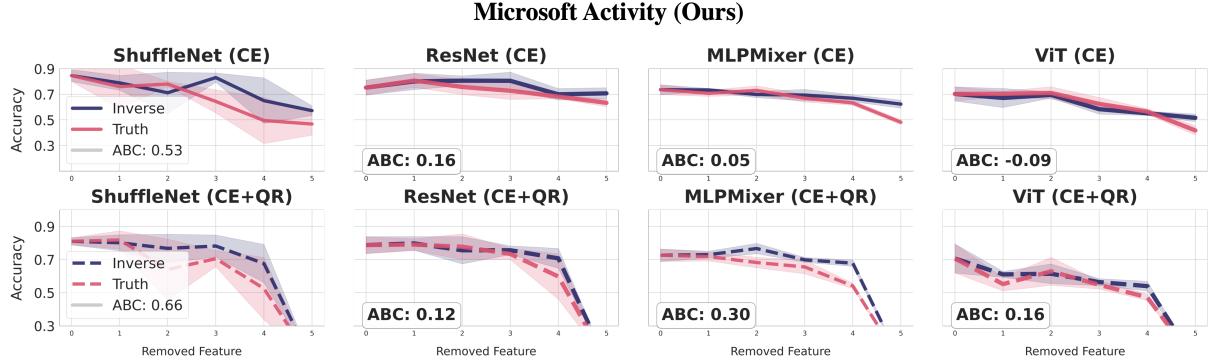


Figure 12: The ROAR plot on the Microsoft Activity task using CAFO. (Top row) Only CE(cross entropy) loss is applied. (Bottom row) CE+QR-Ortho Loss is applied. We observe that applying CE+QR-Ortho Loss leads to increased ABC (Area Between Curves) for all models except ResNet.

## C.1 GI Metric Calculation

**Area between curve (ABC).** Here, we assume that we have a training set which is split into five-fold cross-validation (CV), and a left-out test set. Each CV fold serves as a validation set against this test set. We start with a complete set of features, performing five-fold CV to derive feature rankings based on Global Importance (GI) scores from each run, which are then averaged to establish the final GI rankings. These features are arranged in descending order for the ‘truth’ rank and ascending order for the ‘inverse’ rank. Sequentially, we remove each feature from both training and test sets to assess model performance, hypothesizing that removing a critical feature decreases accuracy, while removing a non-essential feature doesn’t impact performance significantly. After eliminating all features, we calculate the ABC metric by measuring the area between the two resulting performance curves ( $f(x)$ ,  $g(x)$ ),

$$\text{ABC} \triangleq \int_a^b (f(x) - g(x)) dx \approx \frac{1}{2} \sum_{i=1}^{n-1} (f(x_i) - g(x_i) + f(x_{i+1}) - g(x_{i+1})) \Delta x_i$$

**Drop in accuracy (DA).** The drop in accuracy (DA) calculates the percentage decrease in accuracy of a model when  $K\%$  percentage of the most important features (denoted as  $K_{\text{acc}}$ ) identified by the model are dropped, compared to the base accuracy (when no features are dropped; denoted as  $\text{base\_acc}$ ). This is done by subtracting the accuracy after dropping  $K\%$  of the total features from the base accuracy and then dividing by the base accuracy. The result is multiplied by 100 to express it as a percentage. Here, we set  $K = 20\%$ . This metric helps in understanding the impact of removing the most important features on the model’s performance. Mathematically, DA is given as

$$\text{DA} \triangleq \left( \frac{\text{base\_acc} - K_{\text{acc}}}{\text{base\_acc}} \right) \times 100$$

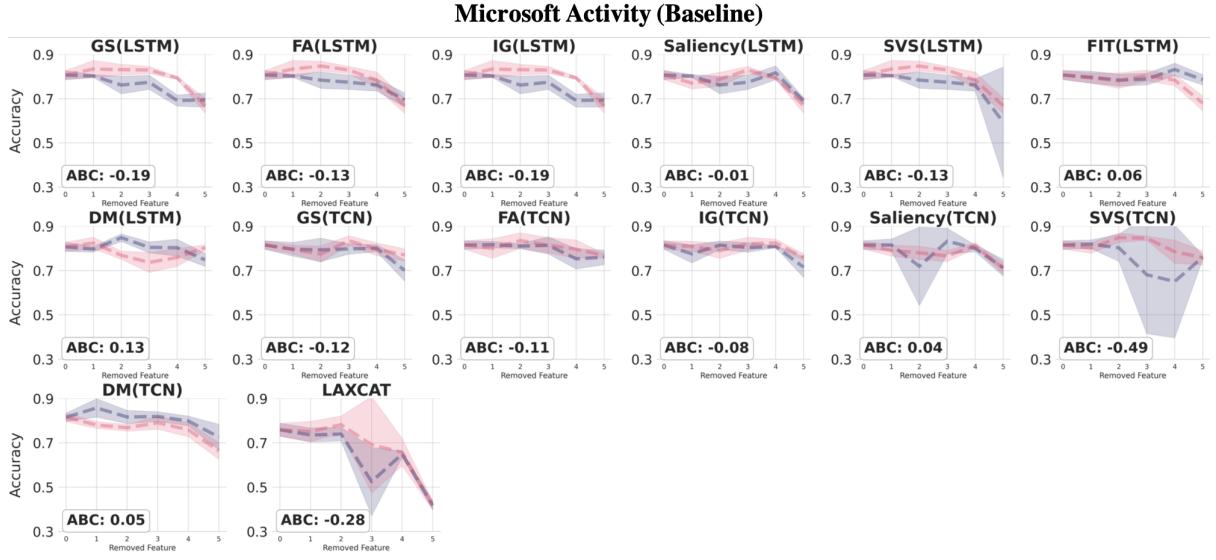


Figure 13: The ROAR plot on the Microsoft Activity task for all baseline models.

**Weighted drop in accuracy (WDA).** While DA provides a snapshot of the degraded model performance at a specific  $K$ , the weighted drop in accuracy (WDA) complements DA by considering the impact of dropping each feature individually and combines these impacts in a weighted manner. To compute WDA, consider  $D$  as the total number of features in the dataset. For each feature indexed by  $d$  (where  $d$  ranges from  $0, \dots, D - 1$ ), the weight  $w_d$  is calculated as follows:

$$w_d = \frac{1}{D - 1} \times (D - d - 1)$$

In this formula, the weight  $w_d$  decreases linearly with the index  $d$ , reflecting the diminishing marginal impact of removing additional features. For instance, given 14 features as in the Gilon task, the weights are given as 1.0, 0.923, 0.846, 0.769, 0.692, 0.615, 0.538, 0.461, 0.384, 0.307, 0.230, 0.153, 0.077. The WDA is calculated as below. Here,  $\text{base\_acc}$  is the base accuracy, and  $\text{d\_acc}$  represents the accuracy of model when  $d$  features are removed.

$$\text{WDA} \triangleq \sum_{d=0}^{D-1} (\text{base\_acc} - \text{d\_acc}) w_d$$

**Spearman  $\rho_s$  and Kendall  $\rho_k$ .** As in the area between curve (ABC) metric computation, we assume a training-set consisting of five-fold CV, with a left-out test set. Using each fold as a validation set, we can obtain five feature ranks. We calculate the pairwise spearman correlation and Kendall correlation, and state the average value.

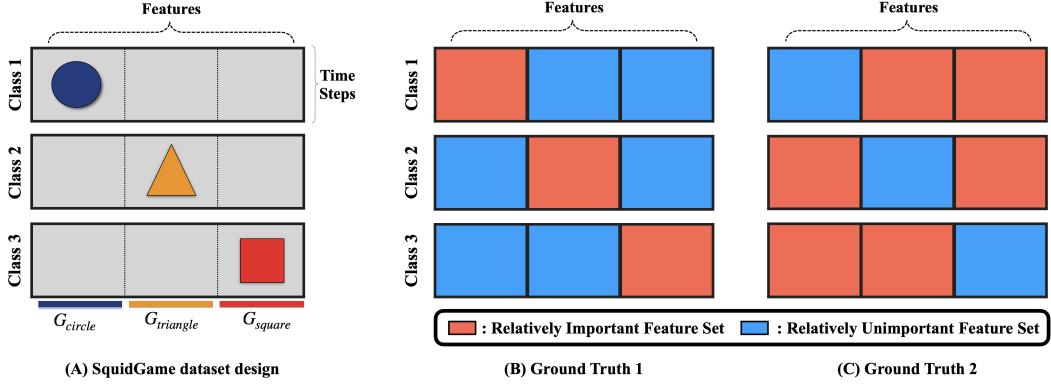
## D CWRI FULL RESULTS

Panel A. Performance Comparison of Various Models on CE and QR loss		SquidGame			WhichFinger		
Models	Methods	F1(↑)	Jaccard(↑)	IACC(↑)	F1(↑)	Jaccard(↑)	IACC(↑)
ShuffleNet	GS	0.689±0.08	0.531±0.09	0.649±0.08	0.590±0.09	0.424±0.10	0.596±0.10
	SVS	<b>0.811±0.14</b>	<b>0.699±0.18</b>	<b>0.853±0.11</b>	<b>0.635±0.03</b>	<b>0.466±0.04</b>	<b>0.636±0.04</b>
	Saliency	0.533±0.12	0.371±0.12	0.489±0.06	0.601±0.10	0.436±0.10	<b>0.644±0.06</b>
	FA	0.765±0.07	0.624±0.10	0.818±0.06	0.552±0.06	0.383±0.06	0.588±0.07
	IG	0.690±0.08	0.531±0.10	0.649±0.09	0.582±0.08	0.414±0.08	0.588±0.09
	CE	0.561±0.08	0.394±0.09	0.518±0.15	0.429±0.03	0.273±0.03	0.530±0.02
<b>CE+QR(Ours)</b>		<b>0.983±0.02</b>	<b>0.967±0.03</b>	<b>0.978±0.02</b>	<b>0.679±0.02</b>	<b>0.514±0.03</b>	0.606±0.08
ResNet	GS	0.747±0.05	0.598±0.06	0.693±0.08	0.630±0.04	0.462±0.05	0.636±0.05
	SVS	<b>0.758±0.11</b>	<b>0.621±0.15</b>	<b>0.800±0.10</b>	0.626±0.03	0.456±0.03	0.634±0.03
	Saliency	0.489±0.10	0.329±0.10	0.467±0.05	0.588±0.06	0.419±0.06	0.604±0.06
	FA	0.739±0.14	0.603±0.19	0.789±0.12	<b>0.670±0.01</b>	<b>0.503±0.01</b>	<b>0.696±0.02</b>
	IG	0.746±0.06	0.597±0.07	0.693±0.08	0.632±0.04	0.463±0.05	0.638±0.05
	CE	0.524±0.11	0.362±0.11	0.580±0.13	0.357±0.20	0.232±0.15	0.530±0.03
<b>CE+QR(Ours)</b>		<b>0.987±0.01</b>	<b>0.974±0.01</b>	<b>0.982±0.01</b>	<b>0.724±0.05</b>	<b>0.570±0.06</b>	<b>0.698±0.07</b>
Mixer	GS	0.929±0.09	0.877±0.15	0.918±0.10	<b>0.859±0.01</b>	<b>0.753±0.02</b>	<b>0.862±0.01</b>
	SVS	<b>0.994±0.01</b>	<b>0.988±0.03</b>	<b>0.996±0.01</b>	0.778±0.07	0.641±0.09	0.784±0.08
	Saliency	0.514±0.08	0.349±0.07	0.453±0.04	0.726±0.03	0.571±0.03	0.744±0.02
	FA	<b>0.987±0.02</b>	<b>0.975±0.03</b>	<b>0.991±0.01</b>	0.732±0.10	0.586±0.13	0.752±0.09
	IG	0.929±0.09	0.877±0.15	0.918±0.10	<b>0.861±0.01</b>	<b>0.756±0.02</b>	<b>0.864±0.01</b>
	CE	0.596±0.07	0.427±0.08	0.736±0.04	0.496±0.08	0.333±0.07	0.600±0.05
<b>CE+QR(Ours)</b>		0.949±0.03	0.904±0.05	0.933±0.03	0.709±0.05	0.551±0.06	0.660±0.08
ViT	GS	0.861±0.08	0.763±0.13	0.842±0.09	0.685±0.09	0.527±0.10	0.690±0.09
	SVS	<b>0.883±0.11</b>	<b>0.803±0.17</b>	<b>0.902±0.09</b>	<b>0.811±0.07</b>	<b>0.686±0.09</b>	<b>0.812±0.07</b>
	Saliency	0.526±0.08	0.359±0.07	0.484±0.05	<b>0.722±0.08</b>	<b>0.570±0.10</b>	<b>0.750±0.05</b>
	FA	0.779±0.04	0.640±0.06	0.809±0.05	0.674±0.07	0.512±0.07	0.684±0.06
	IG	0.861±0.08	0.763±0.13	0.842±0.09	0.702±0.06	0.544±0.07	0.710±0.07
	CE	0.700±0.10	0.546±0.12	0.796±0.08	0.520±0.04	0.352±0.04	0.536±0.02
<b>CE+QR(Ours)</b>		<b>0.925±0.04</b>	<b>0.863±0.07</b>	<b>0.898±0.05</b>	0.531±0.06	0.363±0.05	0.572±0.02
Panel B. Performance Evaluation Model with Different Interpretability Methods		SquidGame			WhichFinger		
Models	Methods	F1(↑)	Jaccard(↑)	IACC(↑)	F1(↑)	Jaccard(↑)	IACC(↑)
LSTM	GS	0.692±0.05	0.531±0.06	0.756±0.04	<b>0.709±0.03</b>	0.549±0.04	0.732±0.03
	SVS	0.719±0.05	0.564±0.07	0.778±0.04	0.527±0.04	0.358±0.04	0.584±0.03
	Saliency	0.475±0.04	0.312±0.03	0.484±0.08	0.615±0.05	0.445±0.05	0.636±0.03
	FA	0.516±0.03	0.348±0.03	0.647±0.01	0.590±0.06	0.421±0.06	0.624±0.05
	IG	0.691±0.05	0.529±0.05	0.758±0.03	0.707±0.04	0.548±0.05	<b>0.734±0.03</b>
	FIT	0.622±0.10	0.457±0.10	0.533±0.05	0.625±0.07	0.458±0.07	0.608±0.09
<b>DM</b>		<b>0.753±0.04</b>	<b>0.605±0.05</b>	<b>0.798±0.04</b>	0.704±0.13	<b>0.556±0.15</b>	0.676±0.13
TCN	GS	0.597±0.04	0.426±0.04	0.709±0.03	<b>0.685±0.06</b>	<b>0.523±0.07</b>	<b>0.706±0.06</b>
	SVS	0.659±0.02	0.491±0.02	0.751±0.02	0.580±0.04	0.409±0.04	0.606±0.04
	Saliency	0.444±0.00	0.286±0.00	0.444±0.00	0.523±0.04	0.355±0.04	0.546±0.03
	FA	<b>0.737±0.01</b>	<b>0.584±0.02</b>	<b>0.796±0.01</b>	0.607±0.03	0.436±0.04	0.632±0.03
	IG	0.591±0.03	0.420±0.03	0.702±0.02	0.667±0.07	0.504±0.08	0.692±0.06
	DM	0.493±0.05	0.328±0.05	0.553±0.12	0.594±0.04	0.423±0.04	0.602±0.05
None	Random	0.400±0.06	0.252±0.05	0.501±0.05	0.498±0.06	0.334±0.05	0.499±0.05
	LAXCAT	<b>0.568±0.05</b>	<b>0.398±0.05</b>	<b>0.518±0.07</b>	<b>0.545±0.07</b>	<b>0.377±0.07</b>	<b>0.570±0.06</b>

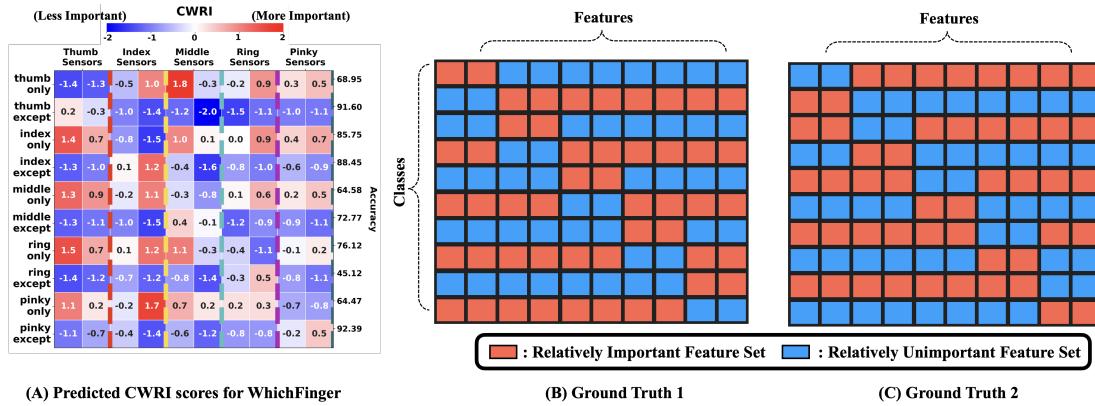
Table 5: Performance Evaluation of CWRI Metrics. This table presents the evaluation of CWRI metrics using F1 Score, Jaccard, and Accuracy (IACC). We use the term IACC to differentiate with model accuracy. The metrics are measured based on the predicted class-wise importance of each explainer methods against the established ground truth class-wise importance. All experiments were conducted with a five-fold cross validation.

## E EVALUATION OF CWRI METRICS

During the evaluation of the CWRI measure, we encountered a dilemma regarding the establishment of ground truth for each class. While we can identify the exact features relevant to each class, the model is not obligated to base its predictions on these features exclusively. For instance, flexing the thumb (class 1-Thumb only; Fig. 18)) implies that the remaining four fingers remain still, which also serves as a useful identifier for the class. This deviates from spurious correlation, as the association between features and class predictions is not merely coincidental.



**Figure 14: Ground Truth Design of SquidGame.** (A) The design of SquidGame task. (B) & (C). Two distinct ground truth label types are established for the SquidGame task. In this context, the red feature sets signify more critical or influential features, whereas the blue feature sets represent those that are comparatively less important or impactful.



**Figure 15: An example of a predicted CWRI score for the WhichFinger task.** (A) The CWRI score is predicted using MLP-Mixer. (B) & (C). Two distinct ground truth label types are established for the WhichFinger task. In this context, (A) follows the ground truth label of (C) rather than (B).

In order to quantitatively assess the CWRI score, we generated two distinct sets of ground truth labels, as illustrated in Fig. 14 and Fig. 15. Each of these ground truth label sets was utilized to evaluate the predicted CWRI score. We selected the ground truth label that produced a higher F1 score, as it represented the only feasible approach for evaluating the CWRI score. This decision was grounded in the understanding that the model may have considered these feature sets crucial during that particular run.

Throughout our experimentation, we observed that a model adheres to a consistent ground truth label, notwithstanding variations in seed and cross-validation folds. Nevertheless, the choice of ground truth may shift between different architectures and the application of QR-Ortho Loss. Determining which ground truth label the model will adopt remains an open question and constitutes a challenge that we intend to investigate in future work.

## F DATASETS

Dataset	#Samples	Length	#Users	# Features	# Class
Gilon	47,647	160	72	14	7
MS	14,201	200	93	6	10
SquidGame	54,000	32	-	30	3
WhichFinger	18,010	120	19	10	10

Table 6: Statistics of the datasets used in our experiment

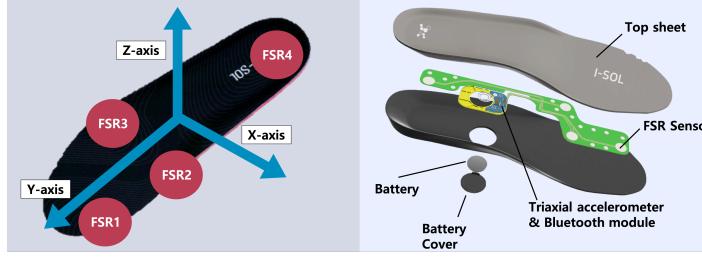


Figure 16: Smart insole used to collect the Gilon Activity task. Used with permission [22].

**Gilon Activity** [22]. This dataset comprises smart insole measurements gathered from 72 distinct users in South Korea. The smart insole records 14 different sensor measures, split evenly between the left and right feet. These metrics include three-dimensional accelerometer readings (X, Y, Z) and data from four force-sensitive resistors (FSRs). These participants engaged in seven specific activities while equipped with the smart insole. The activities include standing still (0), walking on the ground (1), walking on a treadmill at a constant speed (2), running on a treadmill at a constant speed (3), performing lunges (4), squats (5), and jumping jacks (6). The original study established predefined train/test splits based on the users. As a result, the training set is composed of data from 50 users, amounting to 33,368 samples, while the test set consists of data from 22 users, totaling 14,279 samples. Further, within the training set, the data is divided into five-folds. Each fold uses a different group of users as a validation set, ensuring a balanced evaluation across the dataset. The dataset was segmented into windows size of 160 with an overlapping length of 120. The five-fold cross validation in our research was conducted by using each validation split with the fixed test set. The dataset (<https://github.com/Gilon-Inc/GILON-MULTITASK-DATASET>) can be accessed by request. The dataset is protected under CC-BY-NC-ND-4.0 license.

**MS Activity** [28]. The original dataset (exercise\_data\_50.0000\_singleonly.mat) contains 114 users performing 72 distinct gym exercises, recorded in 4,686 sessions with wearable arm sensors. The armband includes a 3-axis accelerometer (X, Y, Z) and gyroscope (X, Y, Z). To address the imbalance in the number of sessions across activities, we refined the dataset by selecting exercises with session counts ranging between 25 and 70, based on an empirical analysis of the data distribution. We then focused on activities featuring similar physical motions to effectively showcase our class-wise importance metric. This resulted in a curated set of activities, including Bicep Curl (0), Biceps Curl with Band (1), Jump Rope (2), Plank (3), Pushups (4), Squat (5), Squat with Hands Behind Head (6), Squat Jump (7), Walk (8), and Walking Lunge (9). The data was segmented into non-overlapping windows of 200 samples each. The final processed dataset comprises training (65 users; 10,892 samples) and testing subsets (28 users; 3,309 samples), with the training data further divided into five-fold cross validation, following the approach used in the Gilon dataset. The raw dataset can be accessed in (<https://github.com/microsoft/Exercise-Recognition-from-Wearable-Sensors>).

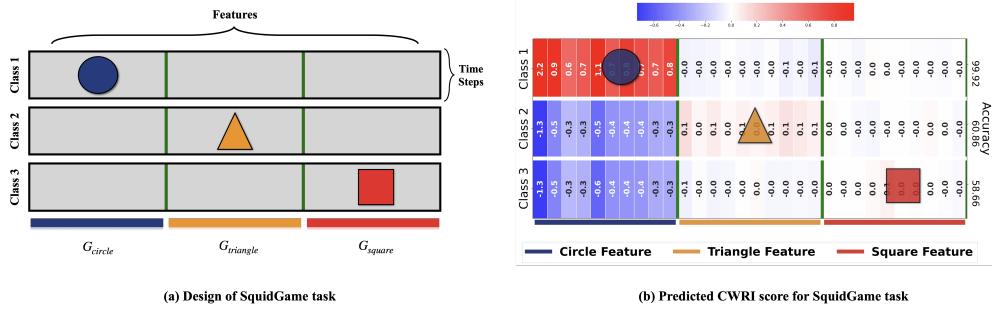


Figure 17: (a) The design of SquidGame task. (b) The predicted CWRI score for the SquidGame.

**SquidGame.** The SquidGame task is a synthetic 3-class classification dataset designed to validate the efficacy of CWRI scores. This task features distinct, non-overlapping feature sets, each representing a specific class. Initially, all features are populated with random Gaussian noise. These time series instances are then treated as  $T \times D$  images, and an empty mask is created with circle, triangle, and square shapes. Each mask is filled with characteristic time signals from the sinusoidal series, with the length, size, and center coordinates generated randomly for each instance. Here,  $T = 32$  and  $D = 30$ . Feature indices are divided into three groups:  $G_{circle} = \{1, \dots, 10\}$ ,  $G_{triangle} = \{11, \dots, 20\}$ , and  $G_{square} = \{21, \dots, 30\}$ .  $G_{circle}$  indicates that feature indices 1 to 10 contain a circular mask with characteristic time signals representing the first class. The complement of the circle masks in indices 1 to 10 and the remaining indices (11 to 30) are filled with Gaussian noise for the first class. This scheme is also applied to the second and third classes. As a result, only the first class contains characteristic time signals in feature indices 1 to 10, while the other classes have random Gaussian noises. The dataset's name was inspired by the popular Netflix series "Squid Game."

**WhichFinger.** The WhichFinger task utilizes a real-world smart glove dataset that we have gathered in order to verify the effectiveness of the CWRI scores. In this task, we set  $T = 120$  and  $D = 10$ . A comprehensive discussion of the dataset's details can be found in Appendix G, which is dedicated to this particular topic.

## G WHICHFINGER DATASET

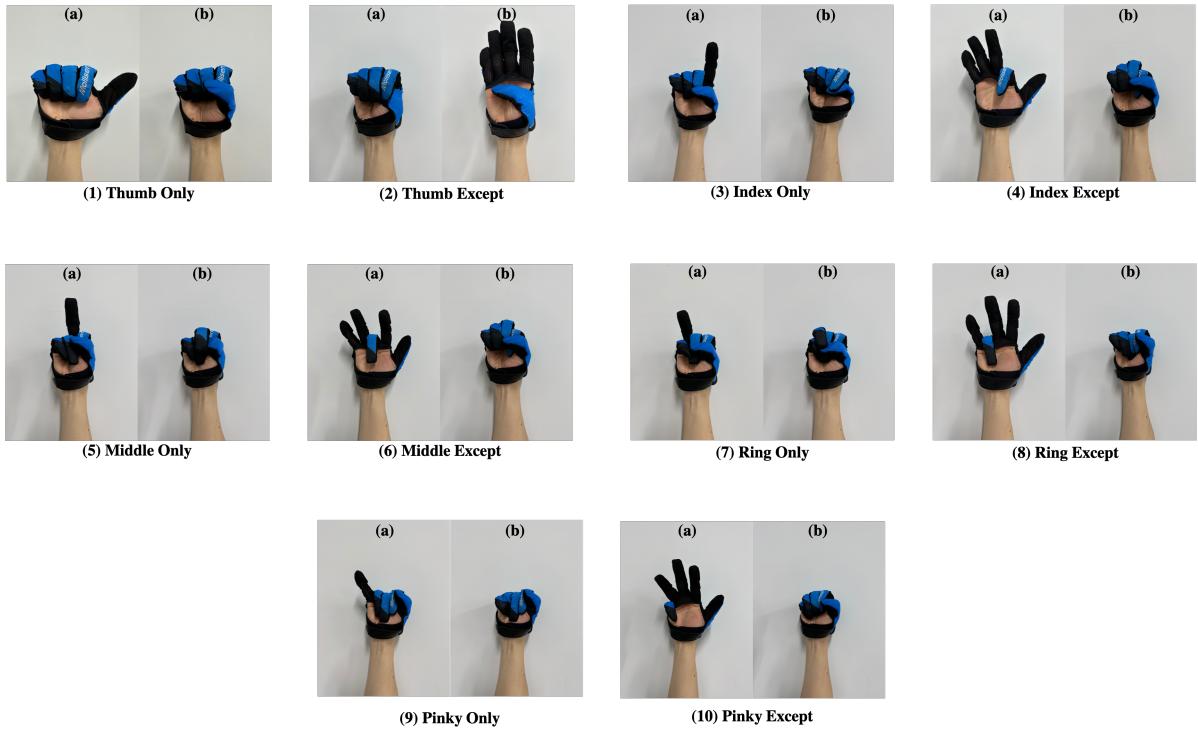


Figure 18: The figure illustrates the detailed finger movement for each class. (a) and (b) for each class are repeated for more than one minute.

The WhichFinger Dataset is a multivariate time series (MTS) dataset, designed for eXplainable Artificial Intelligence (XAI) applications. This dataset offers comprehensive information on the data collection process for each class, as well as the features relevant to specific classes, which facilitates the validation of the CWRI measure. We created this dataset because, to the best of our knowledge, no public MTS datasets met the following three criteria: (1) strong prior knowledge or information regarding each feature's contribution to specific classes, (2) a sufficient number of classes ( $C \geq 2$ ) and features ( $D \geq 2$ ), and (3) an adequate number of samples ( $N \geq 1,000$ ). In this section, we describe the detailed data collection process and the preprocessing steps involved in the creation of the dataset.

### G.1 Recruitment of Subjects

We recruited 20 volunteer participants through flyer advertisements and provided each participant with a fee of approximately \$7 (USD) for their participation in the experiment. The experiment was approved by the UNIST Institutional Review Board (IRB) (UNISTIRB-23-008-A).

## G.2 Smart Glove

**Sensors.** The smart glove is equipped with an integrated soft sensor that captures precise finger movements. Each finger contains two sensors, resulting in a total of 10 sensor recordings. These sensors detect changes in resistance as the fingers flex and relax. The data is recorded at approximately 66.7 Hz.

## G.3 Data Collection Process

**Experiment Condition.** Participants were instructed to wear the smart glove on their right hand. The supervisor sat adjacent to each participant, closely monitoring the data collection process during the entire experiment. The supervisor also provided a thorough explanation of the experimental procedures and specific actions required from the participants.

**Experiment Procedure.** A total of 10 unique finger movements, corresponding to 10 different categories, were recorded. The finger movements captured included: (1) Thumb only, (2) Thumb except, (3) Index only, (4) Index except, (5) Middle only, (6) Middle except, (7) Ring only, (8) Ring except, (9) Pinky only, and (10) Pinky except. These specific finger movements are depicted in detail in Fig. 18. For each category, participants carried out the actions for one minute. Upon completing every two categories, a one-minute break was provided, or extended if requested by the participant. The entire data recording process took approximately 20 minutes for each participant to complete.

As the experiment aimed to identify specific features contributing to each class, it was necessary to minimize undesired finger movements. However, due to the interconnected nature of finger muscles, it is impossible to keep the other fingers completely still while moving one finger. Therefore, participants were encouraged to use their left hand to support and constrain the fingers that needed to remain stationary. When required, duct tape was also employed to reinforce these constraints. Despite these measures to reduce extraneous movements, they could not be entirely eliminated, leading to sensor measurements that still captured some unintended finger motions.

**Data Exclusion.** One participant had to be excluded from the study as a result of data records being lost during the data collection process, resulting in 19 user information in the final dataset.

## G.4 Data Preparation for Model Training

**Data Shape.** We dropped the first and last 150 recordings (approximately 2 seconds) from each class measurement. The resulting dataset for each class contains on average  $4190 \pm 268$  recordings (rows) for each finger activity.

**Data Preprocessing.** We adopted a similar preprocessing approach for the recorded data as described in the work of [22]. The multivariate time series (MTS) data was segmented into consecutive 2-second time windows (120 rows) with a 75% overlap (80 rows) between windows. This segmentation was performed within each class and participant. Consequently, for each class, we obtained an average of 102 MTS instances, calculated as  $(\frac{4190 - 120}{120 - 80} + 1) = 102$  for each class recording.

**Data Split.** Out of the 19 participants, four were randomly chosen for the test set. The remaining 15 participants were divided into five groups, with each group serving as a validation set. This resulted in a 5-fold cross-validation (CV) setup.

## G.5 License for Data Use

We apply the Creative Commons Attribution-NonCommercial 4.0 International License.

## H IMPLEMENTATION DETAILS

### H.1 Model Training

In our experiments, we fixed all random seeds to 42 and used the AdamW [4] optimizer with learning rates of 0.002 for Gilon and SquidGame, 0.005 for WhichFinger, and 0.01 for Microsoft Activity. Learning rate scheduling was not used.

### H.2 CAFO

Throughout our experiments, we set the expansion filter number  $\gamma = 3$  in the DepCA module. The task-specific hyperparameter  $\lambda$  in QR-Ortho loss requires a grid search, which we perform for  $\lambda \in \{0.1, 0.2, 0.5, 1.0\}$  in each task, selecting based on the best validation accuracy. Consequently, we used  $\lambda = 0.5$  for Gilon,  $\lambda = 0.1$  for Microsoft, and  $\lambda = 1.0$  for both SquidGame and WhichFinger.

### H.3 Deep Architectures

**ShuffleNet.** We utilized grouped convolution, setting the group number to 3, and configuring the output channels as follows:  $[D, 24, 120, 240, 480]$ . Here,  $D$  is the number of input feature channels.

**ResNet.** We used the original implementation of the ResNet but with 9 layers.

**MLP-Mixer.** For MLP-Mixer, we used the implementation from <https://github.com/lucidrains/mlp-mixer-pytorch>. We set the number of Mixer layers to 3 and the feed-forward layer dimension to 256.

**Vision Transformer.** We used the implementation from <https://github.com/lucidrains/vit-pytorch>. We configured the Transformer architecture with three blocks, each containing a multi-head attention mechanism with three heads. Additionally, the dimensions of all feed-forward layers were set to 256. For both MLP-Mixer and Vision Transformer, the patch size was set to 1/10th of the original input image for Gilon, Microsoft, and WhichFinger, while 1/8th was used for SquidGame due to its smaller image size.

### H.4 Baselines

All explainers produce attributions for each time step, yielding an attribution size of  $\mathbb{R}^{T \times D}$ . We average the attributions feature-wise and utilize these values to compute metrics in our study.

**FIT.** We trained a feature generator using a gated recurrent unit (GRU) as in the original implementation of [40]. The hidden dimension was set to 256, and 1 layer was used for GRU. We generated ten monte carlo samples for each time stamp.

**DynaMask.** We used the original implementation of [7]. However, due to the immense computational complexity in optimizing the perturbation mask, we reduced the optimization step to 50 for each instance.

**LAXCAT.** We set a task specific kernel and stride size for 1-D convolution. As the original paper did not release the code, we used a third party implementation from <https://github.com/Shuheng-Li/UniTS-Sensory-Time-Series-Classification>.

**Others.** We used the implementations from the CAPTUM library [2].

### H.5 Hardware and Software

We conducted experiments on three GPU types: Nvidia TITAN RTX-24GB, Tesla V100-PCIE-32GB, and RTX A6000-48GB, utilizing PyTorch 1.8.1 and PyTorch-Lightning 1.6.5.

## I QR FEATURE ORTHOGONALITY REGULARIZATION ALGORITHM

### Algorithm 1 QR-Ortho Loss Algorithm

---

```

1: Input: Training data  $\mathbf{K} = \{(\mathbf{X}^{(i)}, y^{(i)})\}_{i=1}^N$ , Instances in class  $c$  denotes  $\mathbf{K}_c$ , Batch size  $B$ , Class set  $C$ , The matrix of class prototypes
    $\bar{\mathbf{A}} \in \mathbb{R}^{C \times D}$ , Sum of the number of off-diagonal elements in  $\mathbf{R}$  denoted as  $r$ , Strength of orthogonality  $\lambda$ , Depthwise Channel Attention
   denoted as DepCA, Image encoder denoted as RP
2: For mini-batch  $\{(\mathbf{X}^{(i)}, y^{(i)})\}_{i=1}^B \subseteq \mathbf{K}$  do                                ▷ Orthogonality phase for training data
3:    $\mathcal{X}^{(i)} = \mathbf{RP}(\mathbf{X}^{(i)})$                                          ▷ Encode raw time series using RP
4:    $\mathbf{a}^{(i)} = \mathbf{DepCA}(\mathcal{X}^{(i)})$                                          ▷ Calculate channel attention score
5:    $\bar{\mathbf{a}}_c = \frac{1}{|\mathbf{K}_c|} \sum_{i \in \mathbf{K}_c} \mathbf{a}^{(i)}$                                          ▷ Calculate class prototypes for channel attention score
6:    $\bar{\mathbf{A}} = [\bar{\mathbf{a}}_i]_{i=1}^C$                                          ▷ Stack  $\bar{\mathbf{a}}_c$  in a row-wise manner
7:   Decompo $\bar{\mathbf{A}} = \mathbf{QR}$                                          ▷ QR decomposition
8:    $\mathcal{L}_{\text{QR}} = \frac{1}{r} \sum_{i < j} |\mathbf{R}_{ij}|$                                          ▷ Compute QR-Ortho Loss
9:    $\mathcal{L}_{\text{Total}} = \mathcal{L}_{\text{CE}} + \lambda \mathcal{L}_{\text{QR}}$                                          ▷ Total loss is the sum of cross-entropy (CE) Loss with QR-Ortho Loss
10: end for

```

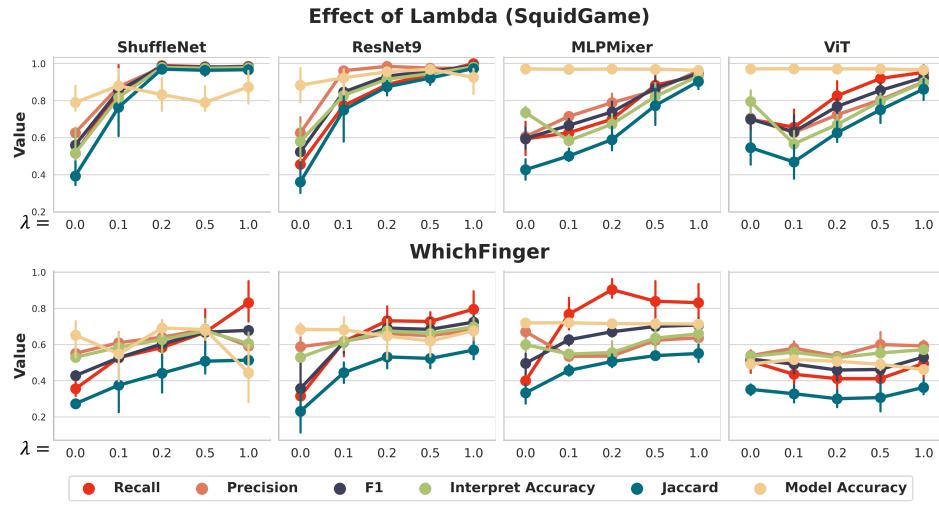
---

## J PSEUDO VARIABLES

We generate random pseudo signals from three different time series process using the TimeSynth [27] library. The pseudo signal was generated for each MTS instance. We used the default setting from the library.

- (1) **WhiteNoise**. Gaussian noise with a zero mean and a 0.3 standard deviation
- (2) **Sinusoidal**. Sine waves with an amplitude of one and a frequency of 0.25
- (3) **Gaussian Process**. Matern kernel was used with nu=1.5

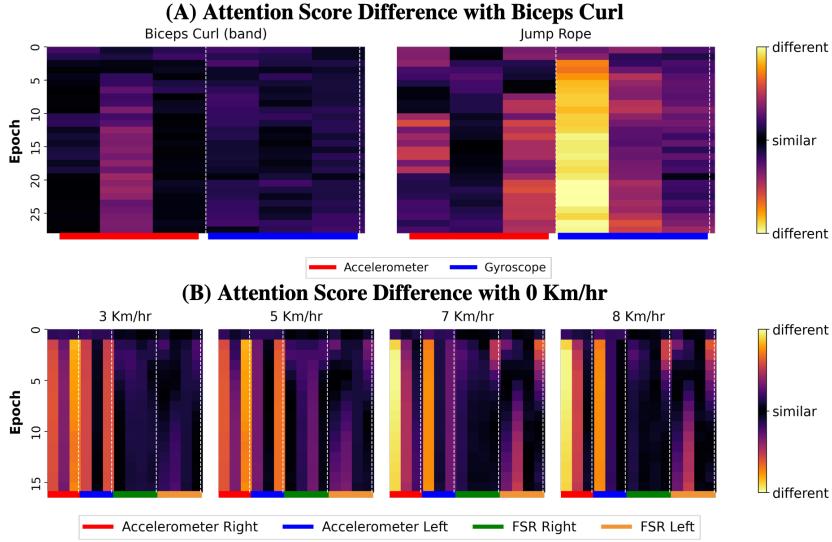
## K EFFECT OF $\lambda$



**Figure 19:** (Top row) CWRI related metrics are visualized for the SquidGame task. Here,  $\lambda = 0$  is equivalent to using the CE loss only. We observe a general upward trend as  $\lambda$  is increased. (Bottom row) The general upward trend observed in the SquidGame is weaker in the WhichFinger.

We conduct a grid search of  $\lambda \in \{0, 0.1, 0.2, 0.5, 1.0\}$ , a task-specific hyperparameter that controls the strength of QR-Ortho Loss (Eq. (4)), for SquidGame and WhichFinger tasks. As shown in Fig. 19, metrics related to the evaluation of CWRI generally improve with larger  $\lambda$ .

However, excessive orthogonality can decrease model accuracy, as exemplified by the ShuffleNet on the WhichFinger task, which shows that identifying an optimal  $\lambda$  is a critical consideration.



**Figure 20: Visualization of Attention Score Difference. The Y-axis is the epoch and the X-axis corresponds to the feature. A greater disparity in attention scores is indicated by a more intense color brightness.**

## L ALIGNMENT WITH DOMAIN KNOWLEDGE

In the domains of vision and natural language processing, XAI methodologies have aimed to demonstrate the alignment of their explanations with human intuition. This is typically achieved by superimposing attention maps onto images highlighting class-related features [7] or visualizing the intensity of attention signals connecting pairs of related words [9]. Our methodology adopts a parallel strategy but by visualizing the differences between attention vectors derived from closely related classes. By doing so, we show that the explanations made with CAFO align with domain knowledge but also provide a deeper insight into the subtleties differentiating one class from another.

We employ real-world datasets, namely Gilon and MS, to illustrate this methodology. For visual analysis, we gather class attention prototypes  $\tilde{\mathbf{a}}_c \in \mathbb{R}^D$  (Eq. (2)) for each epoch  $E$ , forming a class attention matrix  $\tilde{\mathbf{A}}_c \in \mathbb{R}^{E \times D}$ . Here, we visualize  $|\tilde{\mathbf{A}}_c - \tilde{\mathbf{A}}_{c'}| (c \neq c')$ . Our first validation, showcased in Fig. 20-A, tests the hypothesis that *similar activities should yield similar attention scores*. The comparison between (a) Bicep Curl (w/o band) and Bicep Curl with Band, and (b) Bicep Curl (w/o band) and Jump Rope from the MS data, supports this. The attention scores of similar activities (Bicep Curl (w/o band) and Bicep Curl with Band) show minor divergence, whereas there is a large difference in the attention scores when comparing Bicep Curl to Jump Rope.

Subsequently, we verify that *accelerometer features should be the main feature in differentiating varying speed ranges* in Fig. 20-B. For this experiment, we utilized a downstream regression task from the Gilon, where the objective is to regress the speed at which the users are moving on a treadmill. Here, we train with mean square error loss and without QR-Ortho regularizer on the Gilon dataset. We visualize the difference between the 0km/hr (stationary position) and the remaining speed labels (3, 5, 7, and 8 km/hr). We first observe that the attention divergence is large for accelerometer features across all speed ranges. Moreover, this divergence escalates with larger speed difference, denoted by the brightness in the heatmap. This change in color difference highlights the role of accelerometer features in differentiating speed ranges, a finding that is consistent with established knowledge.

## REFERENCES

- [1] Niima Es-Sakali, Moha Cherkaoui, Mohamed Oualid Mghazli, and Zakaria Naimi. 2022. Review of predictive maintenance algorithms applied to HVAC systems. *Energy Reports* 8 (2022), 1003–1012.
- [2] Narine Kokhlikyan, Vivek Miglani, Miguel Martin, Edward Wang, Bilal Alsallakh, Jonathan Reynolds, Alexander Melnikov, Natalia Kliushkina, Carlos Araya, Siqi Yan, et al. 2020. Captum: A unified and generic model interpretability library for pytorch. *arXiv preprint arXiv:2009.07896* (2020).
- [3] Jianguan Liu, Daliang Shi, Guannan Li, Yi Xie, Kuining Li, Bin Liu, and Zhipeng Ru. 2020. Data-driven and association rule mining-based fault diagnosis and action mechanism analysis for building chillers. *Energy and Buildings* 216 (2020), 109557.
- [4] Ilya Loshchilov and Frank Hutter. 2017. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101* (2017).
- [5] Setu Madhavi Namburu, Mohammad S Azam, Jianhui Luo, Kihoon Choi, and Krishna R Pattipati. 2007. Data-driven modeling, fault diagnosis and optimal sensor selection for HVAC chillers. *IEEE transactions on automation science and engineering* 4, 3 (2007), 469–473.
- [6] Jeffrey Schein and Steven T Bushby. 2006. A hierarchical rule-based fault detection and diagnostic method for HVAC systems. *Hvac&r Research* 12, 1 (2006), 111–125.
- [7] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. 2017. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, 618–626.

- [8] Fan Tang. 2010. *HVAC system modeling and optimization: a data-mining approach*. Ph.D. Dissertation. The University of Iowa.
- [9] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems* 30 (2017).