This is the author accepted manuscript version of: Y. Tian et al., "Toward Learning Model-Agnostic Explanations for Deep Learning-Based Signal Modulation Classifiers," in IEEE Transactions on Reliability
The final published version can be found at: https://doi.org/10.1109/TR.2024.3367780

Towards Learning Model-agnostic Explanations for Deep Learning-based Signal Modulation Classifiers

Yunzhe Tian, Dongyue Xu, Endong Tong, *Member, IEEE*, Rui Sun, Kang Chen, Yike Li, Thar Baker, *Senior Member, IEEE*, Wenjia Niu, and Jiqiang Liu, *Senior Member, IEEE*

***Abstract*—Recent advances in deep learning (DL) have brought tremendous gains in signal modulation classification. However, DL-based classifiers lack transparency and interpretability, which raises concern about model's reliability and hinders the wide deployment in real-word applications. While explainable methods have recently emerged, little has been done to explain the DL-based signal modulation classifiers. In this work, we propose a novel model-agnostic explainer, MASE, which provides explanations for the predictions of black-box modulation classifiers. With the subsequence-based signal interpretable representation and in-distribution local signal sampling, MASE learns a local linear surrogate model to derive a class activation vector which assigns importance values to the timesteps of signal instance. Besides, the constellation-based explanation visualization is adopted to spotlight the important signal features relevant to model prediction. We furthermore propose the first generic quantitative explanation evaluation framework for signal modulation classification to automatically measure the faithfulness, sensitivity, robustness and efficiency of explanations. Extensive experiments are conducted on two real-world datasets with four black-box signal modulation classifiers. The quantitative results indicate MASE outperforms two state-of-the-art methods with 44.7% improvement in faithfulness, 30.6% improvement in robustness and 44.1% decrease in sensitivity. Through qualitative visualizations, we further demonstrate the explanations of MASE are more human interpretable and provide better understanding into the reliability of black-box model decisions.**

***Index Terms*—Explainable AI, model reliability, deep learning, black-box model, interpretability, modulation classification.**

I. Introduction

SIGNAL modulation classification, essential in both civilian and military wireless systems, identifies the modulation type of received radio signals to understand their communication schema [1]–[3]. As a typical pattern recognition problem, considerable efforts have been put into this research area. In prior work, modulation classification has been achieved based on carefully hand-crafted feature extractors [4]–[6], largely relying on domain knowledge. Recently, with the prosperity of deep learning (DL), DL-based modulation classification has become increasingly prevalent, achieving remarkable performance improvements [7]–[11]. Without manual feature extraction, DL-based signal modulation classifiers, fed with the raw radio signals, enable automatic learning the higher-level information hidden in the data and performing classification in an end-to-end fashion.

Despite the effectiveness of DL-based modulation classification, the inherent opaqueness aggravates the untrustworthiness and unreliability of models, posing a significant hindrance to their large-scale deployment [12]. Compared to conventional feature-based modulation classification, the complex internal structure and decision process of DL-based classifiers are not human-understandable, causing domain experts may not trust the model predictions [13]. To address this, *explainability* has been recognized as a critical tool to build the understandable, trustworthy and reliable interaction between human and models, and the field of explainable artificial intelligence (XAI) thrived [14]–[16]. However, existing XAI methods, such as EVA [17] and Ensemble XAI [18] are mainly for computer vision tasks, with limited work in signal modulation classification. Huang *et al.* [19] and Chen *et al.* [20] make the initial attempts to increase the transparency of DL-based signal modulation classification by visualizing the signal features which are extracted and chosen for classification.

There are two issues that exist in the aforementioned literature. Firstly, they visualize the DL-based signal modulation classifiers with model-specific XAI methods (i.e., Grad-CAM [21] and MASK [22]), which requires the white-box access to model internals like model structure and gradients. However, due to the business and military constraints, the access to modulation classifiers in need of explanations is typically limited to an inference API, providing only prediction scores [23]. Thus, it is urgent to develop a model-agnostic XAI method which is general and flexible for any black-box signal modulation classifier. Second, in the existing work, the effectiveness of XAI method is assessed subjectively via visualization, with explanations evaluated based on author's domain knowledge of relevant signal features. However, for practical applications, this assessment is labor-intensive, time-consuming and unreliable. Therefore, an objective, quantitative, and comprehensive evaluation framework becomes necessary to assess the explanations for signal modulation classification.

The recent survey [12] reveals significant advancements in model-agnostic XAI. Prevalent methods like LIME [24], SHAP [25], and Anchors [26] typically begin by converting

This work is supported by the Fundamental Research Funds for the Central Universities of China under Grant No. 2023YJS031, 2023JBMC055, and 2023JBZY036, the National Natural Science Foundation of China under Grant No. 62372021, the Hebei Natural Science Foundation under Grant No. F2023105005, the Central Funds Guiding the Local Science and Technology Development under Grant No. 236Z0806G, and the 'Top the List and Assume Leadership' project in Shijiazhuang, Hebei Province, China. *(Corresponding authors: Endong Tong; Wenjia Niu.)*

Yunzhe Tian, Dongyue Xu, Endong Tong, Rui Sun, Kang Chen, Yike Li, Wenjia Niu and Jiqiang Liu are with the Beijing Key Laboratory of Security and Privacy in Intelligent Transportation, Beijing Jiaotong University, Beijing 100044, China (e-mail: {tianyunzhe, dongyuexu, edtong, ruisun, chenkang, yikeli, niuwj, jqliu}@bjtu.edu.cn.)

Thar Baker is with the School of Architecture, Technology and Engineering, University of Brighton, Brighton, BN2 4GJ, UK (email: t.shamsa@brighton.ac.uk).

the raw data into an interpretable representation, consisting of interpretable components that are understandable to humans, such as super-pixels in image classification. These methods then perturb the raw data to either train a local linear surrogate model, compute Shapley values from coalitional game theory, or identify a decision rule that anchors the prediction sufficiently, thereby deriving model-agnostic explanations. Finally, these explanations, along with raw data, are visualized to highlight areas that are most influential in the model's decision-making. However, developing a model-agnostic XAI method for signal modulation classification faces three key challenges. (1) *How to identify interpretable components within signal modulation data?* Unique attributes of signal modulation data should be accounted for, such as its time-domain characteristics. (2) *How to generate in-distribution perturbations in signal modulation data?* An out-of-distribution perturbed sample causes classifier's prediction unreliable and untrustworthy. (3) *How to visualize model-agnostic explanations for signal modulation classification?* Visualizing feature importance in signal waveform diagrams is less intuitive, especially for longer signal lengths.

In this paper, to address the above challenges, we develop **M**odel-**A**gnostic **S**ignal modulation classification **E**xplainer (MASE), a novel model-agnostic XAI method for black-box DL-based modulation classifiers. MASE generates a class activation vector for a given signal sample as the explanation, highlighting each signal timestep's importance for the model's predicted class. Specifically, MASE extends LIME [24] with three innovations to address the aforementioned three challenges in signal modulation classification. First, by identifying the consecutive signal subsequences as interpretable components, MASE trains a linear surrogate model over the subsequence-based interpretable signal representation. Second, to avoid out-of-distribution perturbed samples, an in-distribution local signal sampling mechanism is proposed with the timestep and signal-to-noise ratio (SNR)-specific replacement noise. Third, for better explanation visualization, the constellation-based explanation visualization with activation threshold is adopted, providing more perceptiveness on signal modulations. Fig. 1 shows the difference between the model-specific and model-agnostic XAI methods in signal modulation classification. The model-specific explainer exploits model's inner working, while the model-agnostic explainer, free from model access requirements, utilizes query feedback via inference APIs for black box explainability.

Then, to address the lack of quantitative explanation assessments in signal modulation classification, we propose the first generic quantitative explanation evaluation framework focusing on explanation faithfulness, sensitivity, robustness and efficiency. For more reliable faithfulness evaluation, a novel metric, Normalized Area Over the Most Relevant Perturbation Curve (NAOPC), is proposed. In sensitivity and robustness evaluation, we introduce perturbation-to-signal ratio (PSR)-based metrics, which controls perturbation magnitude with relative power instead of the L_p norm. Additionally, IFIA [27], a powerful adversarial XAI attack, is adopted for generating adversarial perturbations in robustness evaluation.

With the quantitative explanation evaluation framework, we evaluate MASE with two state-of-the-art XAI methods (i.e., Grad-CAM [21] and MASK [22]) across two well-known modulation classification datasets (i.e., RadioML2016.10A [28] and RadioML2018.01A [29]) and for four types of black-box DL-based signal modulation classifiers (i.e., CNN-based, ResNet-based, LSTM-based classifiers, and Transformer-based [30]). We find that MASE (1) is more faithful to the classifiers, with an average of 44.7% improvement in explanation faithfulness; (2) is less sensitive to the random noise, with the explanation sensitivity reduced by 44.1%; (3) is more robust to adversarial attacks, showing an average improvement of 30.6% in explanation robustness; and (4) shows high efficiency, with average 2.10s and 4.49s of each explanation for the two datasets, respectively. Furthermore, after close visual examinations on explanation visualizations derived for both correct and wrong predictions, we demonstrate that MASE could provide more human interpretable explanations and allow human to better understand black-box classifier's behaviors, promoting model's reliability and trustworthiness. The main contributions of this paper are as follows:

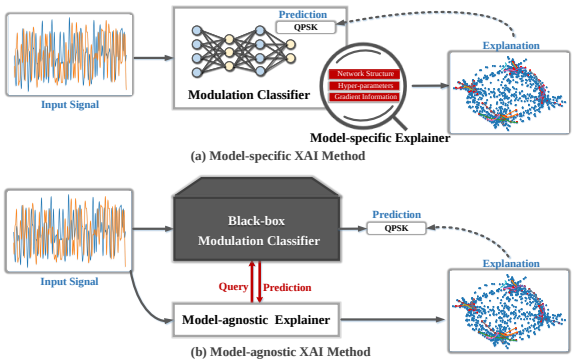


Fig. 1. Illustrations of model-specific and model-agnostic XAI methods in the signal modulation classification context.

- We propose a model-agnostic XAI method for black-box DL-based signal modulation classifiers, in which some delicate innovations based on LIME, e.g., subsequence-based interpretable signal representation, in-distribution local signal sampling and constellation-based explanation visualization, are proposed for adapting to signal data.
- We present the first generic quantitative evaluation framework for signal modulation classification explainers, consisting of the novel metrics, e.g., Normalized AOPC and PSR-based Max-Sensitivity, to enable reliable judgments in faithfulness, sensitivity, robustness and efficiency.
- With ten quantitative and qualitative experiments on two datasets and four classifiers to be explained, we demonstrate that MASE outperforms two state-of-the-art alternatives, successfully generating faithful, non-sensitive, robust and understandable explanations, and increasing the reliability and trustworthiness of opaque DL models.

The rest of the paper is organized as follows. In Section II, we introduce preliminary knowledge and problem definition. Section III and Section IV describe the proposed explainable method and quantitative explanation evaluation framework, respectively. Then, extensive experiments are done to validate the proposed method in Section V. In Section VI, we discuss the related works. Finally, Section VII concludes this work.

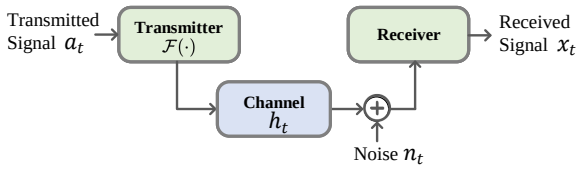


Fig. 2. Illustration of a wireless communication system consisting of a transmitter, a channel and a receiver.

II. Preliminaries and Problem Definition

In this section, we introduce the basic concepts of signal modulation classification and formalize the model-agnostic explanation problem for signal modulation classifiers.

A. Signal Modulation Classification

In a wireless communication system shown in Fig. 2, the received signal x_t with the time t can be expressed as

$$x_t = \mathcal{F}(a_t) \times h_t + n_t, \tag{1}$$

where a_t is the transmitted signal, $\mathcal{F}$ is a modulation function, h_t is a channel impulse response and n_t is the additive white Gaussian noise. The transmitter modulates the transmitted signal a_t on a carrier wave using $\mathcal{F}$ based on a specific modulation type [31]. In the receiver, accurate recovery of a_t from the modulated wave x_t involves identifying the modulation type. Thus, the objective of the signal modulation classification is to predict the modulation type of $\mathcal{F}$ using the received signal x_t, from a candidate modulation set $\mathcal{C}$.

Without extracting expert features from the received signal, in the DL-based signal modulation classifier, the input is the raw in-phase and quadrature (IQ) sequence (i.e., $x_t = (I_t, Q_t)$) or the transformed amplitude and phase (AP) sequence (i.e., $x_t = (A_t, \phi_t)$) [32], as calculated by

$$\begin{cases} A_t = \sqrt{I_t^2 + Q_t^2} \\ \phi_t = \arctan(Q_t / I_t) \end{cases}. \tag{2}$$

Given a segment of received signal $\boldsymbol{x}$ with length N, the core of DL-based signal modulation classifier is to learn a mapping function $f : \boldsymbol{x} \to \mathcal{Y}$, where $\mathcal{Y} = \{y_c \in [0,1] | c \in \mathcal{C}\}$ and y_c is the prediction probability of the signal $\boldsymbol{x}$ belonging to the modulation type c. The top predicted modulation type c^* is the one with the highest probability, i.e., $c^* = \arg\max_{c \in \mathcal{C}} y_c$.

B. Problem Definition

In this paper, we focus on the model-agnostic explainability problem for black-box DL-based signal modulation classification, to increase the model's transparency and understand the reliability of model decisions. Given any black-box classifier f and a signal instance of interest $\boldsymbol{x}$, we aim to learn a model-agnostic explainer Φ to provide an explanation e for the decision $f(\boldsymbol{x}) = c$. We assume that the black-box classifier f can be queried at will to obtain new prediction examples.

Following the prior work in this field [19], the explanation e is represented by a class activation vector $\boldsymbol{w} = \{w_i \in [0,1] | i \in \{0, 1, \cdots, N-1\}\}$, where each element w_i indicates the importance, relevance and contribution of corresponding input x_i on the prediction $f(\boldsymbol{x}) = c$. Thus, the model-agnostic explainer Φ is defined as

$$\Phi : f_c \times \boldsymbol{x} \to \boldsymbol{w}. \tag{3}$$

III. Methodology

In this section, we present **M**odel-**A**gnostic **S**ignal modulation classification **E**xplainer (MASE), the first model-agnostic method for explaining black-box signal modulation classifiers.

A. Method Overview

The overview of MASE method is illustrated in Fig. 3, consisting of four key components: (1) the *Subsequence-based Interpretable Signal Representation* converts the raw time-series representation to a subsequence-based binary vector via a signal segmentation algorithm, indicating the "presence" or "absence" of subsequences. Compared to a single signal point, the consecutive signal subsequence captures the signal time-domain features and is more intuitive for human. (2) The *In-distribution Local Signal Sampling* generates local signal samples by replacing the randomly picked subsequences with in-distribution non-informative noises, instead of constant value-based or random value-based noises which could be out-of-distribution. (3) The *Local Linear Explanation Generation* trains a local linear surrogate model on subsequence-based binary signal representation using newly sampled local samples and their black-box predictions, where the model's coefficients, serving as explanations, weight the attributions of different signal subsequences for the prediction of the given signal sample. (4) The *Constellation-based Explanation Visualization* highlights the important signal subsequences in the mapped constellation diagram for enhanced perceptive understanding of modulation type.

In MASE, there are two motivations for training a local surrogate model for each signal sample, rather than a global surrogate model. Firstly, it's difficult to accurately approximate a DL-based classifier's global decision function with simple interpretable models. Secondly, global models focus on explaining the model's overall behavior, whereas local models offer specific explanations for each sample's prediction.

From the above description, it can be seen that the proposed explainable method does not require any intrinsic information of the signal modulation classifier to be explained. Explanations are derived via queries to the black-box model. Thus, MASE is model-agnostic and shows more generality.

B. Subsequence-based Interpretable Signal Representations

In order to derive an explanation allowing human to understand, it is necessary to segment the raw data sample into a set of interpretable components which incorporate the human domain knowledge and make sense to humans. For example, an image instance is segmented into a set of super-pixels (i.e., a contiguous patch of similar pixels), possibly enriched with some semantic information and corresponded to human-understandable objects [33]. With these interpretable components, raw data representation is conveyed to the interpretable data representation which is a binary vector indicating the presence or absence of each interpretable component.

In MASE method, we propose a subsequence-based interpretable signal representation, where the consecutive subsequences are identified as valuable and interpretable components for the signal sample. Formally, given a signal sample $\boldsymbol{x}$

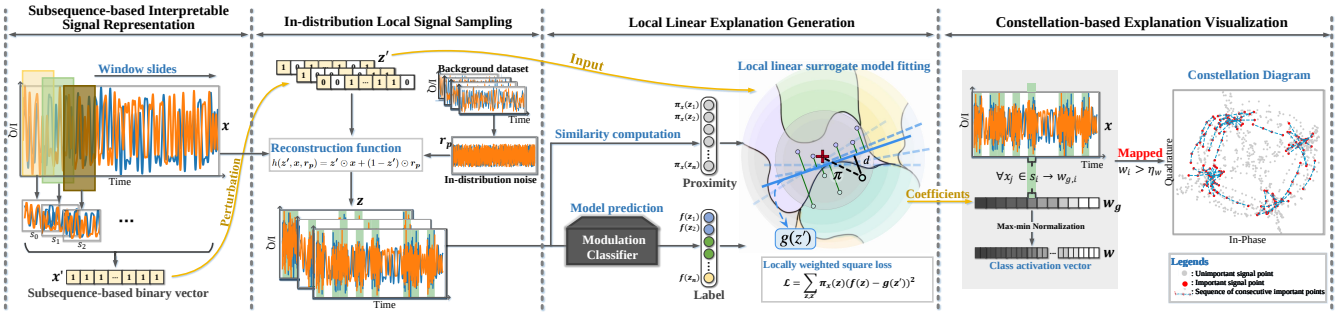


Fig. 3. Overview of the proposed model-agnostic explainable method for the signal modulation classification.

being explained, we first segment it into a set of subsequences $\boldsymbol{s} = \{s_0, s_1, \cdots, s_{d-1}\}$ via a signal segmentation algorithm $\Psi(\boldsymbol{x})$. Here, we adopt the sliding window-based segmentation algorithm to transform the signal data into a set of subsequences, thanks to its effectiveness, efficiency and prevalence in time series analytics [34]. Now let τ and β be the length and stride length of the sliding window, respectively. For a signal $\boldsymbol{x}$ with length N, the size of subsequences set d can be formulated as $d = \lfloor \frac{N-\tau}{\beta} \rfloor + 1$. Each subsequence $s_i \in \boldsymbol{s}$ is represented as $s_i = \{x_{i\times\beta}, \cdots, x_{i\times\beta+\tau}\}$.

Based on the subsequence set $\boldsymbol{s}$, the subsequence-based interpretable signal representation is written as a binary vector $\boldsymbol{x'} \in \{0,1\}^d$, where the i^{th} element indicates the presence or absence of the corresponding subsequence s_i. For instance, the interpretable signal representation for the raw signal sample $\boldsymbol{x}$ is an all-ones vector, i.e., $\boldsymbol{x'} \in \mathbb{1}^d$.

C. In-distribution Local Signal Sampling

In order to capture and explain the local behavior of the black-box signal modulation classifier f, a key step is to generate a number of local samples which are nearby to the raw signal sample $\boldsymbol{x}$. On the top of the subsequence-based interpretable signal representation, we develop an in-distribution local signal sampling mechanism.

Firstly, we perturb the raw signal $\boldsymbol{x}$'s interpretable representation $\boldsymbol{x'} \in \mathbb{1}^d$, by flipping an arbitrary number of ones to zeros, to generate a local sample $\boldsymbol{z'} \in \{0,1\}^d$. Then, we reconstruct the local signal sample in the raw representation $\boldsymbol{z} = h(\boldsymbol{z'})$, where $h(\cdot)$ is defined as the reconstruction function. In the reconstruction, the interpretable component (i.e., signal subsequence) $\boldsymbol{s_i}$ with $\boldsymbol{z'_i} = 1$ will be kept. Otherwise, it will be replaced with a non-informative mask in order to deactivate this component for prediction. In the image domain, the non-informative mask can be the constant predefined pixels (such as gray pixels) or the random pixels [24]. However, due to the characteristics of the signal sample, constant value-based replacement and random value-based replacement often create out-of-distribution local samples, causing classifier's prediction untrustworthy. Hence, in order to solve the above challenge, we generate SNR-specific replacement noise sequences from a "background" dataset that exhibits the same data distribution with the raw signal sample $\boldsymbol{x}$, serving as the in-distribution and non-informative masks. In practice, a subset of the dataset, from which the raw signal sample $\boldsymbol{x}$ is drawn, is typically used as the background dataset [35].

More specifically, considering the in-distribution non-informative replacement noise may vary by timestep and SNR, thus we generate the replacement noises $\boldsymbol{r}$ on a per-SNR and timestep-by-timestep bias. For each SNR p, we collect signal samples from the background dataset $\mathcal{D}$ to estimate the mean μ_i and standard deviation σ_i at each timestep $i \in [0, N-1]$. Then, the timestep and SNR-specific replacement noise $\boldsymbol{r_p}$ is generated following the Gaussian distribution $\mathcal{N}(\cdot,\cdot)$, as shown in Equation 4. With $\boldsymbol{r_p}$ as the replacement mask, the reconstruction function $h(\cdot)$ can be formulated as Equation 5, where $\odot$ is the Hadamard product.

$$\boldsymbol{r_p} = \{r_{p,i} \sim \mathcal{N}(\mu_i, \sigma_i^2) | i \in [0, N-1]\} \tag{4}$$

$$h(\boldsymbol{z'}, \boldsymbol{x}, \boldsymbol{r_p}) = \boldsymbol{z'} \odot \boldsymbol{x} + (1 - \boldsymbol{z'}) \odot \boldsymbol{r_p} \tag{5}$$

With the SNR-specific replacement noise sequences and the reconstruction function, we generate a lot of in-distribution local signal samples. Moreover, we construct a local signal dataset $\boldsymbol{\mathcal{Z}} = \{(\boldsymbol{z_i}, \boldsymbol{z'_i}) | i \in [0, M-1]\}$ containing the raw representation and interpretable representation of each local signal sample, where M is the number of generated local signal samples. The dataset $\boldsymbol{\mathcal{Z}}$ will be used for training an interpretable linear surrogate model to get locally faithful explanations in the following subsection.

D. Local Linear Explanations Generation

Next, we introduce how to train a linear model g as the local surrogate of the black-box signal modulation classifier f around the given signal sample $\boldsymbol{x}$, to produce an explanation for the prediction $f_c(\boldsymbol{x})$. In order to explain the importance of each interpretable component (i.e., signal subsequence, in our case), the linear surrogate model g needs to act over absence/presence of the interpretable components and emulate the prediction of f. Based on the interpretable signal representation, g is defined as $g(\boldsymbol{z'}) = \boldsymbol{w_g} \cdot \boldsymbol{z'}$, whose domain is $\{0,1\}^d$. If g is good enough for local approximation, then each coefficient $w_{g,i} \in \boldsymbol{w_g}$ weights the importance of corresponding subsequence s_i, which is treated as an explanation.

In order to encourage g to learn the decision surface of f around $\boldsymbol{x}$, there are two important keys. Firstly, for a sample $\boldsymbol{z}$ from local signal dataset $\boldsymbol{\mathcal{Z}}$, the prediction of g should be similar with the prediction of f, i.e., minimizing

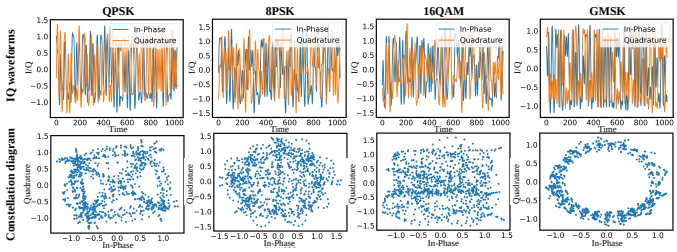


Fig. 4. Signal IQ waveforms and corresponding constellation diagrams of four modulation types at SNR = 30dB in RadioML2018.01A dataset.

the distance between $g(\boldsymbol{z'})$ and $f_c(\boldsymbol{z})$. Here, $f_c(\boldsymbol{z})$ is the probability that local signal sample $\boldsymbol{z}$ belonging to modulation type of interest $c \in \mathcal{C}$. Moreover, considering a sample with a larger similarity to $\boldsymbol{x}$ is more important for learning locally faithful explanations. Thus, each sample $\boldsymbol{z}$ is weighted according to its proximity to $\boldsymbol{x}$. The proximity function is defined as $\pi_{\boldsymbol{x}}(\boldsymbol{z})$, capturing the locality around $\boldsymbol{x}$. Thus, we train g by minimizing the locally weighted square loss $\mathcal{L}$:

$$\mathcal{L}(f, g, \pi_{\boldsymbol{x}}) = \sum_{\boldsymbol{z}, \boldsymbol{z'} \in \boldsymbol{Z}} \pi_{\boldsymbol{x}}(\boldsymbol{z})(f_c(\boldsymbol{z}) - g(\boldsymbol{z'}))^2. \tag{6}$$

In our setting, we use an exponential kernel defined on Frobenius distance as the proximity function, i.e., $\pi_{\boldsymbol{x}}(\boldsymbol{z}) = exp(-\|\boldsymbol{x} - \boldsymbol{z}\|_F^2)$.

The coefficients (i.e., $w_{g,0}, \cdots, w_{g,d-1}$) of the well-trained model g weight the attribution of signal subsequences (i.e., $s_0, \cdots, s_{d-1}$) for the prediction $f_c(\boldsymbol{x})$. Furthermore, we transform the coefficient vector $\boldsymbol{w_g} \in \mathbb{R}^d$ to the class activation vector $\boldsymbol{w} \in [0, 1]^N$ by a max-min normalization and assigning all timesteps belonging to a subsequence (i.e., $\forall x_j \in s_i$) with the same importance $w_{g,i}$. Due to the overlaps between subsequences, a timestep may belong to multiple subsequences. We assign the overlapped timestep with the highest subsequence importance. Finally, with a linear surrogate model over the interpretable signal representation, MASE achieves local explanations for a black-box signal modulation classifier.

E. Constellation-based Explanation Visualization

With the derived class activation vector $\boldsymbol{w}$, we propose a constellation-based explanation visualization for signal modulation classification. Constellation diagram is a widely used signal representation [36], which maps the signal sample points x_i into scattering points on a 2-D complex plane (aka IQ plane) in the rectangular coordinate system, i.e., $x_i = (I_i, Q_i)$. Compared to waveform, the constellation diagram of the signal sample provides more perceptive information about modulation type, as shown in Fig 4.

We highlight the signal sample point x_i with the attribution greater than the activation threshold η_w (i.e., $w_i > \eta_w$) in the constellation diagram, meaning that these signal points are significant for recognizing the modulation type. Furthermore, to emphasize the time-domain signal features, we connect the consecutive highlighted constellation points via blue lines. In this way, the constellation-based explanation visualization only spotlights the signal features relevant to model predictions, providing a concise and understandable explanation for human, and facilitating the understanding into reliability and trustworthiness of the model decisions. Algorithm 1 summarizes the proposed model-agnostic explainable method for black-box signal modulation classifiers.

Algorithm 1 **M**odel-**A**gnostic **S**ignal modulation classification **E**xplainer (MASE)

Input: Raw signal sample $\boldsymbol{x}$ with SNR p; Modulation type of interest c; Black-box classifier f; Number of local samples M; Proximity function π; Signal segmentation function Ψ; Reconstruction function h; Background signal dataset $\mathcal{D}$;
Output: Class activation vector $\boldsymbol{w}$;

Initialize local sample dataset $\boldsymbol{Z} \leftarrow \{\}$
// Subsequence-based interpretable signal representation
Generate signal subsequence set $\boldsymbol{s} = \{s_0, s_1, \cdots, s_d\}$ via $\Psi(\boldsymbol{x})$
Represent $\boldsymbol{x}$ using the interpretable signal representation $\boldsymbol{x'} \in \mathbb{1}^d$
// In-distribution local signal sampling
for $i = 1$ to M **do**
    Randomly perturb $\boldsymbol{x'}$ to generate $\boldsymbol{z'_i} \in \{0, 1\}^d$
    Generate the SNR-specific replacement noise sequence $\boldsymbol{r_p}$ from the background signal dataset $\mathcal{D}$
    Reconstruct the signal $\boldsymbol{z_i}$ from $\boldsymbol{z'_i}$ via

$$\boldsymbol{z_i} = h(\boldsymbol{z'_i}, \boldsymbol{x}, \boldsymbol{r_p}) = \boldsymbol{z'_i} \odot \boldsymbol{x} + (1 - \boldsymbol{z'_i}) \odot \boldsymbol{r_p}$$

    Compute the proximity $\pi_{\boldsymbol{x}}(\boldsymbol{z_i})$ between $\boldsymbol{z_i}$ to $\boldsymbol{x}$
    Obtain the black-box prediction $f_c(\boldsymbol{z_i})$
    $\boldsymbol{Z} \leftarrow \boldsymbol{Z} \cup \{\boldsymbol{z_i}, \boldsymbol{z'_i}\}$
end for
// Local Linear Explanation Generation
Initialize a linear model $g(\boldsymbol{z'}) = \boldsymbol{w_g} \cdot \boldsymbol{z'}$
Train g via minimizing the locally weighted square loss $\mathcal{L}$:

$$\mathcal{L}(f, g, \pi_{\boldsymbol{x}}) = \sum_{\boldsymbol{z_i}, \boldsymbol{z'_i} \in \boldsymbol{Z}} \pi_{\boldsymbol{x}}(\boldsymbol{z_i})(f_c(\boldsymbol{z_i}) - g(\boldsymbol{z'_i}))^2$$

Transform coefficient $\boldsymbol{w_g} \in \mathbb{R}^d$ to $\boldsymbol{w} \in [0, 1]^N$ via max-min normalization and assignment for each timestep
return $\boldsymbol{w}$

IV. Quantitative Explanation Evaluation Framework

This section presents the quantitative explanation evaluation framework for signal modulation classification. The overview is shown in Fig 5 including (a) explanation faithfulness, (b) explanation sensitivity, (c) explanation robustness and (d) explanation efficiency.

A. Explanation Faithfulness

The faithfulness of an explanation is defined as the explanation's ability to accurately rank the signal timesteps by their importance for prediction, such that when perturbing important timesteps, model's prediction confidence score will drop rapidly. Under this intuition, we adopt the *Most Relevant First* (MoRF) perturbation curve [37] to evaluate how fast the $f_c(\boldsymbol{x})$ decreases, when we progressively perturb the most relevant timesteps in the raw signal sample $\boldsymbol{x}$, sorted by the derived explanation $\boldsymbol{w}$ (i.e., the class activation vector).

Formally, according to the class activation vector $\boldsymbol{w}$, we derive an ordered sequence of timesteps $\mathcal{O} = \{o_1, \cdots, o_N\}$, where o_k is the k^{th} most relevant timestep. Thus, for all indices of $\mathcal{O}$, the property $(i < j) \Leftrightarrow (w_{o_i} > w_{o_j})$ holds. We perturb the most relevant timesteps by replacing their

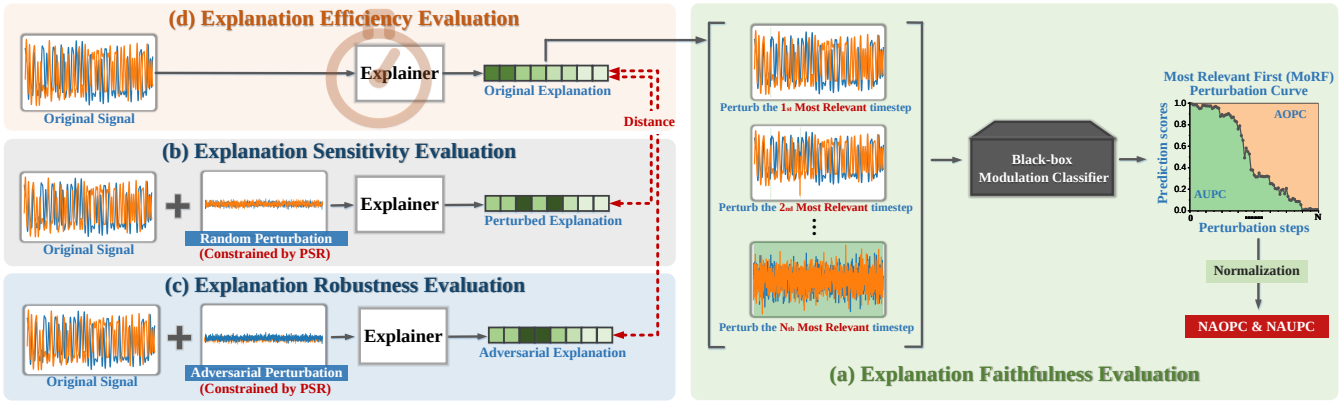
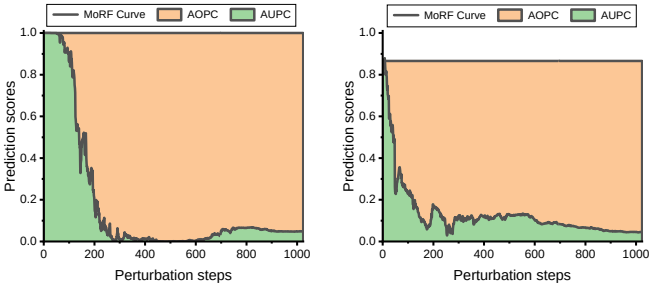


Fig. 5. Overview of quantitative explanation evaluation framework for the signal modulation classification.



(a) AOPC=847.1, AUPC=177.9; NAOPC=0.826, NAUPC=0.174

(b) AOPC=753.3, AUPC=134.3; NAOPC=0.849, NAUPC=0.151

Fig. 6. A motivating example shows AUPC and AOPC may give opposite indications about which explanation is more faithful.

values with a timestep and SNR-specific noise sequence $\boldsymbol{r_p}$, to avoid out-of-distribution perturbation and spurious model classification (see details in Section III-C). Based on the k most relevant timesteps, the perturbed signal sample $\boldsymbol{x}_{MoRF}^{(k)}$ is generated via the following recursive formulation:

$$\boldsymbol{x}_{MoRF}^{(k)} = \Omega(\boldsymbol{x}_{MoRF}^{(k-1)}, o_k, \boldsymbol{r_p}), \qquad k \in [1, 2, \cdots, N] \tag{7}$$

where $\boldsymbol{x}_{MoRF}^{(0)} = \boldsymbol{x}$ and the function $\Omega(\boldsymbol{x}_{MoRF}^{(k-1)}, o_k, \boldsymbol{r_p})$ replaces the value of $\boldsymbol{x}_{MoRF}^{(k-1)}$ at the timestep o_k with the value of noise $\boldsymbol{r_p}$ at o_k. The faster the perturbation curve $f_c(\boldsymbol{x}_{MoRF}^{(k)})$ decreases, the more faithful the explanation $\boldsymbol{w}$ is.

To quantitatively evaluate the explanation faithfulness, there are two common metrics to quantify the degree of curve decrease: the *Area Over the MoRF Perturbation Curve* (AOPC) [38], [39] and the *Area Under the MoRF Perturbation Curve* (AUPC) [40], [41]. Based on the trapezoidal rule, the AOPC and AUPC are calculated by Equation 9 and Equation 10, respectively. A smaller AUPC or a larger AOPC indicates a more faithful explanation. However, for two MoRF curves with different initial scores (i.e., prediction probability of the raw signal sample), AUPC and AOPC may give opposite indications. As shown in Fig. 6, the left explanation with a larger AOPC indicates the left is more faithful. However, the right explanation with a lower AUPC shows the right explanation is more faithful. The discrepancy arises from the differing initial scores on MoRF perturbation curves, leading to varying total areas and, consequently, inconsistent evaluations.

$$Diff(k) = f_c(\boldsymbol{x}_{MoRF}^{(0)}) - f_c(\boldsymbol{x}_{MoRF}^{(k)}) \tag{8}$$

$$AOPC = \sum_{k=1}^{N} \frac{Diff(k-1) + Diff(k)}{2} \tag{9}$$

$$AUPC = \sum_{k=1}^{N} \frac{f_c(\boldsymbol{x}_{MoRF}^{(k-1)}) + f_c(\boldsymbol{x}_{MoRF}^{(k)})}{2} \tag{10}$$

To solve this problem, we design the *Normalized AOPC* (NAOPC) and *Normalized AUPC* (NAUPC), which normalize AOPC and AUPC based on their total areas, as calculated in Equation 11 and Equation 12, respectively. Since the total area is the sum of the areas above and below the MoRF perturbation curves, this normalization ensures the sum of NAOPC and NAUPC equals 1 (i.e., $NAOPC + NAUPC = 1$), indicating that an explanation with a larger NAOPC or a smaller NAUPC is more faithful. By considering the proportion of AOPC and AUPC relative to their total area, NAOPC and NAUPC effectively mitigate the impact of varying initial scores, thereby offering consistent and reliable evaluations for explanation faithfulness across different signal samples and signal modulation classifiers. As shown in Fig. 6, both NAOPC and NAUPC consistently indicate the right is more faithful.

$$NAOPC = \frac{AOPC}{f_c(\boldsymbol{x}_{MoRF}^{(0)}) \times N} \tag{11}$$

$$NAUPC = \frac{AUPC}{f_c(\boldsymbol{x}_{MoRF}^{(0)}) \times N} \tag{12}$$

B. Explanation Sensitivity

The explanation sensitivity measures the variability of an explanation in response to random input perturbations, common in real-world wireless systems. A lower sensitivity is preferable, ensuring the explanation would not be affected significantly, as the raw input varies slightly. Conversely, a high sensitivity could undermine human's trust in the explanations. In our evaluation framework, we adopt two metrics, *Max-Sensitivity* [42] and *Delta NAOPC* [43], to measure explanation sensitivity under random noise perturbations.

In wireless communication, perturbation imperceptibility is typically measured by the relative power of the perturbation

with respect to the received signal [44], instead of L_p norm commonly used in the image domain. Thus, we define $\boldsymbol{P_\epsilon}$ as a set of imperceptible signal perturbations constrained by the perturbation-to-signal ratio (i.e., PSR = ϵ). PSR is denoted as the power ratio of the perturbation $\boldsymbol{p}$ to the raw received signal $\boldsymbol{x}$ [45], as shown in Equation 13, where $pow(\boldsymbol{p})$ and $pow(\boldsymbol{x})$ represent their respective powers. In the evaluation, we generate $\boldsymbol{p} \in \boldsymbol{P_\epsilon}$ with in-distribution timestep and SNR-specific Gaussian noise $\boldsymbol{r_p}$, as shown in Equation 14, where ϵ is the desired PSR and the power of $\frac{\boldsymbol{r_p}}{\sqrt{pow(\boldsymbol{r_p})}}$ is 1.

$$PSR = 10 \times log_{10}(\frac{pow(\boldsymbol{p})}{pow(\boldsymbol{x})}) \tag{13}$$

$$\boldsymbol{p} = \sqrt{pow(\boldsymbol{x}) \times 10^{\frac{\epsilon}{10}}} \times \frac{1}{\sqrt{pow(\boldsymbol{r_p})}} \times \boldsymbol{r_p} \tag{14}$$

Max-Sensitivity (MS) is used to measure the maximum change between the original explanation $\boldsymbol{w} = \Phi(f_c, \boldsymbol{x})$ and the perturbed explanation $\boldsymbol{w_p} = \Phi(f_c, \boldsymbol{x} + \boldsymbol{p})$, where a minor random perturbation $\boldsymbol{p} \in \boldsymbol{P_\epsilon}$ is injected to the raw input signal $\boldsymbol{x}$. With Monte-Carlo sampling, MS for signal modulation classification explanation is calculated as Equation 15, where $D_{cos}(\cdot, \cdot)$ refers to the cosine distance between two class activation vectors. Moreover, *Delta NAOPC* (ΔNAOPC) is introduced to assess the impact of random noise perturbation on the faithfulness of explanations. It measures the difference in explanation's NAOPC before and after perturbations, as calculated as Equation 16. Together, MS and ΔNAOPC provide a comprehensive evaluation and understanding into how noise affect the explainable methods. An explanation characterized by low MS and ΔNAOPC is more desired, since it has more reliability and stability against random perturbations.

$$MS = \max_{\boldsymbol{p} \in \boldsymbol{P_\epsilon}} D_{cos}(\boldsymbol{w}, \boldsymbol{w_p}) \tag{15}$$

$$\Delta NAOPC = \max_{\boldsymbol{p} \in \boldsymbol{P_\epsilon}} (NAOPC(\boldsymbol{w}) - NAOPC(\boldsymbol{w_p})) \tag{16}$$

C. Explanation Robustness

Recent advances in adversarial explainable AI (AdvXAI) have revealed significant vulnerabilities in state-of-the-art explainable methods, raising serious concerns on their reliability and security [46]. Research shows that imperceptible adversarial perturbations can drastically change the derived explanations, even though the predicted label remain unchanged. This concern is particularly critical in explanations for signal modulation classification, which are extensively employed in security-sensitive applications such as military wireless systems. Therefore, this section focuses on evaluating the robustness of explainable methods against AdvXAI attacks for signal modulation classification.

As an initial effort in explanation robustness evaluation for signal modulation classification, we employ a well-known adversarial attack method in AdvXAI, the Iterative Feature Importance Attacks (IFIA), to generate adversarial perturbations. First introduced in [27], IFIA aims to perturb the feature attribution map by decreasing the relative importance of the k initially most important input features, while ensuring the predictions remain unchanged. Note that, similar to the random perturbations created for explanation sensitivity evaluation, the adversarial perturbations generated by IFIA are constrained by PSR = ϵ to ensure perturbation imperceptibility. Formally, the generation of adversarial perturbations, denoted as $\boldsymbol{p_{adv}}$, can be defined as Equation 17, where $D_{top_k}(\cdot, \cdot)$ measures the distance between the top-K features of two explanations.

$$\begin{aligned} &\arg\max_{\boldsymbol{p_{adv}}} D_{top_k}(\Phi(f_c, \boldsymbol{x}), \Phi(f_c, \boldsymbol{x} + \boldsymbol{p_{adv}})) \\ &\text{s.t.} \quad PSR(\boldsymbol{p_{adv}}, \boldsymbol{x}) = \epsilon, \quad f(\boldsymbol{x}) = f(\boldsymbol{x} + \boldsymbol{p_{adv}}) \end{aligned} \tag{17}$$

Then, we adopt *Max-Sensitivity* (MS) to measure the maximum change in explanation, and use *Delta NAOPC* (ΔNAOPC) to measure the maximum changes in explanations faithfulness when adversarial perturbations $\boldsymbol{p_{adv}}$ are injected to the raw input signal $\boldsymbol{x}$. An explainable method with low MS and ΔNAOPC under adversarial perturbations is preferred, indicating enhanced robustness against AdvXAI attacks.

D. Explanation Efficiency

For explanation efficiency, we use a generic metric *Average Computation Time* (ACT) to measure the average time of each explanation generation, as calculated by dividing the total time elapsed T_{total} by the number of signal samples N_{total},

$$ACT = T_{total}/N_{total}. \tag{18}$$

Note that, the signal length N, classifier complexity $O(f)$ and explainer complexity $O(\Phi)$ are the key factors affecting the time taken to generate an explanation. Explainers with lower ACT are more promising, especially in real-time applications.

V. Experimental Results and Analysis

In this section, we first introduce the experimental setup, and then investigate the hyper-parameter sensitivity of MASE. Lastly, we present the quantitative and qualitative evaluations to validate the effectiveness of MASE.

A. Experimental Setup

1) Datasets: The evaluation of the proposed method is carried out on two publicly available modulation classification datasets: RadioML2016.10A [28] and RadioML2018.01A [29]. The RadioML2016.10A dataset consists of total 22,000 signal samples including 11 modulations, where each signal sample is a base-band I/Q matrix of 2×128. The RadioML2018.01A dataset is larger, consisting of 24 modulations and 2,555,904 signal samples, where the sampling length is 1024. The detail parameters of two datasets can be found in Table I. For each dataset, the samples are split into training, validation, and test set with a ratio of 7:1:2.

TABLE I
PARAMETERS OF RADIOML2016.10A AND RADIOML2018.01A.

Parameter	RadioML2016.10A	RadioML2018.01A
Modulation types	11 classes (AM-DSB,AM-SSB, 8PSK,BPSK,CPFSK, QPSK,WBFM,GFSK, 64QAM,16QAM, 4PAM)	24 classes (OOK,4ASK,8ASK,BPSK,QPSK, FM,8PSK,16PSK,32PSK,16APSK, 32APSK,64APSK,128APSK,16QAM, 32QAM,64QAM,128QAM,256QAM, AM-SSB-WC,AM-SSB-SC,GMSK, AM-DSB-WC,AM-DSB-SC,OQPS)
SNR(dB)	-20:2:18	-20:2:30
Sample length	128	1024
Number of samples	220,000	2,555,904

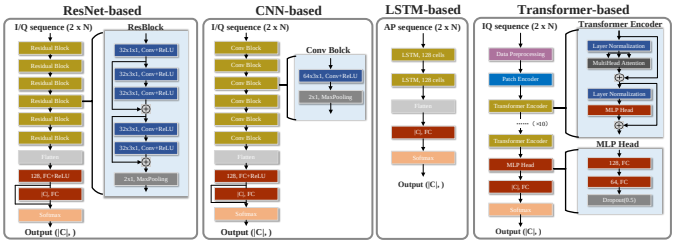
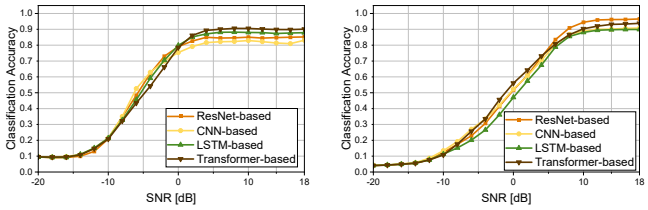
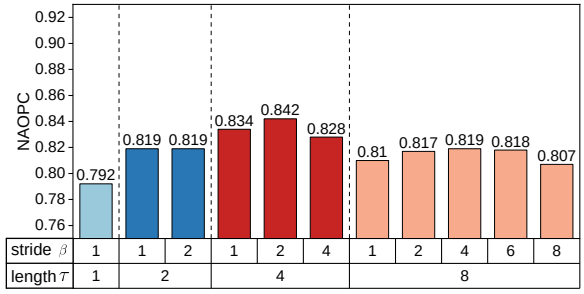


Fig. 7. Model structures of ResNet-based, CNN-based, LSTM-based, and Transformer-based signal modulation classifiers to be explained.

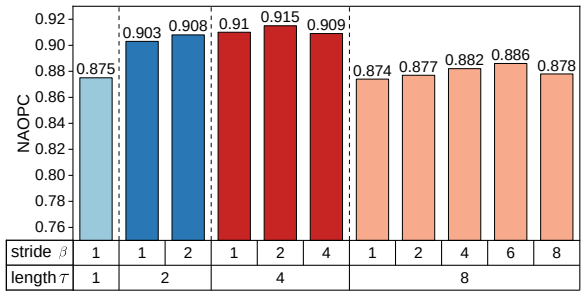


(a) RadioML2016.10A (b) RadioML2018.01A

Fig. 8. The classification accuracy w.r.t. SNR of different signal modulation classifiers on (a) RadioML2016.10A and (b) RadioML2018.01A datasets.



(a) RadioML2016.10A



(b) RadioML2018.01A

Fig. 9. Performance of the NAOPC ↑ metric under different (τ, β) combinations for the ResNet-based classifier on (a) RadioML2016.10A and (b) RadioML2018.01A datasets.

2) Signal Modulation Classifiers: For each dataset, we train a CNN-based, a ResNet-based, a LSTM-based, and a Transformer-based model to serve as signal modulation classifiers in need of explanations. The model structures are shown in Fig. 7. For the CNN-based, ResNet-based and Transformer-based classifiers, the input is the raw I/Q sequences. While for the LSTM-based classifier, we feed the transformed AP sequences (see Equation 2) as input, since the LSTM-based classifier performs poorly with I/Q inputs, as reported in [32]. The classification accuracy with respective to (w.r.t.) SNR is shown in Fig. 8, and all the models obtain nearly state-of-the-art performance on their respective datasets.

3) Baselines: In prior work [19], the authors adopt Grad-CAM [21] to explain the CNN-based and ResNet-based signal modulation classifiers, and employ MASK [22] for the LSTM-based classifier. Thus, in this evaluation, we compare MASE with these two state-of-the-art explainable methods: (a) Grad-CAM [21], a gradient-based method, generates a coarse localization map to highlight significant parts of inputs, by using the gradient data from the last convolutional layer or LSTM layer; and (b) MASK [22], a perturbation-based method, dynamically optimizes a perturbation vector to discover which parts of the raw input most affect its output score.

4) Metrics: We evaluate the MASE method and the baselines with the quantitative explanation evaluation framework presented in Section IV. Specifically, we employ two widely used metrics, *Area Over the MoRF Perturbation Curve* (AOPC) and *Area Under the MoRF Perturbation Curve* (AUPC), along with our two novel metrics, *Normalized AOPC* (NAOPC) and *Normalized AUPC* (NAUPC), to assess explanation faithfulness. Furthermore, we use *Max-Sensitivity* (MS) and *Delta-NAOPC* (ΔNAOPC) both to evaluate explanation sensitivity to random noise perturbation and explanation robustness against adversarial noise perturbation. Lastly, *Average Computation Time* (ACT) is used for explanation efficiency.

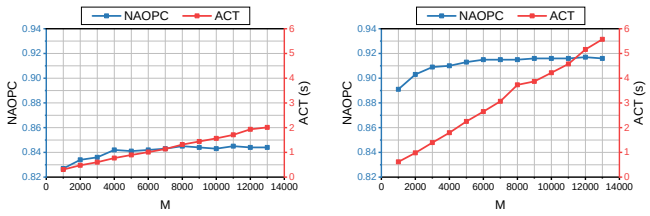
5) Implementation Details: We implement the proposed method MASE using Keras with TensorFlow [47] as the backend. For each dataset, we evaluate explainable methods on signal samples randomly selected from the test set. And we adopt test set as the background signal dataset $\mathcal{D}$ for generating SNR-specific replacement noise sequence $\boldsymbol{r_p}$. All the reported results are the average of five runs. For each method, we have used grid search to compute the optimum values of hyperparameters to get the best possible results.

All experiments are conducted on a Linux server with GPU (GeForce RTX 3090), and its operating system is Ubuntu 16.04.1. The Python and Keras versions are 3.7.0 and 2.6.0.

B. Hyper-parameters Analysis

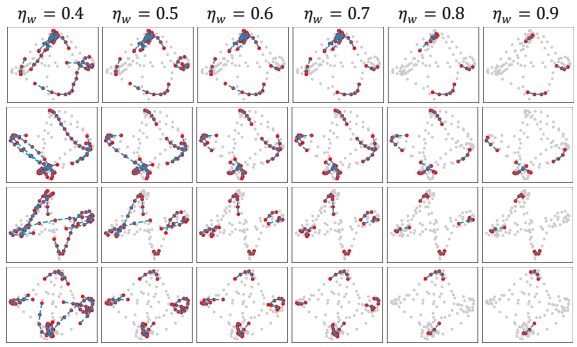
In this section, we systematically investigate the sensitivity of four main hyper-parameters of MASE: the length τ and stride length β of sliding windows, the number of generated local samples M, and the activation threshold η_w.

1) Analysis of τ and β: In the sliding window-based subsequence segmentation algorithm, the (τ, β) pair determines the generated signal subsequence set and further interpretable signal representation. We test different combinations for the ResNet-based signal modulation classifier on two datasets and corresponding NAOPC values are shown in Fig. 9. We can see that as the sliding length τ increases, the results show a trend of increasing first and then decreasing. Too small or large length would harm the model, because a short sliding window may not cover a complete signal feature, while a longer will include excessive information. For the stride length β, as we gradually increase it, the performance grows

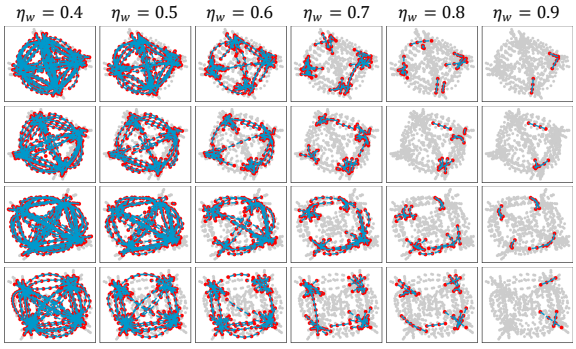


(a) RadioML2016.10A (b) RadioML2018.01A

Fig. 10. Performance of the NAOPC ↑ and the ACT ↓ metrics under different M for the ResNet-based classifier on (a) RadioML2016.10A and (b) RadioML2018.01A datasets.



(a) RadioML2016.10A



(b) RadioML2018.01A

Fig. 11. Explanation visualization of QPSK modulation with different η_w on (a) RadioML2016.10A and (b) RadioML2018.01A datasets.

since a larger β could reduce the number of generated signal subsequences and the dimension of the interpretable signal representation. Nevertheless, when β is larger than the optimal value, increasing β will hurt the performance probably because many significant signal subsequences are skipped and ignored. On both datasets, MASE achieves optimal performance at the length $\tau = 4$ and the stride length $\beta = 2$.

2) Analysis of M: In the MASE method, the number of in-distribution local signal samples M determines the volume of training data for local surrogate model, which is the key hyper-parameter on explanation faithfulness and explanation efficiency. We vary the value of M from 1000 to 13000 and report corresponding NAOPC and ACT values in Fig. 10. As we can see, the NAOPC value improves with the increase of the number of generated local samples M, and the performance tends to be stable once the M reaches around 6000. However, the ACT value increases nearly linearly with M. Therefore, we employ $M = 6000$ to balance the trade-off between faithfulness and efficiency.

3) Analysis of η_w: In constellation-based explanation visualization, the activation threshold η_w determines the number of highlighted signal features. Fig. 11 shows explanations from the ResNet-based classifier under different η_w. The visualizations with a small threshold (i.e., η_w=0.4) spotlight a large number of signal features, providing a poor interpretability for human. However, a high threshold (i.e., η_w=0.9) filters most of the relevant signal features. Hence, in order to provide concise explanation visualization as well as highlight these relevant signal features, we set $\eta_w = 0.7$ in the following evaluations.

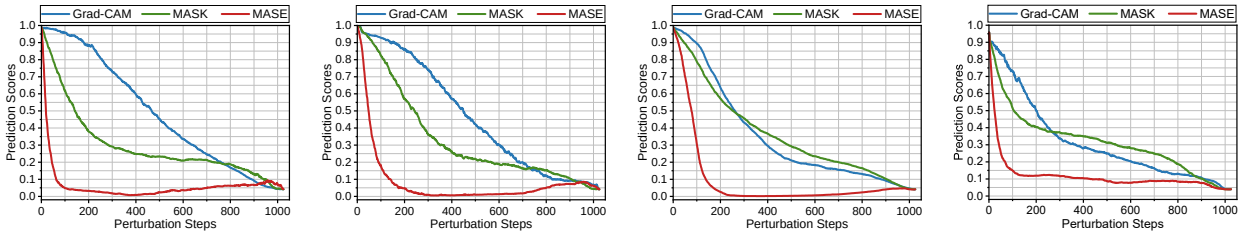
C. Quantitative Evaluation

In this section, we perform quantitative evaluations on explanation faithfulness, sensitiveness, robustness and efficiency, showing the superiority of MASE in comparison to the state-of-the-art baselines.

1) Explanation Faithfulness Analysis: To evaluate the local faithfulness of explanations to the classifier's behaviors, we present the AOPC, AUPC, NAOPC, and NAUPC metrics in Table II, where each method is evaluated for four modulation classifiers and two datasets. Based on the results, we make the following observations: (a) MASE consistently and significantly outperforms all baselines by learning better rankings for timestep importance. For RadioML2016.10A dataset, MASE achieves average NAOPC performance gains over the best baseline by 46.7%, 61.6%, 55.5%, and 51.7% for ResNet-based, CNN-based, LSTM-based and Transformer-based classifiers, respectively. And for RadioML2018.01A dataset, the improvements are 37.0%, 32.8%, 39.8%, and 32.2%. It demonstrates the linear surrogate model, trained over the subsequence-based signal representations, is locally faithful to the black-box classifier. (b) Among the baselines, the MASK method generally outperforms the gradient-based method (i.e., Grad-CAM), which proves that compared with learning a perturbation, the gradient-based values are prone to miscalculate the sensitivity of the model's output to the features represented in the input signal. (c) AOPC and AUPC sometimes yield conflicting indications when evaluating explanation faithfulness across different classifiers, as seen with MASE for the ResNet-based and CNN-based classifiers under the RadioML2016.10A dataset. The larger AOPC for the ResNet-based classifier (i.e., 97.87) indicates the explanations for ResNet-based classifier are more faithful. Conversely, the lower AUPC for the CNN-based classifier (i.e., 15.903) implies higher faithfulness in the CNN-based classifier's explanation. This discrepancy arises due to the different initial scores on the MoRF perturbation curves for the two classifiers, leading to various total areas. Our proposed NAOPC and NAUPC address this issue by normalizing AOPC and AUPC against their total area, providing more consistent and reliable evaluations for explanation faithfulness across different classifiers. For instance, both NAOPC and NAUPC (i.e., 0.858 and 0.142, respectively) indicate the MASE's explanations for the CNN-based classifier are more faithful compared to those for ResNet-based classifier.

TABLE II
EXPLANATION FAITHFULNESS COMPARISON OF DIFFERENT EXPLAINABLE METHODS FOR DIFFERENT CLASSIFIERS (BEST RESULT IN BOLD, AND SECOND BEST UNDERLINED). '*Imp.*' MEANS THE NAOPC'S RELATIVE IMPROVEMENT PERCENTAGE OF OUR METHOD AGAINST THE BEST BASELINE.

Dataset	Methods	ResNet-based Classifier				CNN-based Classifier				LSTM-based Classifier				Transformer-based Classifier			
		AUPC↓	AOPC↑	NAUPC↓	NAOPC↑	AUPC↓	AOPC↑	NAUPC↓	NAOPC↑	AUPC↓	AOPC↑	NAUPC↓	NAOPC↑	AUPC↓	AOPC↑	NAUPC↓	NAOPC↑
RadioML 2018.01A	**Grad-CAM**	476.691	543.904	0.466	0.534	499.467	513.922	0.492	0.508	355.366	658.069	0.349	0.651	321.124	658.981	0.316	0.684
	MASK	341.026	679.569	0.332	0.668	300.973	712.416	0.296	0.704	371.826	641.607	0.363	0.637	322.181	657.925	0.330	0.670
	MASE (ours)	**88.941**	**931.654**	**0.085**	**0.915**	**68.378**	**945.015**	**0.065**	**0.935**	**92.997**	**920.435**	**0.090**	**0.910**	**94.606**	**885.5**	**0.096**	**0.904**
	Imp.	-	-	-	37.0%	-	-	-	32.8%	-	-	-	39.8%	-	-	-	32.2%
RadioML 2016.10A	**Grad-CAM**	48.143	68.091	0.433	0.567	55.853	55.005	0.524	0.476	64.227	57.172	0.533	0.467	49.024	68.712	0.441	0.559
	MASK	46.625	69.61	0.426	0.574	50.243	60.615	0.469	0.531	53.299	68.183	0.440	0.560	46.621	71.114	0.426	0.574
	MASE (ours)	**18.374**	**97.87**	**0.158**	**0.842**	**15.903**	**94.955**	**0.142**	**0.858**	**15.518**	**105.963**	**0.129**	**0.871**	**14.423**	**103.313**	**0.141**	**0.859**
	Imp.	-	-	-	46.7%	-	-	-	61.6%	-	-	-	55.5%	-	-	-	51.7%



(a) CNN-based classifier (b) ResNet-based classifier (c) LSTM-based classifier (d) Transformer-based classifier

Fig. 12. MoRF perturbation curves on RadioML2018.01A dataset for (a) CNN, (b) ResNet, (c) LSTM and (d) Transformer-based signal modulation classifiers.

TABLE III
EXPLANATION SENSITIVITY COMPARISON OF DIFFERENT EXPLAINABLE METHODS UNDER RANDOM PERTURBATION AT PSR=-20DB (BEST RESULT IN BOLD, AND SECOND BEST UNDERLINED). '*Dec.*' MEANS THE RELATIVE DECREASE PERCENTAGE OF OUR METHOD AGAINST THE BEST BASELINE.

Dataset	Methods	ResNet-based Classifier		CNN-based Classifier		LSTM-based Classifier		Transformer-based Classifier	
		ΔNAOPC ↓	MS ↓	ΔNAOPC ↓	MS ↓	ΔNAOPC ↓	MS ↓	ΔNAOPC ↓	MS ↓
RadioML 2018.01A	**Grad-CAM**	0.064 ± 0.049	0.056 ± 0.038	0.046 ± 0.032	0.073 ± 0.080	0.036 ± 0.031	0.042 ± 0.028	0.051 ± 0.055	0.054 ± 0.086
	MASK	0.076 ± 0.064	0.634 ± 0.119	0.065 ± 0.086	0.568 ± 0.145	0.048 ± 0.124	0.595 ± 0.189	0.049 ± 0.052	0.568 ± 0.166
	MASE (ours)	**0.022 ± 0.023**	**0.043 ± 0.023**	**0.026 ± 0.029**	**0.059 ± 0.074**	**0.010 ± 0.012**	**0.034 ± 0.048**	**0.011 ± 0.027**	**0.025 ± 0.023**
	Dec.	65.6%	23.2%	43.5%	19.2%	72.2%	19.0%	77.6%	53.7%
RadioML 2016.10A	**Grad-CAM**	0.032 ± 0.05	0.005 ± 0.012	0.030 ± 0.036	0.002 ± 0.004	0.032 ± 0.038	0.007 ± 0.012	0.066 ± 0.085	0.006 ± 0.003
	MASK	0.076 ± 0.097	0.630 ± 0.239	0.097 ± 0.094	0.581 ± 0.263	0.056 ± 0.042	0.535 ± 0.235	0.091 ± 0.082	0.606 ± 0.108
	MASE (ours)	**0.017 ± 0.02**	**0.001 ± 0.001**	**0.016 ± 0.014**	**0.001 ± 0.002**	**0.017 ± 0.023**	**0.003 ± 0.005**	**0.008 ± 0.01**	**0.003 ± 0.006**
	Dec.	46.90%	80.0%	46.7%	50.0%	46.9%	57.1%	87.9%	50.0%

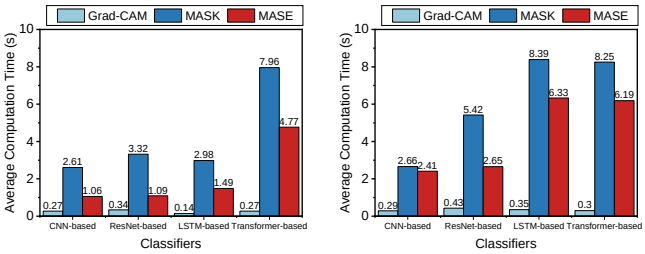
Furthermore, Fig. 12 illustrates MoRF perturbation curves for all methods on RadioML2018.01A dataset. At each step, the most relevant timestep is perturbed based on the derived explanation until the whole timesteps of the signal are perturbed. It is obvious that MASE achieves the fastest decrease in the curve, indicating that the explanations more accurately rank the timesteps of signal by their importance for prediction and thus are more faithful to the classifier, which is consistent with the quantitative results demonstrated in Table II.

2) Explanation Sensitivity Analysis: To investigate the explanation sensitivity, we next compare all methods via the MS and ΔNAOPC metrics, which measure the maximum change in explanations and the difference in explanation faithfulness resulting from random input perturbations. For both datasets, we inject random perturbations with PSR=-20dB to the signal instances, and report the sensitivity performance in Table III. We observe the following phenomena: (a) MASE generally outperforms all baselines, achieving the lowest MS and ΔNAOPC values, showing a superior stability against random perturbations. It is mainly because the key components of MASE, including subsequence-based interpretable signal representation, in-distribution local signal sampling and linear explanation generation, are not significantly affected by minor random perturbations. (b) MASK shows the highest MS to the perturbations, which is probably because there are many randomness and uncertainties in the optimization of saliency perturbations. In comparison, Grad-CAM shows more competitive performance. (c) Despite the noticeable variation in the MS values among Grad-CAM, MASK, and MASE (for instance, with average of 0.056, 0.491, and 0.040 respectively in the RadioML2018.01A dataset), their ΔNAOPC remain consistently small, averaging 0.049, 0.059, and 0.017. This indicates that minor random perturbations do not significantly compromise the faithfulness of the explanations.

3) Explanation Robustness Analysis: To evaluate explanation robustness of MASE and two other baseline methods, we report their MS and ΔNAOPC performance against adversarial perturbation in Table IV. These adversarial perturbations are generated by IFIA method [27] with a PSR=-20dB constraint to ensure impeccability. It is evident to draw the following findings. (a) With the same perturbation impeccability constraint (i.e., PSR=-20dB), the MS and ΔNAOPC values under adversarial perturbation are significantly larger than those under random perturbations. This firstly reveals the vulnerability

TABLE IV
EXPLANATION ROBUSTNESS COMPARISON OF DIFFERENT EXPLAINABLE METHODS UNDER ADVERSARIAL PERTURBATION AT PSR=-20DB (BEST RESULT IN BOLD, AND SECOND BEST UNDERLINED). '*Dec.*' MEANS THE RELATIVE DECREASE PERCENTAGE OF OUR METHOD AGAINST THE BEST BASELINE.

Dataset	Methods	ResNet-based Classifier		CNN-based Classifier		LSTM-based Classifier		Transformer-based Classifier	
		ΔNAOPC ↓	MS ↓	ΔNAOPC ↓	MS ↓	ΔNAOPC ↓	MS ↓	ΔNAOPC ↓	MS ↓
RadioML 2018.01A	**Grad-CAM**	0.386 ± 0.068	0.396 ± 0.053	0.294 ± 0.049	0.706 ± 0.074	0.191 ± 0.037	**0.302 ± 0.037**	0.208 ± 0.034	0.433 ± 0.082
	MASK	0.126 ± 0.014	0.625 ± 0.041	0.197 ± 0.011	0.536 ± 0.091	0.145 ± 0.057	0.72 ± 0.089	0.144 ± 0.009	0.685 ± 0.143
	MASE (ours)	**0.095 ± 0.077**	**0.29 ± 0.04**	**0.123 ± 0.012**	**0.145 ± 0.031**	**0.113 ± 0.061**	0.357 ± 0.132	**0.107 ± 0.012**	**0.154 ± 0.054**
	Dec.	24.6%	26.8%	37.6%	72.9%	22.1%	-18.2%	25.7%	64.6%
RadioML 2016.10A	**Grad-CAM**	0.295 ± 0.075	0.471 ± 0.379	0.196 ± 0.06	0.191 ± 0.244	0.265 ± 0.056	0.297 ± 0.206	0.29 ± 0.025	**0.256 ± 0.15**
	MASK	0.260 ± 0.024	0.689 ± 0.082	0.247 ± 0.057	0.693 ± 0.154	0.203 ± 0.113	0.68 ± 0.082	0.241 ± 0.006	0.715 ± 0.103
	MASE (ours)	**0.127 ± 0.054**	**0.308 ± 0.257**	**0.159 ± 0.07**	**0.183 ± 0.076**	**0.150 ± 0.132**	**0.175 ± 0.166**	**0.148 ± 0.007**	0.308 ± 0.105
	Dec.	51.2%	34.6%	18.9%	4.2%	26.1%	41.1%	38.6%	-20.3%



(a) RadioML2016.10A (b) RadioML2018.01A

Fig. 13. Average Computation Time (ACT) ↓ comparison of different methods on (a) RadioML2016.10A and (b) RadioML2018.01A datasets.

of explainable methods for signal modulation classification against AdvXAI attacks. (b) Across both datasets, our MASE demonstrates superior robustness, evidenced by its lowest MS and ΔNAOPC metrics under adversarial perturbations. Specifically, the MS and ΔNAOPC values of MASE achieve average reductions of 30.6% and 25.7% respectively, compared to the best baseline. We argue that this is because IFIA optimizes adversarial perturbations based on the inherent network structure. Grad-CAM and MASK, as model-specific methods, are closely tied to the network structure to produce explanations, thus are more vulnerable to adversarial perturbations. Conversely, our model-agnostic method, MASE, generates explanations independently of the network structure, thereby showing greater robustness against IFIA.

4) Explanation Efficiency Analysis: In this section, we evaluate the efficiency of the proposed MASE on the two datasets according to the ACT metric. We report the ACT performance in Fig. 13 and have some findings: (a) The gradient-based method (i.e., Grad-CAM) is much faster than other methods, which is mainly attributed to that it only requires a single backpropagation pass. The perturbation-based method MASK seems to take a long time to optimize the saliency perturbation, showing the poorest efficiency among all methods. (b) Though MASE sacrifices efficiency for faithfulness, it still shows high efficiency, with average 2.10s and 4.49s of each explanation for two datasets, respectively. Since the signal sample of RadioML2018.01A is 8 times longer than that of RadioML2016.10A, the average explanation time for a single instance is larger. Besides, in MASE method, the number of generated local samples M is a key adjustable hyper-parameter allowing us to get the desired trade-off between efficiency and faithfulness, as we discussed in Section V-B2.

In conclusion, considering the four quantitative metrics together, MASE considerably outperforms the state-of-the-art alternatives for explaining the signal modulation classifiers.

D. Qualitative Evaluation

In this section, we take a visual examination on derived explanations, showing MASE could provide human interpretable explanations (Section V-D1) and facilitate human to better understand black-box model's decisions (Section V-D2).

1) Visualization of Different Methods: First of all, to evaluate the explanations intuitively, we use the proposed constellation-based visualization mechanism to visualize the explanations derived by different methods in Fig. 14. The grey raw data illustrates the original constellation diagrams of input signals, and the derived explanations for three classifiers are provided on the right. From the plots, we make the following observations, both contributing to the superior human interpretability of the explanations derived by MASE. (a) MASE's explanations are notably more concise than those of baseline methods. For example, in the explanations for the 16APSK modulation sample, MASE highlights fewer signal features, reducing the visual burden for user. This conciseness allows an easier and more effective understanding of the most significant signal features that are relevant to the model's predictions, thereby enhancing interpretability for human users. (b) Explanations produced by MASE display a greater alignment with the domain knowledge of human experts. Take the QPSK modulation sample for instance. There are four reference points in this modulation, each representing a distinct symbol. MASE's explanations effectively highlight signal points proximate to these reference point, closely mirroring the analytical process of human experts who focus on these critical reference points. In contrast, the baseline methods either overlook significant signal features around these key reference point (such as Grad-CAM for CNN-based and LSTM-based classifiers, and MASK for LSTM-based classifier), or highlight an excessive number of irrelevant signal features (such as Grad-CAM for ResNet-based classifier, and MASK for ResNet-based and CNN-based classifiers). This alignment with domain knowledge makes MASE's explanations more intuitive and understandable, particularly for users familiar with this field.

After a close visual inspection on MASE's explanations, we observe that although the CNN-based, ResNet-based

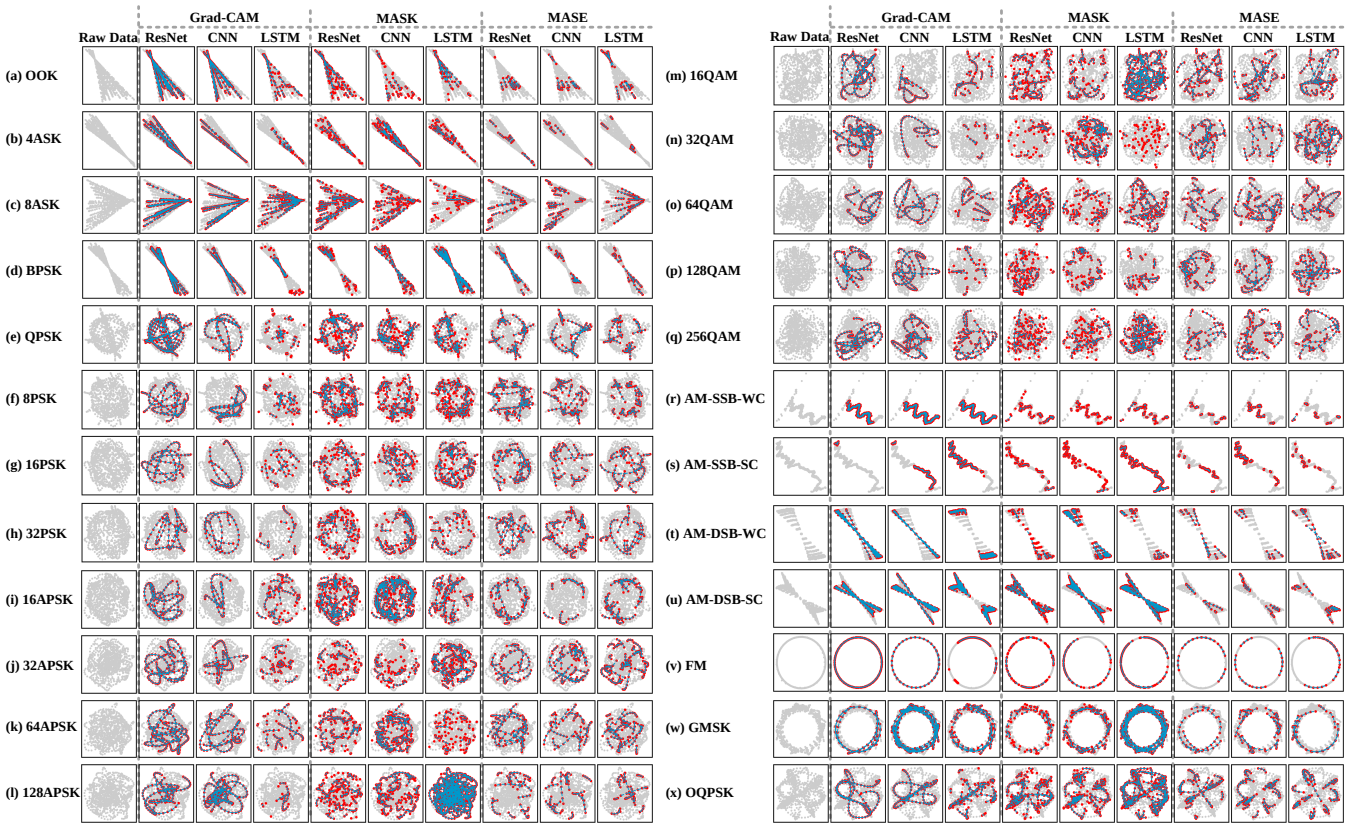
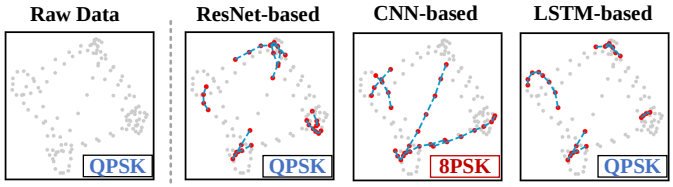
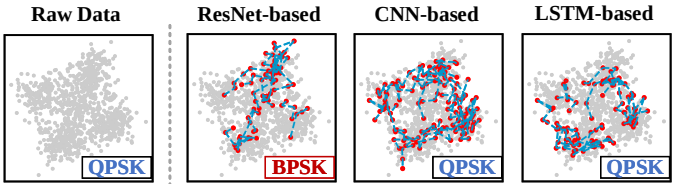


Fig. 14. Explanation visualizations for different modulation types and different classifiers under RadioML 2018.01A dataset.



(a) RadioML2016.10A



(b) RadioML2018.01A

Fig. 15. Visualizing MASE's explanations of the misclassified signal samples.

and LSTM-based modulation classifiers have different model structures and input formats, the signal features relevant to decisions of the three models are generally similar. Note that, with the highlighted signal features in the explanation visualization, different modulation types are easily differentiated.

2) Visualization for Misclassifications: Moreover, we examine instances where the models incorrectly classify modulation types, trying to understand the reasons of model failures by explanations derived by MASE. Two QPSK modulated signal samples from two datasets are shown in Fig. 15, and their corresponding predicted labels are marked in blue (red) for correct (wrong) classification.

The first signal sample is correctly classified by ResNet-based and LSTM-based classifiers as QPSK, as they capture the signal features around four modulation reference points, as the explanation visualizations show. However, the CNN-based classifier fails to capture these relevant features and misclassifies it as the 8PSK modulation. Similarly, for the second signal, the ResNet-based classifier misclassifies it as the BPSK modulation while the other two classifiers successfully classified. The derived explanation visualizations indicate the possible reason for misclassification is that the ResNet-based classifier only concentrates on signal points around two out of the four modulation reference points, whose captured signal features are similar to the BPSK modulation. In comparison, the other two classifiers successfully capture the signal features close to four modulation reference points.

By explaining the reason of model successes and failures, we show that MASE allows human to better understand black-box model decisions. To a certain extent, it could not only facilitate the model reliability and trustworthiness, but also provide insights on model improvements [48].

VI. Related Work

The subsequent literature review provides an overview of the relevant related work focusing on: the signal modulation classification using DL-based methods, and the application of XAI methods for DL-based signal modulation classification.

A. Deep Learning-based Signal Modulation Classification

The rapid advances in DL have facilitated the development of high-performance DL-based signal modulation classifiers [9], [49]. By building a simple four-layer CNN model with I/Q data as inputs, O'Shea *et al.* [50] firstly explore the potential of deep neural networks in signal modulation classification, achieving higher classification accuracy than the expert features-based methods. Afterwards, more advanced DL models with strong feature extraction ability are adopted for signal modulation classification. For example, inspired by the skip connection structures, ResNet-based signal modulation classifiers are proposed, showing better classification accuracy [29], [51]. Considering the temporal correlation features of signals, a few novel RNN-based signal modulation classifiers have been investigated. With the amplitude and phase sequences transformed from I/Q signals as the input, Rajendran *et al.* [32] show an LSTM model is able to achieve high classification accuracy and solve the problem of variable signal length. Besides the sequence representation, other forms of signal representations like image and graph representation [52], [53] are also explored as input to the DL model. Recently, based on attention mechanism, Transformer-based models are able to learn more discriminating features, leading to a potential breakthrough in signal modulation classification [30].

However, while deep learning techniques have continued to provide state-of-the-art performance, one of the primary challenges that stands to hinder this progress is the opaque nature of these complex models. For example, it is difficult to access visually and understand the features extracted by a DL model, aggravating the model's untrustworthiness and unreliability in real-world deployment. To mitigate the issue, in this paper, we study an explainable method for signal modulation classification, which provides human-understandable explanations for model predictions by visualizing the important input signal features extracted by the classifier.

B. Explanations for Signal Modulation Classification

Most prior works have focused on explaining computer vision models, which derives a saliency map to highlight the features in an input that are relevant for a model to issue a prediction, including gradient-based methods [54], perturbation-based methods [17] and surrogate-based methods [55]. In comparison, the explanations for DL-based signal modulation classifiers have not fully studied yet. Prior to our work, Huang *et al.* [19] make the first attempt to visualize the DL-based signal modulation classifiers, where the gradient-based XAI method, Grad-CAM [21] and the perturbation-based XAI method, MASK [22] are adopted to visualize the signal features extracted by CNN-based and LSTM-based classifiers, respectively. Moreover, Chen *et al.* [20] propose a feature explainable signal modulation classification, which uses Grad-CAM method to visualize and compare the hidden layer features extracted by different classifiers including AlexNet, VGG16 and ResNet.

However, the above explanations for signal modulation classification are mainly derived by model-specific XAI methods with white-box access to the classifiers to be explained. In comparison, the proposed MASE is model-agnostic, offering greater flexibility and applicability to any black-box classifiers. Additionally, this paper presents a quantitative framework for evaluating signal modulation classifiers' explanations, addressing the subjective visualization judgment in existing literature.

VII. Conclusion and Future Work

In this paper, we propose MASE, a novel model-agnostic explainable method for DL-based signal modulation classifiers, which is free from model access requirements and could be applied to any classifiers to provide black-box explainability. By training a local linear surrogate model over the subsequence-based interpretable signal representation, MASE derives a class activation vector as the explanation, in which each element indicates the importance value of each signal timestep for model prediction. Besides, constellation diagrams with activation threshold are adopted for providing concise and understandable explanation visualization for human. Additionally, we introduce a generic quantitative explanation evaluation framework for benchmarking explainable methods in signal modulation classification context, from the aspects of faithfulness, sensitivity, robustness and efficiency. In the experiments, we quantitatively and qualitatively compare MASE to two state-of-the-art explainable methods for four DL-based classifiers on two well-known datasets. The results demonstrate that MASE is (1) more faithful to model's behaviors, (2) less sensitive to random noise perturbation, (3) more robust against adversarial XAI attacks, and (4) able to allow human to better understand black-box model decisions.

Therefore, we believe that the MASE method will help to enable the widespread application of complex DL-based signal modulation classifiers in both civilian and military wireless systems, by providing faithful and comprehensive explanations to the users, increasing the trustworthiness and reliability of the opaque models. It is expected to inspire a series of follow-up studies, including but not limited to (1) comparison with newer XAI methods, (2) global explanations with global surroagets, and (3) model improvements based on explanations.

References

[1] F. A. Bhatti, M. J. Khan, A. Selim, and F. Paisana, "Shared spectrum monitoring using deep learning," *IEEE Transactions on Cognitive Communications and Networking*, vol. 7, no. 4, pp. 1171–1185, 2021.

[2] I. Kakalou, K. E. Psannis, P. Krawiec, and R. Badea, "Cognitive radio network and network service chaining toward 5g: Challenges and requirements," *IEEE Communications Magazine*, vol. 55, no. 11, pp. 145–151, 2017.

[3] Y. Lin, H. Zhao, X. Ma, Y. Tu, and M. Wang, "Adversarial attacks in modulation recognition with convolutional neural networks," *IEEE Transactions on Reliability*, vol. 70, no. 1, pp. 389–401, 2021.

[4] O. Dobre, "Survey of automatic modulation classification techniques: classical approaches and new trends," *IET Communications*, vol. 1, pp. 137–156, April 2007.

[5] Q. Shi and Y. Karasawa, "Automatic modulation identification based on the probability density function of signal phase," *IEEE Transactions on Communications*, vol. 60, no. 4, pp. 1033–1044, 2012.

[6] H.-C. Wu, M. Saquib, and Z. Yun, "Novel automatic modulation classification using cumulant features for communications via multipath channels," *IEEE Transactions on Wireless Communications*, vol. 7, no. 8, pp. 3098–3105, 2008.

[7] S. Peng, S. Sun, and Y.-D. Yao, "A survey of modulation classification using deep learning: Signal representation and data preprocessing," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 33, no. 12, pp. 7020–7038, 2022.

[8] W. Xiao, Z. Luo, and Q. Hu, "A review of research on signal modulation recognition based on deep learning," *Electronics*, vol. 11, no. 17, pp. 1–29, 2022.

[9] F. Zhang, C. Luo, J. Xu, Y. Luo, and F.-C. Zheng, "Deep learning based automatic modulation recognition: Models, datasets, and challenges," *Digital Signal Processing*, vol. 129, p. 103650, 2022.

[10] P. Qi, T. Jiang, L. Wang, X. Yuan, and Z. Li, "Detection tolerant black-box adversarial attack against automatic modulation classification with deep learning," *IEEE Transactions on Reliability*, vol. 71, no. 2, pp. 674–686, 2022.

[11] B. Wang, K. Xu, S. Zheng, H. Zhou, and Y. Liu, "A deep learning-based intelligent receiver for improving the reliability of the mimo wireless communication system," *IEEE Transactions on Reliability*, vol. 71, no. 2, pp. 1104–1115, 2022.

[12] R. Dwivedi, D. Dave, H. Naik, S. Singhal, R. Omer, P. Patel, B. Qian, Z. Wen, T. Shah, G. Morgan, and R. Ranjan, "Explainable ai (xai): Core ideas, techniques, and solutions," *ACM Computing Surveys*, vol. 55, no. 9, pp. 1–33, 2023.

[13] M. Nourani, S. Kabir, S. Mohseni, and E. D. Ragan, "The effects of meaningful and meaningless explanations on trust and perceived system accuracy in intelligent systems," in *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, vol. 7, 2019, pp. 97–105.

[14] G. Novakovsky, N. Dexter, M. W. Libbrecht, W. W. Wasserman, and S. Mostafavi, "Obtaining genetics insights from deep learning via explainable artificial intelligence," *Nature Reviews Genetics*, pp. 1–13, 2022.

[15] S. Letzgus, P. Wagner, J. Lederer, W. Samek, K.-R. Müller, and G. Montavon, "Toward explainable artificial intelligence for regression models: A methodological perspective," *IEEE Signal Processing Magazine*, vol. 39, no. 4, pp. 40–58, 2022.

[16] D. Minh, H. X. Wang, Y. F. Li, and T. N. Nguyen, "Explainable artificial intelligence: a comprehensive review," *Artificial Intelligence Review*, vol. 55, no. 5, pp. 3503–3568, 2022.

[17] T. Fel, M. Ducoffe, D. Vigouroux, R. Cadène, M. Capelle, C. Nicodème, and T. Serre, "Don't lie to me! robust and efficient explainability with verified perturbation analysis," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023, pp. 16 153–16 163.

[18] R. Dwivedi, R. Kumar, D. Chopra, P. Kothari, and M. Singh, "An efficient ensemble explainable ai (xai) approach for morphed face detection," *arXiv preprint arXiv:2304.14509*, 2023.

[19] L. Huang, Y. Zhang, W. Pan, J. Chen, L. P. Qian, and Y. Wu, "Visualizing deep learning-based radio modulation classifier," *IEEE Transactions on Cognitive Communications and Networking*, vol. 7, no. 1, pp. 47–58, 2021.

[20] J. Chen, S. Miao, H. Zheng, and S. Zheng, "Feature explainable deep classification for signal modulation recognition," in *IECON 2020 The 46th Annual Conference of the IEEE Industrial Electronics Society*, 2020, pp. 3543–3548.

[21] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-cam: Visual explanations from deep networks via gradient-based localization," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2017, pp. 618–626.

[22] R. C. Fong and A. Vedaldi, "Interpretable explanations of black boxes by meaningful perturbation," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2017, pp. 3429–3437.

[23] J. Bento, P. Saleiro, A. F. Cruz, M. A. Figueiredo, and P. Bizarro, "Timeshap: Explaining recurrent models through sequence perturbations," in *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, 2021, pp. 2565–2573.

[24] M. T. Ribeiro, S. Singh, and C. Guestrin, "Why should i trust you?: Explaining the predictions of any classifier," in *Proceedings of the 22nd ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, 2016, pp. 1135–1144.

[25] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in *Advances in Neural Information Processing Systems*, vol. 30, 2017, pp. 1–10.

[26] M. T. Ribeiro, S. Singh, and C. Guestrin, "Anchors: High-precision model-agnostic explanations," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 32, no. 1, 2018, pp. 1–9.

[27] A. Ghorbani, A. Abid, and J. Zou, "Interpretation of neural networks is fragile," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, no. 01, Jul. 2019, pp. 3681–3688.

[28] T. J. O'shea and N. West, "Radio machine learning dataset generation with gnu radio," in *Proceedings of the GNU Radio Conference*, vol. 1, no. 1, 2016, pp. 1–6.

[29] T. J. O'Shea, T. Roy, and T. C. Clancy, "Over-the-air deep learning based radio signal classification," *IEEE Journal of Selected Topics in Signal Processing*, vol. 12, no. 1, pp. 168–179, 2018.

[30] J. Cai, F. Gan, X. Cao, and W. Liu, "Signal modulation classification based on the transformer network," *IEEE Transactions on Cognitive Communications and Networking*, vol. 8, no. 3, pp. 1348–1357, 2022.

[31] S. Lin, Y. Zeng, and Y. Gong, "Modulation recognition using signal enhancement and multistage attention mechanism," *IEEE Transactions on Wireless Communications*, vol. 21, no. 11, pp. 9921–9935, 2022.

[32] S. Rajendran, W. Meert, D. Giustiniano, V. Lenders, and S. Pollin, "Deep learning models for wireless signal classification with distributed low-cost spectrum sensors," *IEEE Transactions on Cognitive Communications and Networking*, vol. 4, no. 3, pp. 433–445, 2018.

[33] D. Garreau and D. Mardaoui, "What does lime really see in images?" in *Proceedings of the 38th International Conference on Machine Learning*, vol. 139, 2021, pp. 3620–3629.

[34] P. Schäfer, A. Ermshaus, and U. Leser, "Clasp-time series segmentation," in *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*, 2021, pp. 1578–1587.

[35] R. Doddaiah, P. Parvatharaju, E. Rundensteiner, and T. Hartvigsen, "Class-specific explainability for deep time series classifiers," in *IEEE International Conference on Data Mining (ICDM)*, 2022, pp. 101–110.

[36] Y. Mao, Y.-Y. Dong, T. Sun, X. Rao, and C.-X. Dong, "Attentive siamese networks for automatic modulation classification based on multitiming constellation diagrams," *IEEE Transactions on Neural Networks and Learning Systems*, pp. 1–15, 2021.

[37] W. Samek, A. Binder, G. Montavon, S. Lapuschkin, and K.-R. Müller, "Evaluating the visualization of what a deep neural network has learned," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 28, no. 11, pp. 2660–2673, 2017.

[38] R. Tomsett, D. Harborne, S. Chakraborty, P. Gurram, and A. Preece, "Sanity checks for saliency metrics," in *Proceedings of the AAAI conference on Artificial Intelligence*, vol. 34, no. 04, 2020, pp. 6021–6029.

[39] A. Apicella, S. Giugliano, F. Isgrò, and R. Prevete, "Exploiting auto-encoders and segmentation methods for middle-level explanations of image classification systems," *Knowledge-Based Systems*, vol. 255, p. 109725, 2022.

[40] Y. Yang, J. Qiu, M. Song, D. Tao, and X. Wang, "Learning propagation rules for attribution map generation," in *European Conference on Computer Vision*, 2020, pp. 672–688.

[41] I. Kakogeorgiou and K. Karantzalos, "Evaluating explainable artificial intelligence methods for multi-label deep learning classification tasks in remote sensing," *International Journal of Applied Earth Observation and Geoinformation*, vol. 103, p. 102520, 2021.

[42] C.-K. Yeh, C.-Y. Hsieh, A. Suggala, D. I. Inouye, and P. K. Ravikumar, "On the (in)fidelity and sensitivity of explanations," in *Advances in Neural Information Processing Systems*, vol. 32, 2019, pp. 1–12.

[43] W. Jin, X. Li, M. Fatehi, and G. Hamarneh, "Guidelines and evaluation of clinical explainable ai in medical image analysis," *Medical image analysis*, vol. 84, p. 102684, February 2023.

[44] B. Flowers, R. M. Buehrer, and W. C. Headley, "Evaluating adversarial evasion attacks in the context of wireless communications," *IEEE Transactions on Information Forensics and Security*, vol. 15, pp. 1102–1113, 2020.

[45] A. Bahramali, M. Nasr, A. Houmansadr, D. Goeckel, and D. Towsley, "Robust adversarial attacks against dnn-based wireless communication systems," in *Proceedings of the 2021 ACM SIGSAC Conference on Computer and Communications Security*, 2021, pp. 126–140.

[46] M. Noppel and C. Wressnegger, "Sok: Explainable machine learning in adversarial environments," in *IEEE Symposium on Security and Privacy (SP)*, 2024, pp. 1–19.

[47] M. Abadi, P. Barham, J. Chen, Z. Chen, A. Davis, J. Dean, M. Devin, S. Ghemawat, G. Irving, M. Isard *et al.*, "Tensorflow: a system for large-scale machine learning," in *12th USENIX Symposium on Operating Systems Design and Implementation*, 2016, pp. 265–283.

[48] L. Weber, S. Lapuschkin, A. Binder, and W. Samek, "Beyond explaining: Opportunities and challenges of xai-based model improvement," *Information Fusion*, vol. 92, pp. 154–176, 2023.

[49] M. Wang, Y. Lin, Q. Tian, and G. Si, "Transfer learning promotes 6g wireless communications: Recent advances and future challenges," *IEEE Transactions on Reliability*, vol. 70, no. 2, pp. 790–807, 2021.

[50] T. J. O'Shea, J. Corgan, and T. C. Clancy, "Convolutional radio modulation recognition networks," in *Engineering Applications of Neural Networks*, 2016, pp. 213–226.

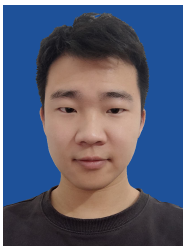
[51] X. Liu, D. Yang, and A. E. Gamal, "Deep neural network architectures for modulation classification," in *51st Asilomar Conference on Signals, Systems, and Computers*, 2017, pp. 915–919.
[52] Y. Liu, Y. Liu, and C. Yang, "Modulation recognition with graph convolutional network," *IEEE Wireless Communications Letters*, vol. 9, no. 5, pp. 624–627, 2020.
[53] Q. Xuan, J. Zhou, K. Qiu, Z. Chen, D. Xu, S. Zheng, and X. Yang, "Avgnet: Adaptive visibility graph neural network and its application in modulation classification," *IEEE Transactions on Network Science and Engineering*, vol. 9, no. 3, pp. 1516–1526, 2022.
[54] I. E. Nielsen, D. Dera, G. Rasool, R. P. Ramachandran, and N. C. Bouaynaya, "Robust explainability: A tutorial on gradient-based attribution methods for deep neural networks," *IEEE Signal Processing Magazine*, vol. 39, no. 4, pp. 73–84, 2022.
[55] H. Tan and H. Kotthaus, "Surrogate model-based explainability methods for point cloud nns," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 2022, pp. 2239–2248.

Yunzhe Tian received his M.S. degree from Beijing Jiaotong University in 2022. He is currently a Ph.D. candidate of cyber security in Beijing Jiaotong University. His main research interests are interpretable machine learning and AI security.

Dongyue Xu received the bachelor's degree from Communication University of China in 2022. She is currently pursuing the graduate degree with Beijing Jiaotong University. Her research interests are AI security and adversarial machine learning.

Endong Tong (Member, IEEE) received the Ph.D. degree from the Chinese Academy of Sciences, Beijing, China, in 2013. He is currently an Assistant Professor with Beijing Jiaotong University. He has published more than 30 research papers in refereed international conferences and journals. His current research interests include AI security, services computing, and data mining.

Run Sui received the bachelor's degree from University of Jinan in 2020. He is currently pursuing the graduate degree with Beijing Jiaotong University. His research interests are machine learning and signal modulation recognition.



Kang Chen received a bachelor's degree from Zhengzhou University in 2020. He is currently pursuing the graduate degree with Beijing Jiaotong University. His research interests are network security and AI security.

Yike Li received bachelor's degree from Hefei University of Technology in 2018. She is currently a Ph.D. candidate of cyber security in Beijing Jiaotong University. Her current research interests include AI security and reinforcement learning.

Thar Baker (Senior Member, IEEE) received his Ph.D. degree in autonomic cloud applications from Liverpool John Moores University (LJMU) in the UK in 2010 and became a Senior Fellow of Higher Education Academy in 2018. He is currently a Professor with The University of Brighton, UK. He was a Reader of cloud engineering and the Head of the Applied Computing Research Group in LJMU between 2013-2020, then in University of Sharjah in UAE between 2020-2022. He has published numerous refereed research papers in multidisciplinary research areas, including parallel and distributed computing, algorithm design, green and sustainable computing, and energy routing protocols.

Wenjia Niu received the bachelor's degree in computer science from Beijing Jiaotong University in 2005, and the Ph.D. degree in computer science from the Chinese Academy of Sciences in 2010. He is currently a Professor with Beijing Jiaotong University. His research interests are AI security, agent, and data mining.

Jiqiang Liu (Senior Member, IEEE) received the B.S. and Ph.D. degrees from Beijing Normal University in 1994 and 1999, respectively. He is currently a Professor with the School of Computer and Information Technology, Beijing Jiaotong University. He has published over 100 scientific papers in various journals and international conferences. His main research interests are trusted computing, cryptographic protocols, privacy preserving, and network security.