Xu *et al. Journal of Cloud Computing* (2024) 13:10
https://doi.org/10.1186/s13677-024-00590-3

Journal of Cloud Computing:
Advances, Systems and Applications

RESEARCH

Open Access

Towards explainability for AI-based edge wireless signal automatic modulation classification

Bo Xu[1], Uzair Aslam Bhatti[1], Hao Tang[1*], Jialin Yan[1], Shulei Wu[5], Nadia Sarhan[2], Emad Mahrous Awwad[3], Syam M. S.[6] and Yazeed Yasin Ghadi[4]

Abstract

With the development of artificial intelligence technology and edge computing technology, deep learning-based automatic modulation classification (AI-based AMC) deployed at edge devices using centralised or distributed learning methods for optimisation has emerged in recent years, and has made great progress in the recognition accuracy and recognisable range of wireless signals. However, the lack of sufficient explanation of these models leads to low accuracy and training efficiency of model training, and their applications and further improvements are limited. Researchers have started to propose interpretable methods for technical analysis of deep learning-based AMC. In this paper, based on the research and application development of interpretable methods in recent years, we review the applicable methods and existing research challenges of interpretable automatic modulation classification. And an interpretable AI-based automatic modulation classification framework is proposed to map the interpretability of automatic modulation classification results by obtaining the contribution of wireless signal features to deep learning network training. Experimental results show that the proposed method possesses the ability to explore the classification mechanism of non-transparent auto-modulated classification networks and has the potential to help edge devices train networks with lower energy consumption and higher accuracy.

Keywords Automatic modulation classification, Explainable methods, Deep learning

*Correspondence:
Hao Tang
melineth@hainanu.edu.cn
[1] School of Information and Communication Engineering, Hainan University, 58 People's Avenue, Haikou, Hainan 570228, China
[2] Department of Quantitative Analysis, College of Business Administration, King Saud University, Riyadh, Saudi Arabia
[3] Department of Electrical Engineering, College of Engineering, King Saud University, P.O. Box 800, Riyadh 11421, Saudi Arabia
[4] Department of Computer Science, Al Ain University, Al Ain, United Arab Emirates
[5] School of Information Science and Technology, Hainan Normal University, Haikou 571158, People's Republic of China
[6] Guangdong-Hong Kong-Macao GBA New Generation Intelligent IoT Research Center, Shenzhen University, Shenzhen, Guangdong 518060, China

Introduction

Pattern recognition has made a lot of progress and gained extensive applications in the field of computer vision [1, 2]. The recognition of unknown wireless signals can be regarded as an important branch of pattern recognition [3]. A series of results have been reported in a range of open researches, which are regarded as AI-based AMC. These works construct various neural network-based AMC models with different model structures and parameters. Compared to traditional methods, AI-based AMC simplifies the conditional assumption of the model and provides a noticeable improvement in classification accuracy and ranges.

However, most AI-based AMC models work as black boxes, which lack interpretation of the model and classification behavior from the perspective of neural networks

and wireless communication. Though these AI-based AMCs achieve more scalable and higher accuracy performance by relying on sophisticated machine-learning classification models trained on massive datasets, them are risk creating and using decision systems that them do not really understand. The untransparent decision systems impacts not only information on ethics but also on accountability [4], on safety [5], and on industrial liability [6]. The explainable AI are widely applied in various scenarios such as signal analysis [7, 8], explainable Robotic Systems [9], and transportation [10, 11]. The interpretation for AI model is also significantly important in wireless communication, i.e., the development of 6G intelligent systems [12].

Due to the limitation in interpretation of AI-based AMC, it is not easy to get knowledge the failure in practice. Even though the failure is determined, the reason for the failure is still unknown. The failure possibly lay on parts of the input or lay on the structure of the AI model. Hence, such problem in non-transparency decrease the performance of modulation classification, especially in the conditions where the signal quality is poor. Therefore, it is still a hard task to solve the specific classification failure and improve the performance of the model in some specific classification. On the other hand, due to the problem in non-transparency, the confidence of the modulation classification result decreases. We are unable to determine whether it is a right classification or a totally wrong classification. So the application of AI-based AMC is limited, especially the sensitive area such as military application, or it may lead to significant loss for the user.

To solve the problem, different scientific communities study the problem of explaining machine-learning decision models [13, 14]. Each community addresses the issue from a different perspective and provides a different meaning to explanation. Such research aimed to explain AI models are regarded as explainable AI (XAI), which make it possible to understand how a decision is made in a dark box [15, 16].

In this paper, we present a novel approach by integrating the concept of attribution to investigate the process of AI-based AMC models in the domain of wireless communication. Our objective is to enhance and optimize these models effectively.By employing this attribution score, we establish a mapping to wireless features that are indicative of specific modulation schemes. This mapping allows us to gain a deeper understanding of how the AI-based AMC model utilizes these features in its decision-making process. By adopting the attribution framework, we provide a more comprehensive and explainable analysis of AI-based AMC models, contributing to their optimization and overall performance improvement in the wireless communication domain. The contributions of this work are summarized as follows:

- Conceptually, We are the first to construct a relatively general interpretation evaluation mechanism for AI-based AMC, exploring the insight of the nontransparent AMC models.
- Technically, we design an attribution-driven explainable AI model for AI-based AMC, which provides the interpretation in the artificial intelligence domain and wireless signal domain. Additional attribution-based AMC model are proposed to improve the performance of current AI-based AMC model.

AMC related works

Likelihood-based AMC

The likelihood-based AMC (LB-AMC) methods are based on the likelihood function of the received signal, which classification by comparing the likelihood ratio with the predefined threshold value [17]. Typically, the LB-AMC are regarded as a multiple hypothesis testing problem whose amount is equal to the number of the modulation types [18]. The probability density functions of the received signal are computed for all hypothesis to determine the modulation types, where Average Likelihood Ratio Test (ALRT) [19, 20], Generalized Likelihood Ratio Test (GLRT) [21], and Hybrid Likelihood Ratio Test (HLRT) [22–24] are adopted.

In [25], a fast recursive algorithm based on the likelihood function and approximations thereof is proposed for the classification of MPSK modulations in white Gaussian noise. And this approach provides a general framework for interpreting previously known structures and creating new ones. In [26], a general maximum likelihood classifier for the linear modulation is proposed. Through the inquiry of maximum likelihood theory, an inference is drawn that the likelihood function of an observation given a reference can closely approximated by a measure of the correlation between empirical and true temporal higher-order moment functions. On the assumption that all signal parameters as well as the noise power, independent data symbols, and rectangular pulse shape are determined, the work in [27] aims to develop a theoretical performance analysis of the generic maximum likelihood classifier to adapt any digital amplitude-phase modulation. To deal with the signal recognition of the orthogonal frequency division multiplexing (OFDM) systems in wireless time division duplex, an automatic modulation classification algorithm based on maximum Likelihood is presented in [28]. In contrast to the signaling-free adaptive modulation technique, the complexity of the model is reduced significantly.

Feature-based AMC

For feature-based AMC (FB-AMC), the modulation format is determined by the observed values of the features. The statistical moments of the signal phase are utilized to automatically classify the modulation types of general M-ary PSK signals in [29]. The theoretical foundation of this approach is that for M-ary PSK signals, the n-th moment (n even) of the phase of the signal is a monotonic increasing function of M. In [30], a simple, low complexity, robust method based on fourth-order cumulants is proposed to classify various digital signaling formats, which is particularly effective for discriminating format subclasses, such as PSK versus PAM versus QAM. This work shows the effectiveness of cumulant-based classification in the hierarchical scheme. To adapt to various channels, including the AWGN channel with unknown phase and the OFDM channel, and the channel with unknown phase and frequency offsets, as well as the non-Gaussian noise channel, the researchers propose a modulation classification approach based on the Kolmogorov-Smirnov (K-S) test [31], which is a non-parametric statistical method to measure the goodness of fit.

AI-based AMC

Two algorithms for analog and digital modulation classification are presented in [32], the first of which utilizes the decision-theoretic approach, and the second of which utilizes the artificial neural network (ANN). Given the evaluation result in this work, both algorithms achieve great performance in the classification of different types of band-limited analog and digitally modulated signals corrupted by band-limited Gaussian noise sequences. To automatically extract features from the long symbol-rate observation sequence along with the estimated SNR, an end to end convolution neural network based automatic modulation classification (CNN-AMC) is proposed in [33]. The CNN-AMC is proved to outperform the conventional feature-based method and obtain a closer approximation to the optimal ML-AMCs. In [34], the researchers provide an extensive dataset of additional radio signal types, a realistic simulation of the wireless propagation environment, over-the-air (OTA) measurement of the new dataset. And the solution in [34] provides an in-depth analysis of many practical engineering design and system parameters that impact the performance and accuracy of the radio signal classifier. New data formats such as gridlike topologies (e.g., images) to represent modulated signals are introduced in [35], and two convolutional neural network based deep learning models, AlexNet and GoogLeNet are applied. The new data formats facilitates the use of prevalent DL network models and frameworks for classification, and the CNN based approach achieves significantly improved performance. Given both high computing cost and large model sizes in deployment of the conventional deep learning-based methods, the researchers propose a deep learning-based lightweight automatic modulation classification method with small model sizes and faster computational speed [36]. The key idea is to enforce scaling factors sparsity via compressive sensing with slight performance loss. Yang et al. [37] proposes a distributed automatic modulated signal classification method based on lightweight networks deployed at the mobile edge computing end, which utilises multiple edge devices to train a global model and share the model weights, which can reduce the communication overhead of distributed learning due to repeated transmission of weight information, while ensuring the classification performance.

In recent years, AI technology has developed rapidly, and it has a wide range of applications in many civil or military scenarios [16, 38]. Although AI has applications in many fields, there are still inherent unknown risks in many application scenarios, such as applications involving medical health, currency, and autonomous driving [39]. The powerful processing of AI is generally considered to be a black box, and there is no reason why this decision can actually be achieved. Due to the lack of transparency and interpretation, the decisions of many AI systems are not reliable or credible, thus limiting the application of AI in sensitive fields. Recently, more and more research has been devoted to explainable AI (XAI), which aims to elucidate the inner principles of the decisions made by AI, and some studies have proved to be effective. The existing XAI methods are classified into four categories: explaining with surrogates, explaining with local perturbations, propagation-based approaches, and meta-explanations [40]. The method explaining with surrogates approximates a complex model using an explainable surrogate function, by sampling near the input of interest, evaluating the neural network at those points, and trying to fit the surrogate function if the input domain of the surrogate function is explainable, then the decision model can be explained [41, 42]. The method explaining with local perturbations is to achieve interpretation by analyzing the model's response to local changes including exploiting gradient information as well as perturbation and optimization based methods, where different degree of influence on prediction from the local perturbations reflect the part that the neural network pays attention to [43–46].

Here, we focus on related work in the field of wireless signal modulation identification using explainable methods. Reference [47] provides explainability for signal modulation recognition by extracting hidden layer features of signals and mapping them to original signal

features. In reference [48], the class activation vector visualization AMC is introduced. Numerical analysis results show that the wireless features learned by the classifier based on CNN and LSTM are somewhat similar to the knowledge of human experts, but short radio samples will lead to low classification accuracy. Reference [49] proposes a feature visualization method using signal clustering, which verifies the feature accuracy extracted by clustering model by visualizing the significant features of different modulated signals. Reference [50] also introduces the class activation gradient to obtain the characteristic heat map of the input signal, and provides a method of explainability measurement. Although the above methods are able to explain AI models, since these methods focus on extracting features through the model and mapping them to the original signal, they are still based on human related experience. Therefore, existing AI-based AMC explanation methods have limited explanation capabilities. In practice, such methods cannot make good explanations in the field of wireless signal processing, which further limits wireless signal processing researchers' improvement of AMC and the reliability of AMC.

Explainability methods

The classification of explainability methods

Based on the duration of the explainability methods and the scope of their effects, we can categorize explainable methods in machine learning into: local explainability and global explainability, self-explainability, and post hoc explainability, as shown in Fig. 1.

1. Local explanation/Global explanation

Local interpretation emphasizes understanding the model's prediction or decision-making process for specific individuals or samples. It focuses on explaining the behavior of the model on a given input or a small group of inputs.In this case, the model can be seen as a black box, without considering the complexity of the model. From a single sample perspective, the predicted values provided by the model may have a linear or even monotonic relationship with certain features. Therefore, local explanation may be more accurate than global explanation.

Global interpretation focuses on understanding the overall behavior, structure, and decision logic of the entire model. It is concerned with the holistic nature of the model rather than the predictive outcomes of specific samples or inputs.The explanation at this level refers to how the model makes decisions based on the entire feature space, model structure, parameters, etc. What features are important and what happens when feature interactions occur. The global explanation of the model can help understand what the distribution of the target variable is for different features.

2. Self-explanation/Post hoc explanation

Self-explanation refers to the ability of a model to automatically provide an explanation for its behavior when making decisions or generating predictions. This type of explanation is typically an integral part of the model itself and does not require additional explanatory steps.

Post hoc explanation refers to the process of explaining the behavior of a model after it has already made predictions or decisions. This explanation usually takes place after the model's output is available, and users or developers need to understand the reasons behind the model's decisions through additional explanatory steps.

The types of explainability methods

1. The gradient-based feature attribution method

To obtain the interpretation of a given AI-based AMC model, we aim to determine the attribution feature of the input. Since we know how an input lead to an output from Eq. 8, we utilize the gradient of the input as our

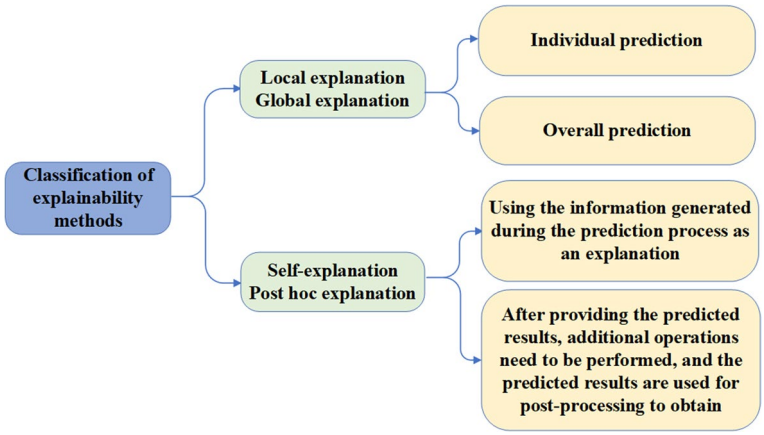

Fig. 1 Classification of explainability methods

measure of the attribution. To quantify the attribution, we define the attribution as a numerical score for each input sample by normalization processing. Specifically, we assign G_i to present the attribution of the ith input sample, which is computing as follows

$$G_i = \frac{\partial O_j}{\partial I_i} \tag{1}$$

where O_j represents the jth output of the neural network which is determined by a specific classification result. From Eqs. 1 and 8, we are able to obtain the paths impact on input sample I_i from G_i. Because G_i is a numerical score, it reflects the contribution degree of paths to the output which I_i goes through. Then for a input of length n and its output O_j, the attributions of the whole input is represented as

$$G = \left[\frac{\partial O_j}{\partial I_0}, ..., \frac{\partial O_j}{\partial I_i}, ..., \frac{\partial O_j}{\partial I_n}\right] \tag{2}$$

With Eq. 2, we obtain the contribution degree distribution of the input samples, i.e., attributions distribution.

Gradient-based attribution methods have the problem of gradient saturation [51]. As shown in Fig. 2, as the input value increases, the output value is large, but the change in output is slow. This means that simply using the gradient to determine the contribution of the input to the output is inaccurate, and many activated neurons might not be marked as highly contributive. Based on this, Integrated Gradients (IG) defines a straight line from the baseline to the input as the path, summing up the gradients at all points on this path as the integrated gradient for the input. The integrated gradient for the ith dimension is expressed as:

$$IG(x_i) = \left(x_i - x_i^{'}\right) \times \sum_{j=1}^{N} \frac{\partial F(x_{ij})}{\partial x_i} \tag{3}$$

Where x_i is the ith dimension of the input x, $x_i^{'}$ is its baseline, and x_{ij} is the jth point on the ith dimension of the line from the baseline to the input. For certain inputs x_i , even if their gradient may be zero, they could have a significant impact on the model output. This is because specific input features might have complex nonlinear relationships with other features, leading to a zero gradient value when considered in isolation, but still having an important influence on the output. In such cases, merely relying on gradient values might overlook the contributions of these features. Therefore, using the integrated gradient method can reveal the true contribution of these

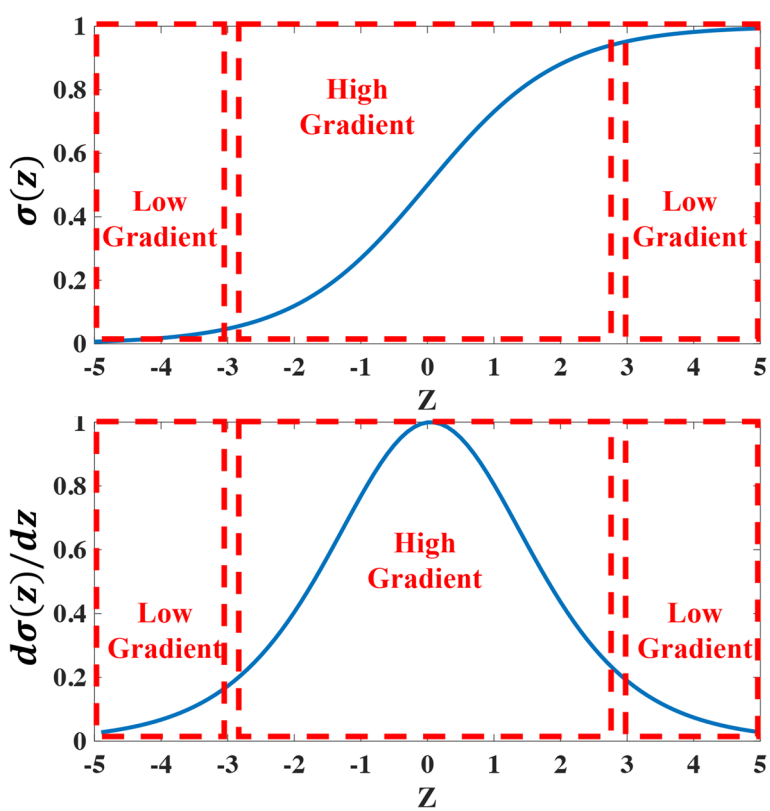

Fig. 2 The problem of gradient saturation

features. Integrated gradients can be accumulated along the path from the baseline to the actual input, thus providing a more comprehensive representation.

Shrikumar et al. [52] proposed DeepLIFT (DL) value as a recursive prediction interpretation method for deep learning. Different from gradient attribution, the DeepLIFT value is the linearized version of the ES value of the deep network. The Deep-Lift method uses the reference difference method of neuron attribution to assign attribution scores. First, set a reference or null value for the input neurons in the network. During forward propagation, calculate the difference between each input neuron and the reference value. Next, estimate the contribution of each input neuron to the output of the subsequent neurons, and associate the difference in the output with the corresponding input difference. Through the back-propagation strategy, distribute the difference correctly to each input neuron, thereby assigning a corresponding contribution score for each neuron in the network.

$$\sum_{i=1}^{N} C_{\nabla x_i \nabla t} = \nabla t \tag{4}$$

Where $C_{\nabla x_i \nabla t}$ represents the difference between t and the reference, attributed to or "lamed on" the difference between x_i and the reference. Notably, when the transfer function of the neuron performs well, the output has a locally linear relationship with its input.

Therefore, the training process of the AI-AMC network using the gradient descent method relies on the IG or DL of input signal features based on the model's classification results. This approach obtains model feature importance scores. Through feature importance analysis, a better understanding of how the model processes input data and the criteria it follows can be achieved.

2. The model-independent explanation method

The model-independent explanation method do not require access to model parameters when providing interpretability in the output. They aim to explain predictions of deep learning models by highlighting important input features. Among them, the Local Interpretable Model-agnostic Explanations (LIME) method involves fitting a locally interpretable surrogate model to a classifier near the target sample [53]. Significance is then calculated based on the parameters of these models.

LIME is a method that explains individual model predictions by constructing a locally approximated model based on a given prediction. Therefore, it is an additive feature attribution method. Let the model to be explained be represented as $f : \Re^d \rightarrow \Re$. Introduce a weighting function $\pi_x(z)$, which measures the proximity between z and x, thus defining the locality around x. Finally, define $\zeta(f, g, \pi_x)$ as the unfaithfulness of g when approximating f in the locality represented by π_x. To ensure interpretability and local fidelity of the model, it is necessary to minimize $\zeta(f, g, \pi_x)$ while ensuring that $\Omega(g)$ is sufficiently low for human comprehension.

Finally, define $\zeta(f, g, \pi_x)$ as the faithfulness when approximating X in the locality represented by X. To ensure interpretability and local fidelity of the model, the goal is to minimize X while ensuring that X is low enough for human comprehension. The explanation given by LIME is as follows:

$$\xi(x) = argmin_{g \in G} \zeta(f, g, \pi_x) + \Omega(g) \tag{5}$$

Where, G represents different explanation families, ζ is the fidelity function, and Ω is the measure of the complexity penalty for g. The fidelity of the explanation model $g(z)$ to the original model $f(h_x(z')$ is enforced by the loss ζ on a set of samples weighted by the local kernel π_x in a simplified input space.

The Shapley regression value method measures the importance of features by the Shapley values of the conditional expectation function of the model to be explained. It allocates significance scores relative to the output by introducing perturbations on the corresponding features. When dealing with multicollinearity, Shapley regression values offer an approach to assess feature importance. When direct model training on all feature subsets becomes infeasible, Expected Shapley (ES) values provide a fast approximation by treating the model output as the expected value. As a result, SHAP explanations exhibit both global and local consistency, ensuring the reliability and stability of the interpretation results. Furthermore, the definition of SHAP values is closely tied to Shapley regression, Shapley sampling, and quantitative feature attributes, while also allowing integration with methods like LIME, DL, and others.

For image data, these methods assign significance scores to pixel positions, associated with a specific classification label. Some researchers have attempted to extend these methods to time series data, assigning significance to time points. However, for certain time series problems, crucial information may be hidden in latent features such as dominant frequencies, state-space model parameters, etc., making it challenging for position-based information extracted from classifiers to explain the importance of these features. Therefore, we believe that gradient-based methods have the potential to provide good interpretability for AI-based automatic identification of wireless signals, especially when dealing with time series data.

Design

This section introduces the design of our explainable AI model for AI-based AMC including three modules: attribution feature determination, wireless attributed

feature conversion, and attribution-based model optimization. As shown in Fig. 3, the whole structure of our model is presented. Through the classification result comparison between the baseline and the input, the gradient estimation is further obtained. Then we are able to determine the attribution features according to the gradient estimation. With the attribution result, the feature conversion module focuses on converting the attribution features in the neural network domain to the attribution features in the wireless signal domaincomputing the reliability degree of the AI-based AMC. Besides, the model optimization, i.e., attribution-based model optimization?figuring out the improvement for the AI-based AMC model by adjusting the model feature with the attribution information.

Attribution input determination

As for a signal receiver in an AMC system, the signal is arbitrary in practice, which means there exists noise, interference, and a mixture of more than one signal. Therefore, our solution to interpret the AI-base AMC model is to attribute the input to the output so that we can determine the attribution feature of an arbitrary input in a classification process. Though we do not know how the an input pass through the neural network and finally turn into one classification, we are able to determine which features contribute the output. This attribution is of significance for a classification model in AI model explanation because it helps to reason what part of the input and why this part contributes to the output.

For the illustration purpose, we analyse a typical network as shown in Fig. 4. We denote the input of the neural network as $I = [I_1, I_2, I_3]$ and the output of the neural network as $O = [O_1, O_2, O_3, O_4]$. And the weight and bias of each connection are denoted as $W_{l,N_{li}N_{(l+1)j}}$ and $b_{l,N_{li}N_{(l+1)j}}$ respectively, where l is the index of related layer and $N_{li}N_{(l+1)j}$ means the from node N_{li} to node $N_{(l+1)j}$. And nodes of the middle layers are denoted $[N_{21}, N_{22}, N_{23}, N_{24}, N_{25}]$ and $[N_{31}, N_{32}, N_{33}, N_{34}, N_{35}]$ respectively. In addition, the input node and output node also can be denoted as N_{li}. Then each neural network node is computed as follows

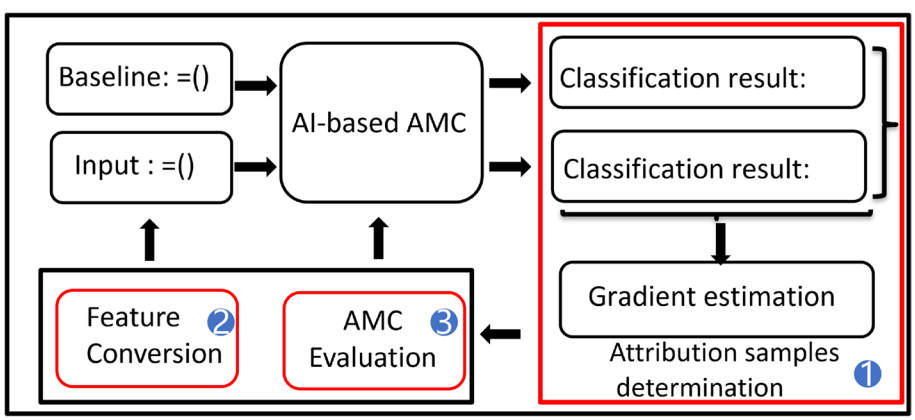

Fig. 3 The structure of XAI model for AI-based AMC

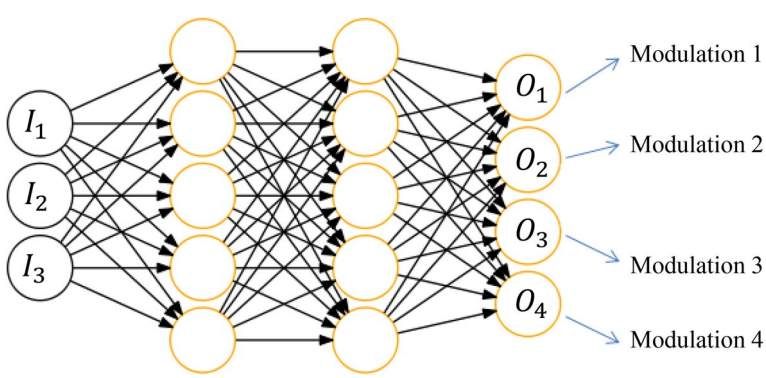

Fig. 4 An example of an AI-based AMC model

$$N_{(l+1)j} = AF\left(\sum_{i=1}^{M_l} N_{li} W_{l,N_{li}N_{(l+1)j}} + b_{l,N_{li}N_{(l+1)j}}\right) \quad (6)$$

where AF is the activation function, and M_l is the amount number of nodes of the layer l. So one output O_j in a neural network is represented as follows

$$O_j = AF\left(\sum_{i=1}^{M_L} N_{Li} W_{L,N_{Li}N_{(L+1)j}} + b_{L,N_{Li}N_{(L+1)j}}\right) \quad (7)$$

where L is the number layers, and M_L is the amount of nodes of the penultimate layer. From another perspective, all the connections are regarded as paths which the neural network nodes go through, and the weight, the bias and the activation are regarded as the impact of a path. So the impact is denoted as $P_{l,N_{li}N_{(l+1)j}}$. Then represent all the neural network nodes of middle layers with weight and bias progressively in Eq. 7, we obtain O_j as follows

$$O_j = \sum_{i_1=1}^{M_1} N_{1i_1} \sum \left(P_{1,N_{1i_1}N_{2i_2}} P_{2,N_{2i_2}N_{3i_3}} ... P_{L,N_{Li_L}N_{(L+1)i_{(L+1)}}}\right) \quad (8)$$

where the range of i_1 to i_L is from 1 to L and $\sum(P_{1,N_{1i_1}N_{2i_2}} P_{2,N_{2i_2}N_{3i_3}} ... P_{L,N_{Li_L}N_{(L+1)i_{(L+1)}})$ represents the total paths impact on node N_{1i_1}, i.e., input sample I_i. From Eq. 8, we observe that all the input samples undergo the impacts of the neural network paths and then lead to the classification result. Therefore, from a neural network perspective, the cause of the output are the various impact of the paths.

In our model, the baseline of a signal input is set as blank signal. Generally, the wireless signal has two channels of information that are independent and represent different information. Therefore, both two channels of information are combined to feed a neural network, but their attributions are considered separately. The signal input is denoted as $x = (I_I, I_Q)$ or $x = (I_A, I_P)$, which is depends on the input set of a AI-based AMC model. $x = (I_I, I_Q)$ is the IQ signal of a original signal and $x = (I_A, I_P)$ is the amplitude and phase of a original signal. For illustration, we denote them both as $x = (IF, IS)$. Then the integrated gradient of the ith dimension is denoted as follows

$$IG(x_i) = \left(IF_i \times \sum_{j=1}^{N} \frac{\partial F(x_{ij})}{\partial IF_i}, IS_i \times \sum_{j=1}^{N} \frac{\partial F(x_{ij})}{\partial IS_i}\right) \quad (9)$$

where IF_i and IS_i are the ith dimension of input $x = (IF, IS)$ respectively. Compared with the gradients of Eq. 2, integrated gradients of the input provide a more accurate measurements of both two channel of the wireless signal. From the perspective of neural network, integrated gradients provide a attribution interpretation of the AMC model.

Wireless attributed feature conversion

From the neural network perspective, the interpretation is achieved due to the attribution distribution. However, it is not straightforward explainable in wireless signal domain because we only obtain the attributions distribution of input sample points instead of any attribution feature that are common recognized in wireless signal domain. And to achieve the goal of the effective utilization of an interpretation model on practical AMC model, it is necessary to convert the interpretation to wireless signal domain. Hence, we aim to convert the attributions distribution to attributing wireless signal features.

As a symbol of a wireless signal is typically modulated by a segment of signal, the signal feature should be obtained by the signal segment consisting of

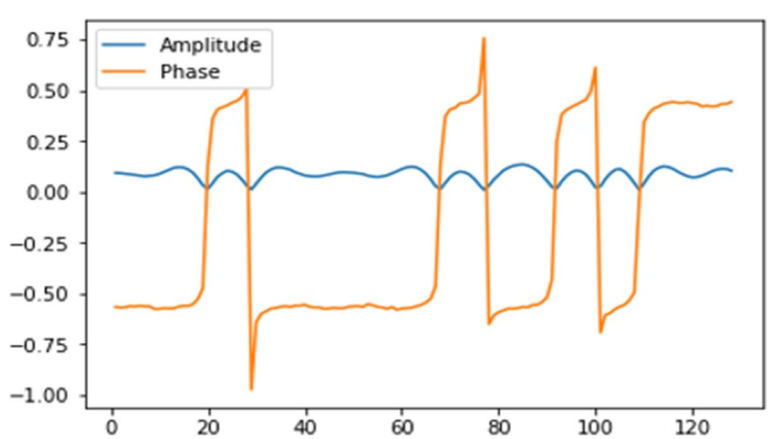

Fig. 5 The original amplitude and phase of QPSK

consecutive discrete input samples. As shown in Fig. 5, the four phases features of the QPSK signal could be identified by the marked signal segments, which is why the modulation of this received signal is recognized as QPSK. Our objective of the feature conversion model is to obtained such signal segments so that the interpretation from the wireless signal perspective is achieved. Different from the processing for the whole received signal, the feature conversion aims to find the attributing signal segment parts. This is necessary since the signal received in practice usually consists of blank parts, noise, and interference. In addition, for the modulation classifier, typically the relatively stable parts contribute to the final classification decision. Consistent with the design to determine the attributing input samples introduced in the last section, the design in this section is to determine the attributing wireless signal features.

To obtain the attributing wireless signal features, we have to determine the attributing signal segments. For inputs to the classifier, these signal segments are regarded as special cut parts. Motivated by this observation, we propose to computing the wireless feature distribution according to the attribution scores from "Attribution input determination" section. We denote a point of a input with attribution scores as $IP = (SI, AS)$ where SI is the signal input(i.e., amplitude, phase,in-phase and quadrature component) and AS is the its corresponding attribution score.

We propose to utilize a series of windows to cut and extract these signal segments which called attention-based windows. On accounting that the lengths of the signal segments are not fixed and same, the windows should be dynamic which guarantees the extraction of the arbitrary attributing signal segments. Excepted representing the attribution, the extraction signal segments obtained through these dynamic windows are also regarded as the preprocessing for inputs. The framework of the feature conversion is presented in Fig. 6, the input cut through the attention-based windows, on the one hand, is fed into the original AI-based AMC model, and on the other hand, is fed into the feature extraction to obtain the wireless signal features. Compared with the original AI-based model, the main difference is the input of the neural network. Since the attributing signal segments are fed, the neural network could achieve higher effectiveness in recognizing modulation formats as a result of interfering signal segments being weakened. Additionally, the accuracy of modulation classification improvement is actually the interpretation, which means the attributing features are correctly located.

Algorithm 1 Attributing signal segment determination

1: **while** x_i *in Input* x **do**
2: $\quad IG(x_i) = (IF_i \times \sum_{j=1}^{N} \frac{\partial F(x_{ij})}{\partial IF_i}, IS_i \times \sum_{j=1}^{N} \frac{\partial F(x_{ij})}{\partial IS_i})$
3: $\quad IG(x) = Insert(IG(x), IG(x_i))$
4: **end while**
5: $Ww_i = \frac{\sum_{j=1}^{WL_i}(G[WI_i[j]])}{WL_i}$

As shown in Fig. 6, the windows to extract the attributing signal segments are achieved by the attention mechanism. The basic idea of this design is that the inputs with similar attention weights are able to be combined and the combinations are regarded as extraction by windows.

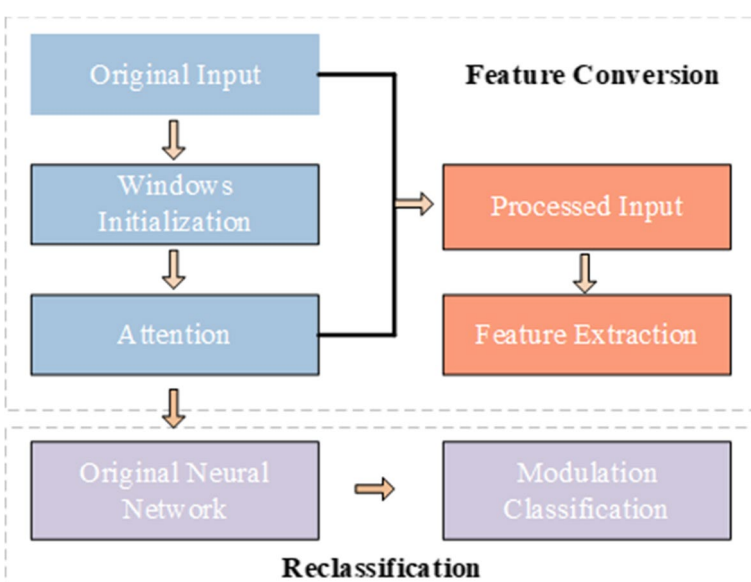

Fig. 6 The framework of the feature extraction model

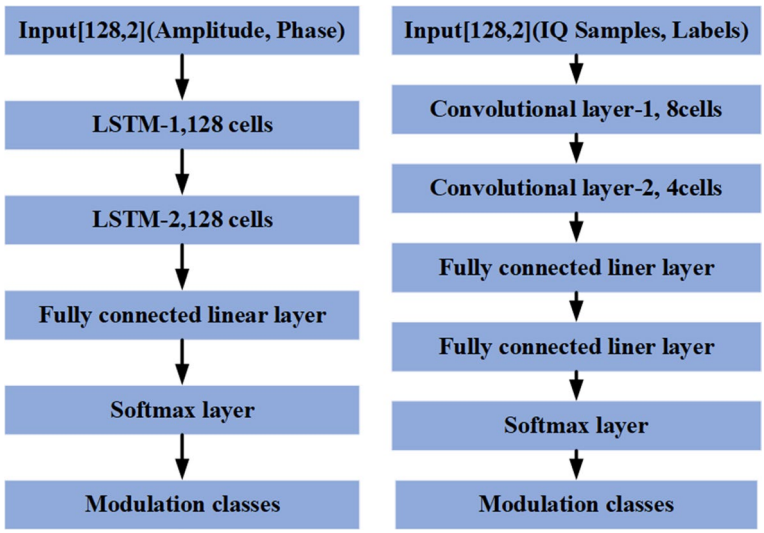

Fig. 7 Different AMC model

Specifically, the window initialization is achieved by the attribution scores of input samples from "Attribution input determination" section. Since the attribution scores represent the contribution degree of input samples to outputs, these attribution scores roughly provides continuity of the input samples with similar scores. Start with the attribution scores of input samples, the initial windows are assigned roughly according to the similarity of attribution scores. The attribution scores are computing only related to the structure of the neural network and the input samples, so the initial windows have to be adjusted to achieve more window assignments. An initial window is denoted as $W_i = [Wl_i, Wr_i, Ww_i]$, where Wl_i and Wr_i represent the left boundary and right boundary of the window i and Ww_i is the weight of the window. The initial weight of the window is the weight for each input samples of the window and is defined as follows

$$Ww_i = \frac{\sum_{j=1}^{WL_i}(G[WI_i[j]])}{WL_i} \tag{10}$$

where $WI_i[j]$ is the $j-th$ index of winodw i and WL_i is its length.

After feature conversion, we are able to reason the classification by locate the wireless features that contribute to the result significantly.

Evaluation

In this section, we present our proposed explainability experimental results, taking only the attribution feature transformation of IG as an example.

Data Source. In our experiments, we adopt the open source radio dataset RML2016.10a [54] to construct the models, which contains 11 modulations, 8PSK, AM-DSB, AM-SSB, BPSK, CPFSK, GFSK, PAM4, QAM16, QAM64, QPSK, WBFM. The length of each data of the dataset are 128 where each 4 samples represent a symbol.

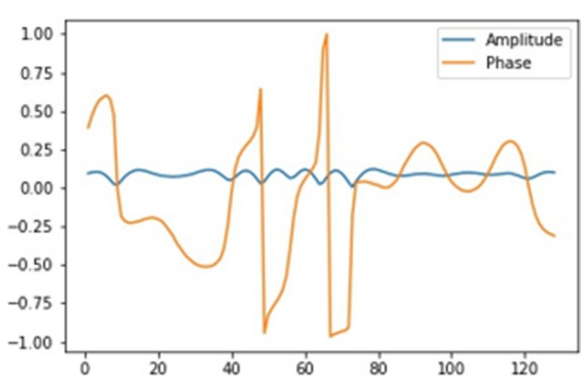

Fig. 8 An original input (8PSK)

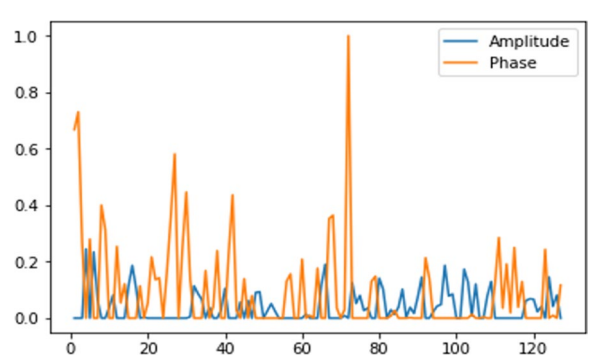

Fig. 9 Attribution of the input

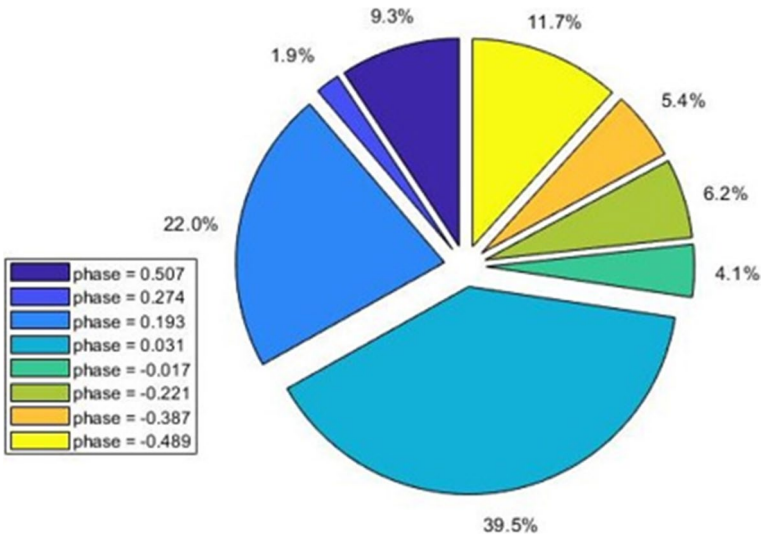

Fig. 10 Feature distribution

And the structure of LSTM-based AMC is shown in Fig. 7.

We utilize Fig. 8 as input, and in this paper we converted the input of the AMC from the original IQ signal to the input in two dimensions, amplitude and phase. The x-axis is the index of samples, and the y-axis is the normalized magnitude and phase, as shown in Fig. 9. The amplitude and phase are input to the neural network as a whole, and after getting the classification results, we calculate the attribution scores for all these input sample

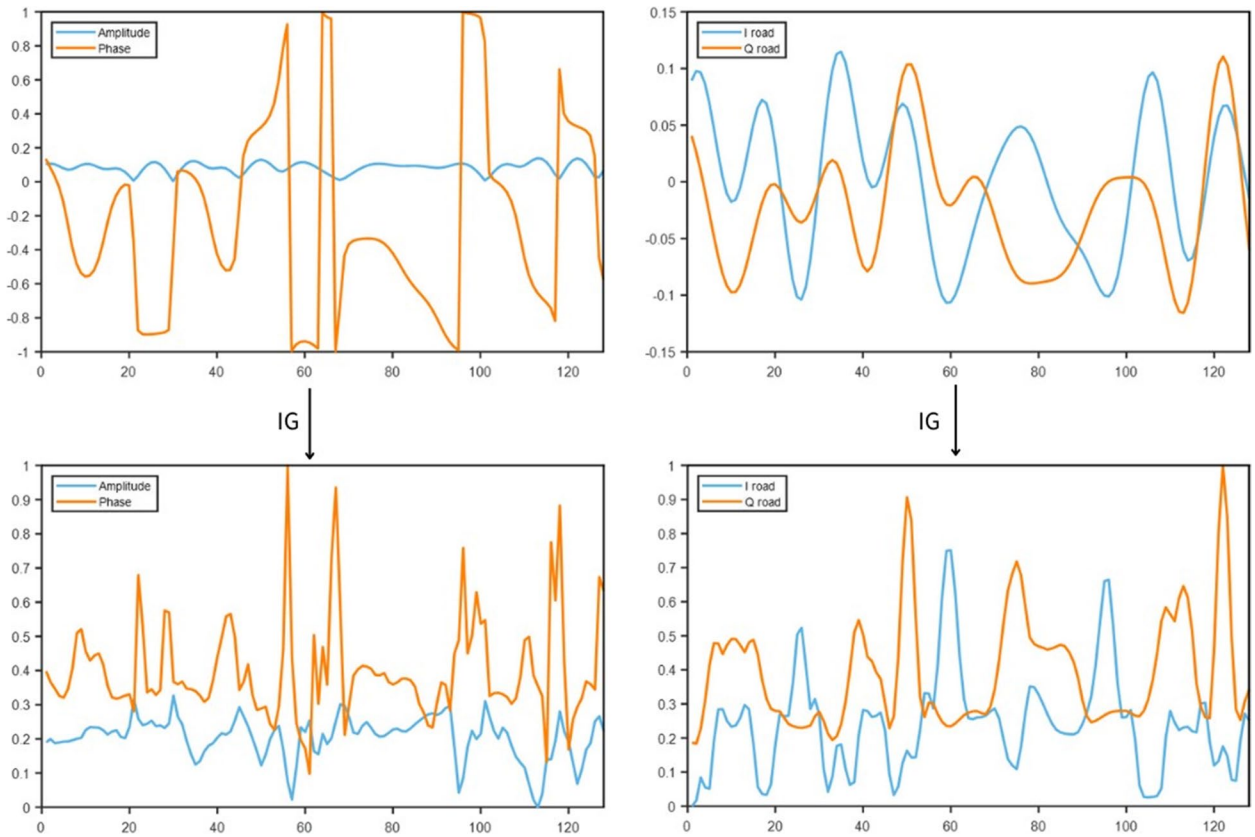

Fig. 11 Attribution performances of the same signal of QAM64 in CNN networks

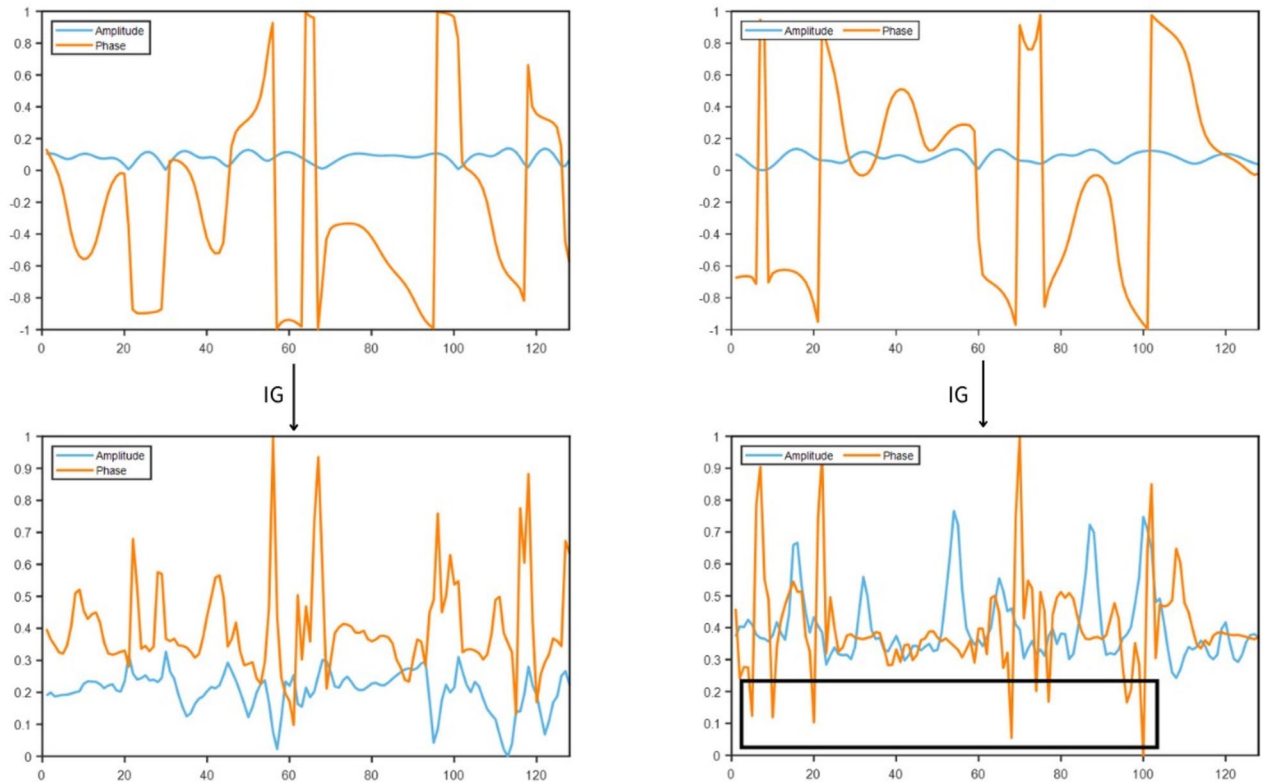

Fig. 12 Different attribution performances of different QAM64 in CNN networks

points, and finally get the result on the right. The x-axis is the index of samples, the y-axis is the attribution score, ranging from 0 to 1.

As shown in Fig. 10, implement the feature conversion model to get the explainability from the wireless communication perspective. Based on the results of wireless signal feature distribution, we will optimize the model and present a more explicit feature distribution compared to traditional methods.

Figure 11 shows the attribution features of the AP and IQ signals of QAM64 in CNN. The CNN network learns different features to output classification results. Figure (a) was mistakenly identified as 8PSK by the network, while Figure (b) was correctly identified as QAM64. Although it is the same modulation signal, it indicates that the input features of the signal are the key to network learning, and correct feature input can help improve the accuracy of network learning.

Figure 12 shows the contribution scores of attribution features of different QAM64 signal samples from the same learning network in CNN. Figure (a) was misjudged as 8PSK, while (b) was Correctly identified as QAM64. The part circled in the box represents the main different attribution features of the two signal samples, indicating that the same type of signal network learns different features, which leads to different classification results.

Conclusion

In this work, we investigated explainability methods for AI-based automatic modulation classification in wireless communication. Firstly, we provided a brief overview of some traditional automatic modulation classification methods. Subsequently, we elucidated the major deep learning architectures for automatic modulation classification based on AI that have emerged in recent years. Following that, we conducted a comprehensive review and comparison of the state-of-the-art interpreters for automatic modulation classification. Lastly, we analyzed several explainability methods and assessed their feasibility in empowering the topic of automatic modulation classification. Deployment of automatic modulation classification networks based on explainable deep learning on mobile edge devices is bound to improve the communication performance as well as the application range of wireless communications.

Authors' contributions
Bo Xu completed the experimental ideas, completed the final software simulation testing, and completed the writing of the paper, as well as conducting formal analysis; Uzair Aslam Bhatti and Hao Tang performed the data analysis; Jialin Yan performed the validation; Shulei Wu reviewed and summarized the research status of the whole field, determined the progressiveness of the method in this paper, and revised the paper; Nadia Sarhan organized the data required for the experiment in this article and assisted in completing the initial draft of the paper; The feasibility of the experimental design was verified by Emad Mahrous Awwad; Syam M.S. participated in the framework design of the paper, completed the chart analysis of the paper, and proofread it; Yazeed Yasin Ghadi provides improvement points and designs experimental models.

Funding
This work was supported by National Natural Science Foundation of China (No.61966013), National Natural Science Foundation of China Foreign Scholar Funds (No.62350410483), Hainan Natural Science Foundation of China (No.620RC602), Hainan Provincial Key Laboratory of Ecological Civilization and Integrated Land-sea Development, Tianjin University - Hainan University Independent Innovation Fund Cooperation Project (No.HDTD202301, 2023XSU-0035) and Hainan University Research Initiation Fund Project (No.KYQD(ZR)23143). The authors present their appreciation to King Saud University for funding this research through Researchers Supporting Program number (RSPD2024R1052), King Saud University, Riyadh, Saudi Arabia.

Availability of data and materials
No datasets were generated or analysed during the current study.

Declarations

Ethics approval and consent to participate
Not applicable.

Competing interests
The authors declare no competing interests.

Received: 29 November 2023 Accepted: 3 January 2024
Published online: 08 January 2024

References

1. Zhang Y, Chen J, Ma X, Wang G, Bhatti UA, Huang M (2024) Interactive medical image annotation using improved attention u-net with compound geodesic distance. Expert Syst Appl 237:121282
2. Bhatti UA, Marjan S, Wahid A, Syam M, Huang M, Tang H, Hasnain A (2023) The effects of socioeconomic factors on particulate matter concentration in china's: New evidence from spatial econometric model. J Clean Prod 417:137969
3. Cheng M, Li D, Zhou N, Tang H, Wang G, Li S, Bhatti UA, Khan MK (2023) Vision-motion codesign for low-level trajectory generation in visual servoing systems. IEEE Trans Instrum Meas 72:1–14
4. Hagendorff T (2020) The ethics of ai ethics: An evaluation of guidelines. Mind Mach 30(1):99–120
5. Amodei D, Olah C, Steinhardt J, Christiano P, Schulman J, Mané D (2016) Concrete problems in AI safety. arXiv preprint arXiv:1606.06565
6. Kim D, Lee C, Park S, Lim S (2022) Potential liability issues of AI-based embedded software in maritime autonomous surface ships for maritime safety in the korean maritime industry. J Mar Sci Eng 10(4):498
7. Wu Y, Zhang L, Bhatti UA, Huang M (2023) Interpretable machine learning for personalized medical recommendations: a lime-based approach. Diagnostics 13(16):2681
8. Chen HY, Lee CH (2020) Vibration signals analysis by explainable artificial intelligence (XAI) approach: application on bearing faults diagnosis. IEEE Access 8:134246–134256
9. Graaf MM, Malle BF, Dragan A, Ziemke T (2018) Explainable robotic systems. In: Companion of the 2018 ACM/IEEE International Conference on HumanRobot Interaction. Association for Computing Machinery, New York, p 387–388
10. Bojarski M, Del Testa D, Dworakowski D, Firner B, Flepp B, Goyal P, Jackel LD, Monfort M, Muller U, Zhang J, et al (2016) End to end learning for self-driving cars. arXiv preprint arXiv:1604.07316
11. Haspiel J, Du N, Meyerson J, Robert Jr LP, Tilbury D, Yang XJ, Pradhan AK (2018) Explanations and expectations: Trust building in automated vehicles. In: Companion of the 2018 ACM/IEEE International Conference on Human-robot Interaction. Association for Computing Machinery, New York, p 119–120
12. Wang S, Qureshi MA, Miralles-Pechuaán L, Huynh-The T, Gadekallu TR, Liyanage M (2021) Explainable AI for b5g/6g: Technical aspects, use cases, and research challenges. arXiv preprint arXiv:2112.04698
13. Wang S, Khan A, Lin Y, Jiang Z, Tang H, Alomar SY, Sanaullah M, Bhatti UA (2023) Deep reinforcement learning enables adaptive-image augmentation for automated optical inspection of plant rust. Front Plant Sci 14
14. Bhatti UA, Tang H, Wu G, Marjan S, Hussain A (2023) Deep learning with graph convolutional networks: An overview and latest applications in computational intelligence. Int J Intell Syst 2023:1–28
15. Khan M, Liu M, Dou W, Yu S (2015) vgraph: graph virtualization towards big data. In: 2015 Third International Conference on Advanced Cloud and Big Data, IEEE, pp 153–158
16. Khan A, Zhang H, Boudjellal N, Ahmad A, Khan M (2023) Improving sentiment analysis in election-based conversations on twitter with elecbert language model. Comput Mater Continua 76(3):3345–3361
17. Xu JL, Su W, Zhou M (2010) Likelihood-ratio approaches to automatic modulation classification. IEEE Trans Syst Man Cybern Part C Appl Rev 41(4):455–469
18. Hameed F, Dobre OA, Popescu DC (2009) On the likelihood-based approach to modulation classification. IEEE Trans Wirel Commun 8(12):5884–5892
19. Abdi A, Dobre OA, Choudhry R, Bar-Ness Y, Su W (2004) Modulation classification in fading channels using antenna arrays. In: IEEE MILCOM 2004. Military Communications Conference, 2004., vol 1. IEEE, pp 211–217
20. Dobre OA, Abdi A, Bar-Ness Y, Su W (2005) Blind modulation classification: a concept whose time has come. In: IEEE/Sarnoff Symposium on Advances in Wired and Wireless Communication, 2005., IEEE, pp 223–228
21. Panagiotou P, Anastasopoulos A, Polydoros A (2000) Likelihood ratio tests for modulation classification. In: MILCOM 2000 Proceedings. 21st Century Military Communications. Architectures and Technologies for Information Superiority (Cat. No. 00CH37155), vol 2. IEEE, pp 670–674
22. Dobre O, Zarzoso J, Bar-Ness Y, Su W (2004) On the classification of linearly modulated signals in fading channel. In: Proc. CISS, pp 71–74
23. Dobre OA, Hameed F (2006) Likelihood-based algorithms for linear digital modulation classification in fading channels. In: 2006 Canadian conference on electrical and computer engineering, IEEE, pp 1347–1350
24. Dobre OA, Hameed F (2007) On performance bounds for joint parameter estimation and modulation classification. In: 2007 IEEE Sarnoff Symposium, IEEE, pp 1–5
25. Huan CY, Polydoros A (1995) Likelihood methods for mpsk modulation classification. IEEE Trans Commun 43(2/3/4):1493–1504
26. Boiteau D, Le Martret C (1998) A general maximum likelihood framework for modulation classification. In: Proceedings of the 1998 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP'98 (Cat. No. 98CH36181), vol 4. IEEE, pp 2165–2168
27. Wei W, Mendel JM (2000) Maximum-likelihood classification for digital amplitude-phase modulations. IEEE Trans Commun 48(2):189–193
28. Häring L, Chen Y, Czylwik A (2010) Automatic modulation classification methods for wireless OFDM systems in TDD mode. IEEE Trans Commun 58(9):2480–2485
29. Soliman SS, Hsue SZ (1992) Signal classification using statistical moments. IEEE Trans Commun 40(5):908–916
30. Swami A, Sadler BM (2000) Hierarchical digital modulation classification using cumulants. IEEE Trans Commun 48(3):416–429
31. Wang F, Wang X (2010) Fast and robust modulation classification via Kolmogorov-Smirnov test. IEEE Trans Commun 58(8):2324–2332
32. Nandi AK, Azzouz EE (1998) Algorithms for automatic modulation recognition of communication signals. IEEE Trans Commun 46(4):431–436
33. Meng F, Chen P, Wu L, Wang X (2018) Automatic modulation classification: a deep learning enabled approach. IEEE Trans Veh Technol 67(11):10760–10772

34. O'Shea TJ, Roy T, Clancy TC (2018) Over-the-air deep learning based radio signal classification. IEEE J Sel Top Signal Process 12(1):168–179
35. Peng S, Jiang H, Wang H, Alwageed H, Zhou Y, Sebdani MM, Yao YD (2018) Modulation classification based on signal constellation diagrams and deep learning. IEEE Trans Neural Netw Learn Syst 30(3):718–727
36. Wang Y, Yang J, Liu M, Gui G (2020) LightAMC: lightweight automatic modulation classification via deep learning and compressive sensing. IEEE Trans Veh Technol 69(3):3491–3495
37. Yang J, Dong B, Fu X, Wang Y, Gui G (2022) Lightweight decentralized learning-based automatic modulation classification method. J Commun 007:043
38. Rafique W, Khan M, Khan S, Ally JS, et al (2023) Securemed: A blockchain-based privacy-preserving framework for internet of medical things. Wireless Commun Mobile Comput 2023:1–14
39. Rafique W, Khan M, Dou W (2019) Maintainable software solution development using collaboration between architecture and requirements in heterogeneous iot paradigm (short paper). In: Collaborative Computing: Networking, Applications and Worksharing: 15th EAI International Conference, CollaborateCom 2019, London, UK, August 19-22, 2019, Proceedings 15, Springer, pp 489–508
40. Samek W, Montavon G, Vedaldi A, Hansen LK, Müller KR (2019) Explainable ai: interpreting, explaining and visualizing deep learning 11700
41. Samek W, Binder A, Montavon G, Lapuschkin S, Müller KR (2016) Evaluating the visualization of what a deep neural network has learned. IEEE Trans Neural Netw Learn Syst 28(11):2660–2673
42. Smilkov D, Thorat N, Kim B, Viégas F, Wattenberg M (2017) Smoothgrad: removing noise by adding noise. arXiv preprint arXiv:1706.03825
43. Ancona M, Ceolini E, Oztireli C, Gross M (2019) Gradient-based attribution methods. Explainable AI: Interpreting, explaining and visualizing deep learning 11700:169–191
44. Montavon G, Samek W, Müller KR (2018) Methods for interpreting and understanding deep neural networks. Digit Signal Process 73:1–15
45. Sundararajan M, Taly A, Yan Q (2017) Axiomatic attribution for deep networks. In: International conference on machine learning, PMLR, pp 3319–3328
46. Zintgraf LM, Cohen TS, Adel T, Welling M (2017) Visualizing deep neural network decisions: Prediction difference analysis. arXiv preprint arXiv:1702.04595
47. Chen J, Miao S, Zheng H, Zheng S (2020) Feature explainable deep classification for signal modulation recognition. In: IECON 2020 The 46th Annual Conference of the IEEE Industrial Electronics Society, pp 3543–3548. https://doi.org/10.1109/IECON43393.2020.9254271
48. Huang L, Zhang Y, Pan W, Chen J, Qian LP, Wu Y (2021) Visualizing deep learning-based radio modulation classifier. IEEE Trans Cogn Commun Netw 7(1):47–58. https://doi.org/10.1109/TCCN.2020.3048113
49. Zhou H, Bai J, Wang Y, Ren J, Yang X, Jiao L (2023) Deep radio signal clustering with interpretability analysis based on saliency map. Digit Commun Netw. https://doi.org/10.1016/j.dcan.2023.01.010
50. Duggal G, Gaikwad T, Sinha B (2023) Dependable modulation classifier explainer with measurable explainability. Front Big Data 5. https://doi.org/10.3389/fdata.2022.1081872
51. Sundararajan M, Taly A, Yan Q (2017) Axiomatic attribution for deep networks. CoRR abs/1703.01365. arXiv:1703.01365
52. Shrikumar A, Greenside P, Shcherbina A, Kundaje A (2016) Not just a black box: Learning important features through propagating activation differences. CoRR abs/1605.01713. arXiv:1605.01713
53. Gaikwad T, Duggal G, Sinha B (2023) Dependable modulation classifier explainer with measurable explainability, vol 5. https://doi.org/10.3389/fdata.2022.1081872
54. O'shea TJ, West N (2016) Radio machine learning dataset generation with gnu radio. In: Proceedings of the GNU Radio Conference, vol 1

Publisher's Note