Constellation Diagram Augmentation and Perturbation-Based Explainability for Automatic Modulation Classification

Anonymous Authors

***Abstract*—Automatic Modulation Classification (AMC) is a cornerstone in ensuring the adaptability and efficiency of modern wireless communication systems. In this paper, we present a novel framework that combines multi-task learning with perturbation-based explainability to address both performance and interpretability challenges in AMC. By transforming the I/Q signal data into enriched constellation diagrams and employing a ResNet-based architecture, our model simultaneously classifies modulation schemes and estimates Signal-to-Noise Ratio (SNR) buckets, leveraging shared feature representations for improved generalization. To enhance interpretability, we systematically perturb high- and low-intensity regions of the constellation diagrams, analyzing their impact on classification accuracy and using the Perturbation Impact Score (PIS) metric. Our experimental evaluation on the RadioML 2018.01A dataset demonstrates the robustness of the proposed framework, achieving high accuracy across diverse modulation types even under challenging noise conditions. Additionally, perturbation analysis reveals critical regions in the constellation diagrams that drive model decisions, offering actionable insights for further optimization.**

***Index Terms*—Modulation classification, constellation diagrams, multi-task deep learning, explainability, perturbation methods, signal-to-noise ratio estimation.**

I. Introduction

Automatic Modulation Classification (AMC) [1] [2] [3] is a fundamental task in wireless communication, enabling the identification of modulation schemes critical for signal decoding, interference mitigation, and dynamic spectrum management. AMC has far-reaching applications across domains such as cognitive radios, public safety, and military operations, where accurate and timely modulation classification directly impacts reliability, efficiency, and security.

Traditional approaches to AMC, including methods based on fourth-order cumulants and Artificial Neural Networks (ANNs) [4], have historically laid the foundation for this field. While these techniques achieved moderate success in controlled settings, their reliance on hand-engineered features and limited adaptability to complex and noisy environments restricted their applicability in real-world scenarios. With the advent of deep learning, Convolutional Neural Networks (CNNs) have demonstrated superior performance in AMC by leveraging their ability to learn rich feature representations directly from raw or preprocessed data [2]. However, despite their high accuracy, these models often operate as "black boxes," providing limited insight into their decision-making processes—a critical drawback in safety-critical applications where trust and transparency are paramount.

The growing need to understand and trust model decisions has driven interest in explainability techniques. Recent advancements, such as Concept Bottleneck (CB) models [5], have aimed to bridge the gap between performance and interpretability in AMC. While CB models have shown promise in explaining model decisions and even classifying modulation schemes unseen during training, they do not address how specific input features or perturbations affect classification accuracy. This gap in understanding limits the deployment of AMC models in high-stakes applications, where decision transparency is essential for regulatory compliance and operational reliability.

In this work, we propose a novel framework that combines constellation diagram augmentation with perturbation-based explainability to enhance the accuracy, robustness, and interpretability of deep learning-based AMC models. Our approach systematically investigates the impact of perturbations on constellation diagrams, identifying critical regions that drive model predictions. By integrating multi-task learning, we simultaneously predict modulation types and SNRs. This addresses the dual needs of modulation classification and channel quality estimation in real-world communication systems.

The key contributions of this work are as follows:

- **Multi-Task Learning for Modulation and SNR Classification**: We introduce a multitask learning framework that jointly predicts the modulation type and SNR, enhancing the model's utility in dynamic wireless environments. This dual-task approach is particularly beneficial for applications requiring adaptive spectrum management and interference mitigation.
- **Constellation Diagram Augmentation**: We propose novel augmentations to traditional constellation diagrams, transforming them into rich visual representations that improve classification robustness, especially for complex modulation schemes under challenging noise conditions.
- **Perturbation-Based Explainability**: To improve interpretability, we employ perturbation-based methods to visualize the impact of specific regions of the constellation diagram on model predictions. This approach provides actionable insights into the features that drive classification decisions, enhancing trust in model outputs.
- **Security Implications and Public Safety Applications**: By quantifying the relationship between perturbation levels and classification performance, we reveal the robustness of deep learning models in wireless communication

contexts. These findings are critical for deploying AMC models in security-sensitive applications.

- **Progressive Perturbation Analysis**: We systematically analyze the degradation in classification accuracy under varying levels of perturbation, providing insights into the model's resilience and guiding future developments in robust and interpretable AMC frameworks.

II. Related Work

AMC is pivotal in wireless communications, enabling the identification of modulation schemes for efficient signal processing and dynamic spectrum management. Traditional approaches, such as those using fourth-order cumulants [4] and ANNs [6], employed statistical analysis and hand-crafted features to identify modulation schemes. Azzouz and Nandi [6] developed decision criteria and signal processing methods for identifying different types of digital modulation, achieving a success rate of over 90% at 10 dB SNR. While foundational, these methods often struggle with scalability and generalization under noisy conditions.

The advent of deep learning has revolutionized AMC, particularly with the introduction of Convolutional Neural Networks (CNNs). Peng et al. [2] leveraged CNN architectures to process constellation diagrams, achieving significant success on datasets like RadioML. These models introduced superior feature extraction capabilities compared to traditional methods but often lacked interpretability.

Residual Networks (ResNets) have further advanced AMC performance. Kumar et al. [7] demonstrated the robustness of ResNet architectures for AMC tasks by introducing a deep residual neural network that achieved high accuracy and resilience in noisy environments. The use of ResNets, which mitigate the problems of vanishing gradients in deep networks, has become a cornerstone in modern AMC pipelines.

Transformer-based models demonstrated state-of-the-art performance on datasets such as RadioML. Kong et al. [8] introduced a Vision Transformer (ViT)-based model that leverages self-attention mechanisms to capture global dependencies in input data, proving especially effective in handling noisy signals.

O'Shea et al. [9] explored over-the-air deep learning approaches for AMC, addressing real-world impairments such as carrier frequency offset and multipath fading. Their work highlighted the challenges in bridging the gap between synthetic datasets like RadioML and real-world signals. Similarly, Zhao et al. [10] proposed a meta-supervised contrastive learning framework to enhance AMC performance, particularly in few-shot and open-set scenarios.

Explainability in AMC has garnered significant attention, driven by the need to interpret and trust model decisions. Most AMC methods treat deep learning models as "black boxes," relying on post-hoc interpretability techniques such as Grad-CAM [11] and Integrated Gradients [12]. While effective, these approaches often lack the granularity needed to understand how specific signal features influence model decisions. Recent studies have highlighted the importance of explainability in AMC, emphasizing the need for integrated methods that offer transparency without compromising performance.

Perturbation-based techniques, widely used in other domains [13]–[15], provide a promising avenue for AMC. By systematically modifying input data, these methods evaluate the impact of specific features on model predictions. Methods such as LIME [14], XRAI [15], and meaningful perturbation [13] provide insights into how specific input features affect predictions. Wong and McPherson [5] introduced the use of concept bottleneck models, bridging the gap between performance and interpretability in AMC. These models first predict predefined concepts, which are then used for the final classification decision, offering inherent explanations for predictions. Our work builds on this foundation, integrating perturbation-based explainability into the AMC pipeline to provide actionable insights into the model's decision-making process.

III. Proposed Approach

A. Conversion of I/Q Signals to Constellation Diagrams

In digital communication systems, signals are represented using in-phase (I) and quadrature (Q) components, forming complex-valued time-domain samples. Constellation diagrams, which plot the I component on the x-axis and the Q component on the y-axis, provide a visual representation of the modulation scheme and serve as input for image-based deep learning models. This transformation has been widely adopted in modulation classification research due to its ability to reveal distinct patterns for different modulation types [2], [6].

1) Transformation Steps: The process of transforming raw I/Q signals into constellation diagrams is as follows:

1) **Binning:** The I and Q components, denoted as $I(i)$ and $Q(i)$ respectively in Eq. 1, are binned into a 224-by-224 grid. Scaling factors s_I and s_Q are applied to map the data into discrete bins as given in Eq. 1.

$$x = \left\lfloor \frac{I(i) - I_{\text{min}}}{s_I} \right\rfloor, \quad y = \left\lfloor \frac{Q(i) - Q_{\text{min}}}{s_Q} \right\rfloor \tag{1}$$

In Eq. 1, I_{min} and Q_{min} are the minimum values of the I and Q ranges. Also, x and y are the binned coordinates.

2) **Smoothing:** A Gaussian smoothing filter is applied to the binned data to reduce noise as given in Eq. 2.

$$C_{\text{smoothed}}(x, y) = \sum_{i,j} C(i, j) \cdot G(x - i, y - j) \tag{2}$$

In Eq. 2, $C(i, j)$ is the intensity at bin (i, j), $G(x, y)$ is the Gaussian kernel and $C_{\text{smoothed}}(x, y)$ is the smoothed data.

3) **Normalization:** As depicted in Eq. 3, the smoothed data is normalized to $C_{\text{final}}(x, y)$ so that the pixel intensity values are in the range [0, 255]:

$$C_{\text{final}}(x, y) = \frac{C_{\text{smoothed}}(x, y)}{\max(C_{\text{smoothed}})} \cdot 255 \tag{3}$$

2) Visualization of Constellation Diagrams: Table I presents examples of constellation diagrams for various modulation types under different SNR levels. These images demonstrate the increasing impact of noise as the SNR decreases, making the points less distinct and the modulation scheme harder to classify.

B. SNR Bucketing

To handle the variability in the SNR, we categorize SNR values into three predefined buckets. For each signal, the appropriate SNR bucket is determined using:

$$\text{Bucket} = \begin{cases} \text{Low}, & \text{if } SNR \in [-20\ dB, -4\ dB] \\ \text{Medium}, & \text{if } SNR \in [-2\ dB, 14\ dB] \\ \text{High}, & \text{if } SNR \in [16\ dB, 30\ dB] \end{cases} \tag{4}$$

These buckets ensure a balanced distribution of samples across noise conditions, as utilized in prior modulation classification studies [2], [8]. It is important to note that the SNR values in the RadioML 2018.01A dataset are restricted to even numbers.

C. Multi-Task Learning Objective

Our framework integrates modulation classification and SNR estimation into a single model using a multi-task learning architecture based on ResNet18 [7]. The architecture includes:

- **Shared Layers:** The shared ResNet18 backbone extracts high-level features from the input constellation diagram.
- **Task-Specific Heads:** Separate fully connected layers predict modulation types and SNR buckets.

Equation 5 gives the overall loss function $\mathcal{L}_{\text{total}}$ as a weighted linear combination (α and β are the balancing weights) of the cross-entropy losses for both the tasks of modulation classification ($\mathcal{L}_{\text{modulation}}$) and SNR estimation ($\mathcal{L}_{\text{SNR}}$).

$$\mathcal{L}_{\text{total}} = \alpha \mathcal{L}_{\text{modulation}} + \beta \mathcal{L}_{\text{SNR}} \tag{5}$$

D. Justification of the Approach

This approach leverages the spatial representation of I/Q data in constellation diagrams, a method validated in earlier works [6], [8]. The use of SNR bucketing ensures robust performance across varying noise levels, while the multi-task learning framework addresses the dual objectives of modulation classification and SNR estimation efficiently. Perturbation-based explainability further provides insights into critical regions of the input, aligning the framework with real-world needs for interpretability and adaptability in communication systems.

IV. Explainability via Perturbation-Based Methods

In our modulation classification framework, we integrate explainability directly by employing a perturbation-based approach. Perturbation-based methods systematically alter specific regions of the input to evaluate their impact on model predictions, offering insights into the model's decision-making process. As highlighted by Ivanovs et al. [16], such methods are particularly advantageous due to their model-agnostic nature and adaptability to various data modalities, including image-like data, text, and reinforcement learning.

Perturbation-based approaches are especially valuable in safety-critical domains, where understanding model behavior is paramount for trust and transparency. Recent advancements in perturbation analysis, such as the work by Fel et al. [17], have demonstrated robust and efficient explainability by systematically exploring the perturbation space to identify critical input features. Similarly, Dineen et al. [18] have proposed unified metrics that evaluate model fidelity and interpretability using perturbation methods, further highlighting the effectiveness of this approach.

In our work, we leverage these methods to interpret the behavior of the modulation classification model by identifying and quantifying the importance of specific regions in the input constellation diagrams.

A. Perturbation Methodology

We implement two key perturbation strategies designed to uncover critical features in the input constellation diagrams. The perturbation process uses percentile-based thresholds for intensity masking. Let $I(x, y)$ represent the intensity of pixel (x, y) in the constellation diagram. For a given perturbation percentage p, the thresholds for masking are calculated as follows:

1) **Masking High-Intensity Regions, Top $p\%$ Brightest Pixels**: The brightest regions of the constellation diagram, corresponding to the highest signal amplitudes, are identified and set to zero. Specifically, pixels with intensities higher than or equal to the 100 - p percentile are masked by setting $I(x, y) = 0$.
2) **Masking Low-Intensity Non-Zero Regions, Bottom $p\%$ Non-Zero Pixels**: The least bright but non-zero regions, representing subtle and often overlooked features, are identified and set to zero. Specifically, pixels with intensities that are greater than 0 and lower than or equal to the p percentile are masked by setting $I(x, y) = 0$. This analysis quantifies the model's reliance on less prominent yet potentially crucial features.

As described in [16] and [17], perturbation-based methods excel in their ability to dynamically query and analyze model behavior across various feature hierarchies. By systematically perturbing high- and low-intensity regions, our methodology provides a comprehensive understanding of how the model interprets and prioritizes different aspects of the input.

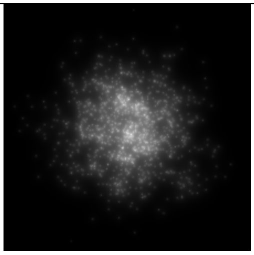
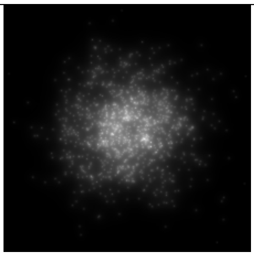
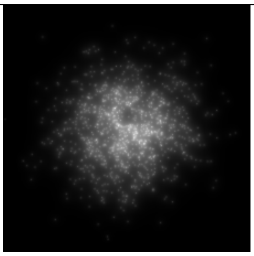
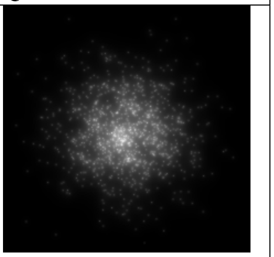
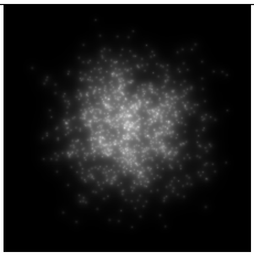
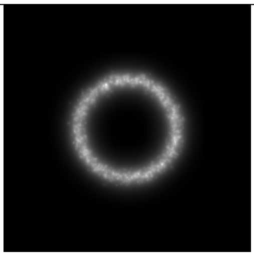
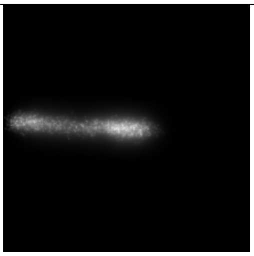
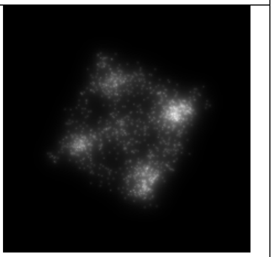
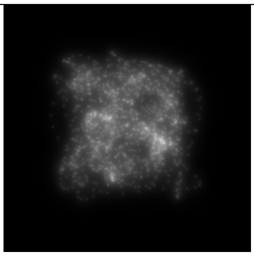
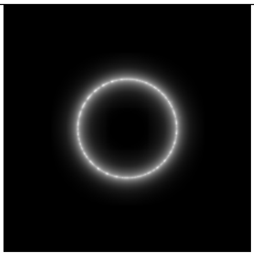
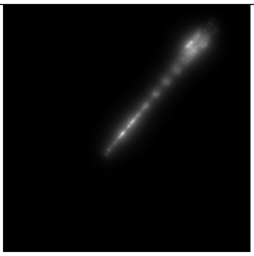
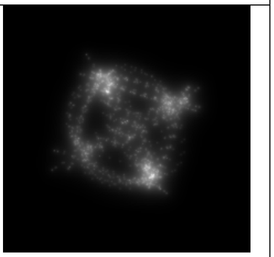
B. Metrics for Evaluating Perturbation Impact

To quantitatively assess the effects of perturbations, we employ the following metrics, inspired by the evaluation practices outlined in [16] and [18]:

1) **Classification Accuracy Drop**: This metric measures the change in accuracy (ΔA) before and after perturbation as given in Eq. 6,

$$\Delta A = A_{\text{original}} - A_{\text{perturbed}} \tag{6}$$

TABLE I: Examples of Constellation Diagrams for Different Modulation Types and SNR Levels

SNR Level	64-QAM	FM	OOK	QPSK
Low SNR (-20 dB)				
Medium SNR (10 dB)				
High SNR (30 dB)				

where A_{original} is the accuracy prior to perturbation, and $A_{\text{perturbed}}$ is the accuracy post-perturbation. While ΔA is generally expected to be positive (indicating a drop in accuracy due to perturbation), it is not guaranteed. If the perturbation improves performance, ΔA could be negative.

2) **Perturbation Impact Score (PIS)**: This score quantitatively assesses the impact of perturbation by normalizing the change in accuracy ΔA by the proportion of the input that was perturbed as depicted in Eq. 7,

$$\text{PIS} = \frac{\Delta A}{f} \tag{7}$$

where f represents the fraction of the input data that was altered. Since f is the fraction of the perturbed input, it is always between 0 and 1. A high PIS can result if (1) ΔA is high indicating a notable change in accuracy due to perturbation and/or (2) f is very low suggesting that even a small perturbation has a substantial effect on performance thereby highlighting the importance of the affected regions.

C. Insights from Perturbation Analysis

As Ivanovs et al. [16] emphasize, perturbation-based methods enable researchers to directly link input features to model outputs in an interpretable and robust manner. Similarly, Fel et al. [17] demonstrated how exhaustive perturbation analysis ensures reliability in explainability results. Also, Dineen et al. [18] proposed advanced metrics to unify static and dynamic scenarios.

By systematically perturbing specific regions of the constellation diagrams, our framework reveals the significance of high- and low-intensity regions, offering actionable insights into feature importance and model behavior. This transparency not only enhances trust in the model but also facilitates further optimization of the training pipeline.

Moreover, the dynamic and iterative nature of our perturbation analysis aligns with the best practices for exploring and interpreting deep learning models in safety-critical applications. This work contributes to the broader field of explainable artificial intelligence by demonstrating the utility of perturbation-based methods in a novel domain.

D. Visualizing Perturbed Constellation Diagrams

To better understand the impact of perturbations, Figure 1 compares an unperturbed constellation diagram with diagrams where 5% of the brightest and dimmest regions are masked. The visual differences highlight the critical role of high-intensity regions in the model's classification accuracy.

V. Experimental Setup

A. Dataset

We specifically use the RadioML 2018.01A dataset which is a widely used benchmark for AMC research. Although the dataset is synthetically generated using GNU Radio, it (1) offers a controlled environment for benchmarking algorithms

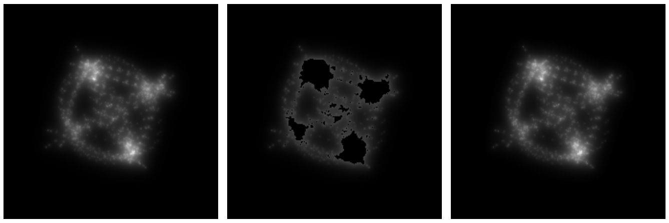


Fig. 1: Comparison of QPSK constellation diagrams at 30 dB SNR: (Left) Unperturbed, (Center) Bright regions perturbed (top 5%), (Right) Dim regions perturbed (bottom 5%).

and (2) poses a challenging testbed in bridging the gap to real-world wireless scenarios by including diverse modulation types and noise conditions. RadioML 2018.01A contains 24 modulation types, including both digital (e.g., BPSK, QPSK, QAM) and analog (e.g., AM, FM) schemes, spanning 26 SNR levels from -20 dB to 30 dB.

For this study, we focus on 20 modulation types and bucket the SNR levels into three predefined SNR buckets as already discussed. Each modulation type is represented by 4096 frames, with each frame consisting of 1024 complex-valued time-series samples.

To transform the raw I/Q data into inputs suitable for image-based neural network architectures, we generate constellation diagrams from the complex signals. These diagrams are augmented using smoothing and normalization techniques to enhance clarity and mitigate the effects of noise, ensuring robust classification in low-SNR scenarios. This preprocessing approach aligns with the methods proposed by Sun and Wang [19], who demonstrated the efficacy of constellation diagram transformation in extracting distinct modulation patterns under noise conditions.

B. Model Architecture

Our model, named `ConstellationResNet`, is based on the ResNet family of architectures and incorporates several modifications to tailor it for the dual tasks of modulation classification and SNR prediction. The details of the architecture are as follows:

- **Base Network**: The model uses the `resnet18` architecture from `torchvision`, initialized with pretrained weights. This ResNet model is well-suited for extracting spatial features from the constellation diagrams.
- **Input Channels**: The first convolutional layer is modified to handle single-channel grayscale inputs instead of the default three-channel RGB inputs. This ensures compatibility with the one-channel constellation diagrams while retaining the ResNet model's feature extraction capabilities.
- **Feature Extraction and Shared Transformation**: The output of the ResNet backbone is processed by a shared fully connected layer (`shared_transform`), which reduces the feature dimension to one-fourth of its original size. This layer is followed by a ReLU activation function, BatchNorm, and a dropout layer with a probability of 0.6 to prevent overfitting.
- **Task-Specific Heads**: Two independent fully connected layers are used as task-specific heads. The `modulation_head` predicts modulation types (20 classes), and the `snr_head` predicts SNR buckets (3 classes). These heads ensure that the model can perform the dual tasks of modulation classification and SNR prediction simultaneously.

The use of Multi-Task Learning (MTL) in our architecture is motivated by its ability to improve feature sharing and reduce overfitting, as discussed in a comprehensive survey [20]. Furthermore, it is demonstrated in [21] that MTL is effective in improving wireless communication tasks, particularly in edge deployments.

The constellation diagrams are preprocessed as grayscale images, which are resized to fit the ResNet architecture's input size. These grayscale images are fed into the network as one-channel tensors, leveraging the modified convolutional layers for feature extraction.

The model takes a single-channel constellation diagram as input and outputs two predictions:

1) **Modulation Output**: A vector of size 20, representing the probabilities of each modulation class.
2) **SNR Output**: A vector of size 3, representing the probabilities of each SNR bucket.

C. Training Configuration

The training regime was designed to evaluate the performance of the proposed model while ensuring robust generalization across various SNR ranges and modulation types. The key aspects of the training setup are as follows:

- **Train/Validation Split**: The dataset is divided into 80% training and 20% validation data, ensuring sufficient samples for both model optimization and performance evaluation.
- **Model Architecture**: We employ a ResNet-18 architecture initialized with pretrained weights. The architecture is modified to handle single-channel grayscale inputs and incorporates a dropout layer with a probability of 0.6 for regularization.
- **Optimizer and Learning Rate**: The Adam optimizer is used to train the model with a base learning rate of 1×10^{-4}. A ReduceLROnPlateau scheduler is employed to adapt the learning rate dynamically based on validation loss.
- **Batch Size**: A batch size of 512 is used for both training and validation phases, balancing computational efficiency and convergence stability.
- **Number of Epochs**: The model is trained for 50 epochs, with early stopping based on validation performance to prevent overfitting.

The explainability framework implemented in this study builds upon the principles of robust perturbation analysis, as demonstrated by Fel et al. [17]. Their work highlights the importance of evaluating the fidelity and robustness of explainability mechanisms, which informed our choice of metrics and validation procedures.

D. Evaluation Metrics

To comprehensively assess model performance, we evaluate the following metrics:

- **Classification Accuracy**: Overall accuracy for predicting modulation types and SNR buckets.
- **Explainability Metrics**: As part of the perturbation-based analysis, we track metrics such as change in accuracy ΔA, PIS, and fidelity of explanations to evaluate how the model relies on specific regions of the constellation diagram for classification.

The PIS was adapted from techniques discussed by Gwak et al. [22], where similar metrics were used for robust fault diagnosis in industrial systems. This ensures that our evaluation framework aligns with state-of-the-art explainability practices.

VI. Results and Discussion

A. Model Performance on the RadioML 2018.01A Dataset

The proposed `ConstellationResNet` model demonstrates strong performance on the RadioML 2018.01A dataset, achieving high classification accuracy across 20 modulation types. Figure 2 presents the normalized confusion matrix for the model, highlighting its performance on each modulation class. The results show that the multi-task learning framework, combined with SNR bucketing and perturbation-based explainability, effectively classifies modulation types even under challenging noise conditions.

The key observations from the confusion matrix are:

1) **High Accuracy for Simple Schemes:** Modulation types such as QPSK (93%), OQPSK (89%), and BPSK (89%) exhibit high classification accuracy due to their distinct constellation patterns and minimal overlap.
2) **Challenges with Higher-Order Modulations:** Performance for complex schemes like 64QAM (66%) and 256QAM (79%) shows a notable drop, reflecting the difficulty in distinguishing densely packed constellation points, particularly under low SNR conditions.
3) **Resilience to Intermediate Modulations:** Intermediate schemes such as 16PSK (86%) and 32APSK (91%) demonstrate balanced performance, suggesting that the model effectively learns from hybrid amplitude-phase relationships.
4) **Misclassification Trends:** Misclassifications tend to occur between modulation types with similar constellation patterns, such as higher-order QAM schemes, reinforcing the need for more robust feature extraction or augmentation techniques.

B. Multi-Task Learning Enhancements

The incorporation of MTL significantly improves model performance by enabling joint training on modulation classification and SNR estimation. Figures 3a, 3b, and 3c illustrate the validation accuracies for SNR, modulation classification, and their combined impact, respectively. The graphs clearly demonstrate that the multi-task learning approach leads to superior results for SNR prediction and combined tasks, although modulation classification by itself slightly favors the single-task approach.

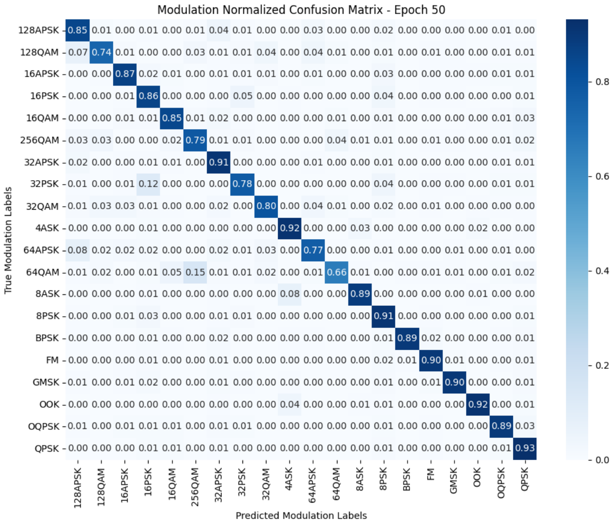
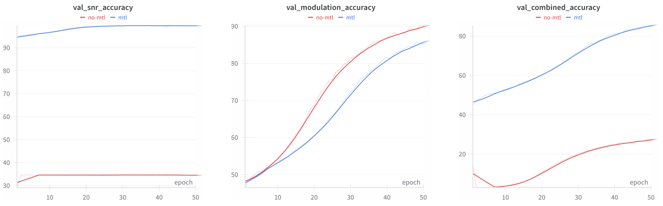


Fig. 2: Confusion Matrix for Modulation Classification across all SNR levels. Higher misclassification rates for complex modulation schemes such as 64QAM and 256QAM are particularly influenced by lower SNR conditions, where densely packed constellation points become harder to distinguish. The band on the right represents the normalized percentage of predictions for each class, indicating the distribution of classification results.



(a) SNR validation accuracy

(b) Modulation validation accuracy

(c) Combined validation accuracy

Fig. 3: Validation accuracy comparison for (a) SNR prediction, (b) modulation classification, and (c) combined tasks under multi-task and single-task learning. Multi-task learning (blue) vs. single-task learning (red).

After 50 epochs, the validation loss for the multi-task model was 0.23, compared to 0.293 for a single-task model trained solely on modulation classification. This lower validation loss for the multi-task model indicates better convergence and generalization, which translates to higher validation accuracy. As shown in Figure 3, the MTL approach consistently outperforms the single-task model in SNR prediction and combined task accuracy. The relationship between the reduced validation loss and improved accuracy is a result of the MTL framework's ability to leverage shared representations to enhance generalization across tasks.

The SNR prediction accuracy (Figure 3a) demonstrates a significant advantage for the MTL approach, showcasing its strength in capturing the interdependence between tasks. Similarly, the combined accuracy (Figure 3c) emphasizes how MTL integrates the interrelated nature of modulation classification and SNR estimation to improve robustness. The lower loss values align with these trends, reinforcing the effectiveness of MTL in jointly optimizing multiple objectives.

Interestingly, while the single-task model exhibits slightly better performance in modulation classification alone (Figure 3b), the overall combined benefits of MTL outweigh this minor trade-off. As noted in Yang [20], MTL allows knowledge transfer between related tasks, which not only reduces overfitting but also enables the model to focus on features that are critical to both tasks. This leads to more robust representations, particularly under scenarios with limited data or challenging conditions, as evidenced by the superior performance of MTL in these results.

C. Comparison with Prior Work

We compare `ConstellationResNet` with NMformer [23], a transformer-based model for AMC using vision transformers. NMformer is evaluated on RGB constellation diagrams and fine-tuned on subsets of data to improve performance under low and high SNRs. Table II summarizes the key differences in accuracy between `ConstellationResNet` and NMformer. Compared to NMformer, `ConstellationResNet` demonstrates:

- **Superior Performance:** Higher accuracy across all SNR scenarios, particularly for low and medium SNRs, attributed to the integration of SNR bucketing and the specialized multi-task architecture.
- **Enhanced Explainability:** Unlike NMformer, which lacks explainability mechanisms, `ConstellationResNet` incorporates perturbation-based explainability to identify critical regions in the constellation diagrams.
- **Efficiency with Single-Channel Input:** While NMformer processes RGB images, `ConstellationResNet` achieves superior results using grayscale constellation diagrams, reducing input dimensionality and computational complexity.

D. Evaluation Under Perturbation Scenarios

To assess the explainability and robustness of the model, experiments were conducted by perturbing key regions of the constellation diagrams. Table III summarizes the results, showing the modulation accuracy, SNR accuracy, combined accuracy, and PIS under various perturbation scenarios. The perturbation analysis shows:

- **PIS for Unperturbed Dataset:** For the unperturbed dataset, the accuracy remains unchanged, resulting in $\Delta A = 0$. Since $f = 0$ in this case (see Eq. 7), the Perturbation Impact Score (PIS) is undefined and is marked as "N/A" in Table III.
- **PIS for 50% Masked Datasets:** When 50% of the brightest or dimmest regions are masked, the large perturbation encompasses much of the input data, causing a near-complete loss of distinguishing features. In this scenario, calculating a meaningful PIS is infeasible, and thus it is also marked as "N/A" in Table III.
- **Decreasing PIS Trends for Bright and Dim Regions:** For both the brightest and dimmest regions, the PIS decreases as the proportion of pixels masked increases from 1% to 5%. This is because the additional pixels masked within these ranges have diminishing importance. For example, masking 1% of the brightest pixels has a disproportionately large impact since these regions contribute the most critical information for classification. As the masked percentage increases, the impact of the additional masked pixels becomes less pronounced, leading to a lower PIS.
- **Impact of Bright Regions:** Masking high-intensity regions significantly affects performance, with modulation accuracy dropping from 91.01% to 20.54% when 5% of the brightest regions are perturbed. This corresponds to a PIS of 14.10, highlighting the critical role these regions play in distinguishing modulation schemes.
- **Impact of Dim Regions:** Perturbing the dimmest regions has a minimal effect on performance, with the PIS as low as 0.95 for 5% of the dimmest regions. This indicates that these regions contribute little to the model's decision-making process, making them less impactful for classification accuracy.

VII. Summary and Conclusions

In this paper, we proposed a novel and robust framework for AMC that integrates constellation diagram augmentation, multi-task learning, and perturbation-based explainability. By transforming raw I/Q signals into enhanced constellation diagrams and leveraging multi-task learning, we simultaneously classify modulation types and estimate SNR levels across varying noise conditions. Compared to traditional methods and transformer models, our framework achieves competitive accuracy while offering direct interpretability.

The integration of perturbation-based explainability experiments show that the brightest regions are critical for classification accuracy, while the dimmest regions contribute minimally thereby highlighting the importance of high-intensity features.

References

[1] T. Huynh-The, Q.-V. Pham, T.-V. Nguyen, T. T. Nguyen, R. Ruby, and M. Zeng, "Automatic modulation classification: A deep architecture survey," *IEEE Access*, vol. 9, pp. 142 950–142 971, 2021.

[2] 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, 2021.

[3] Y. Sun and E. A. Ball, "Automatic modulation classification using techniques from image classification," *IET Communications*, 2022.

TABLE II: Comparison of `ConstellationResNet` with NMformer Across SNR Scenarios. Results are shown for three representative SNR levels (-20 dB, 10 dB, and 30 dB) to illustrate performance trends. Similar trends were observed across other SNR values.

Model	**Accuracy (Low SNR, -20 dB)**	**Accuracy (Medium SNR, 10 dB)**	**Accuracy (High SNR, 30 dB)**	**Explainability Integration**
NMformer (Base Classifier)	71.10%	72.12%	75.77%	No
NMformer (Fine-tuned, Low SNR)	71.60%	70.51%	71.87%	No
`ConstellationResNet` (This Work)	**76.40%**	**85.20%**	**91.01%**	Yes (Perturbation-based)

TABLE III: Performance Metrics Under Different Perturbation Scenarios

Dataset	**Modulation Accuracy (%)**	**SNR Accuracy (%)**	**Combined Accuracy (%)**	**PIS**
Unperturbed	91.01	99.74	90.80	N/A
Top 1% Brightest Masked	56.32	89.09	51.16	34.69
Top 2% Brightest Masked	41.83	82.68	36.14	24.59
Top 5% Brightest Masked	20.54	78.09	16.83	14.10
Top 50% Brightest Masked	5.01	34.63	1.72	N/A
Bottom 1% Dimmest Masked	86.16	99.51	85.79	4.85
Bottom 2% Dimmest Masked	86.15	99.51	85.78	2.43
Bottom 5% Dimmest Masked	86.28	99.51	85.90	0.95
Bottom 50% Dimmest Masked	43.42	83.26	37.80	N/A

[4] E. E. Azzouz and A. K. Nandi, “Algorithms for automatic modulation recognition of communication signals,” *IEEE Transactions on Communications*, vol. 43, pp. 431–436, 1995.

[5] L. J. Wong and S. McPherson, “Explainable neural network-based modulation classification via concept bottleneck models,” in *2021 IEEE 11th Annual Computing and Communication Workshop and Conference (CCWC)*. IEEE, 2021, pp. 0191–0196.

[6] E. E. Azzouz and A. K. Nandi, “Automatic identification of digital modulation types,” *Signal processing*, vol. 47, no. 1, pp. 55–69, 1995.

[7] A. Kumar, K. K. Srinivas, and S. Majhi, “Automatic modulation classification for adaptive ofdm systems using convolutional neural networks with residual learning,” *IEEE Access*, vol. 11, pp. 61 013–61 024, 2023.

[8] Y. Kong, H. Zhang, X. Li, and B. Zhang, “Nmformer: A transformer for noisy modulation classification in wireless communication,” in *Proceedings of the IEEE International Conference on Communications*, 2023.

[9] 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.

[10] J. Zhao, H. Wang, S. Peng, and Y.-D. Yao, “Meta supervised contrastive learning for few-shot open-set modulation classification with signal constellation,” *IEEE Communications Letters*, 2024.

[11] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, “Grad-cam: Visual explanations from deep networks via gradient-based localization,” in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 618–626.

[12] M. Sundararajan, A. Taly, and Q. Yan, “Axiomatic attribution for deep networks,” in *International conference on machine learning*. PMLR, 2017, pp. 3319–3328.

[13] R. C. Fong and A. Vedaldi, “Interpretable explanations of black boxes by meaningful perturbation,” in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 3429–3437.

[14] 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 International Conference on Knowledge Discovery and Data Mining*, ser. KDD ’16. New York, NY, USA: Association for Computing Machinery, 2016, p. 1135–1144. [Online]. Available: https://doi.org/10.1145/2939672.2939778

[15] A. Kapishnikov, T. Bolukbasi, F. Viégas, and M. Terry, “Xrai: Better attributions through regions,” in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 4948–4957.

[16] M. Ivanovs, R. Kadikis, and K. Ozols, “Perturbation-based methods for explaining deep neural networks: A survey,” *Pattern Recognition Letters*, vol. 150, pp. 228–234, 2021. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0167865521002440

[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*, 2023, pp. 16 153–16 163.

[18] J. Dineen, D. Kridel, D. Dolk, and D. Castillo, “Unified explanations in machine learning models: A perturbation approach,” *arXiv preprint arXiv:2405.20200*, 2024.

[19] S. Sun and Y. Wang, “A novel deep learning automatic modulation classifier with fusion of multichannel information using gru,” *EURASIP Journal on Wireless Communications and Networking*, vol. 2023, no. 1, p. 66, 2023.

[20] Y. Zhang and Q. Yang, “A survey on multi-task learning,” *IEEE transactions on knowledge and data engineering*, vol. 34, no. 12, pp. 5586–5609, 2021.

[21] A. Jagannath and J. Jagannath, “Multi-task learning approach for modulation and wireless signal classification for 5g and beyond: Edge deployment via model compression,” *Physical Communication*, vol. 54, p. 101793, 2022.

[22] M. Gwak, M. S. Kim, J. P. Yun, and P. Park, “Robust and explainable fault diagnosis with power-perturbation-based decision boundary analysis of deep learning models,” *IEEE Transactions on Industrial Informatics*, vol. 19, no. 5, pp. 6982–6992, 2022.

[23] A. Faysal, M. Rostami, R. G. Roshan, H. Wang, and N. Muralidhar, “Nmformer: A transformer for noisy modulation classification in wireless communication,” *arXiv preprint arXiv:2411.02428*, 2024.