AMC-Transformer: Automatic Modulation Classification based on Enhanced Attention Model

Yuewen Xu

Abstract—**High-accuracy automatic modulation classification (AMC) is essential for spectrum monitoring and interference-aware access in future 6G systems [1]. We propose AMC-Transformer, which tokenizes raw I/Q sequences into fixed-length patches, augments them with learnable positional embeddings, and applies multi-layer, multi-head self-attention to capture global temporal–spatial correlations without handcrafted features or convolutions. On RadioML2018.01A, our model achieves 98.8% accuracy in the high-SNR regime (SNR at least 10 dB), showing higher accuracy than a CNN and a ResNet reimplementation by 4.44% and 1.96% in relative terms; averaged across all SNRs, it also improves upon MCformer, CNN, and ResNet baselines. Consistent gains are observed on the RadioML2016.10A dataset, further validating robustness across benchmarks. Ablations on depth, patch size, and head count provide practical guidance under different SNR regimes and compute budgets. These results demonstrate the promise of transformer-based AMC for robust recognition in complex wireless environments.**

Index Terms—**Modulation Recognition, Deep Learning, Transformer, Attention Mechanism, IQ Signal.**

I. INTRODUCTION

Wireless signal recognition—also known as automatic modulation classification (AMC)—is pivotal across military and civilian scenarios. It enables identifying modulation types from raw RF signals under limited prior knowledge, supporting dynamic spectrum access (DSA), interference detection, spectrum monitoring, and spectrum coexistence. Moving toward 6G, AMC becomes even more critical for improving spectrum utilization, robustness, and low-overhead (pilot-free) communications. [2–3]

Classical AMC approaches fall broadly into *likelihood-based (LB)* and *feature-based (FB)* families. LB methods (e.g., ML/EM-assisted inference, HLRT/QHLRT variants) can achieve high accuracy in favorable conditions but are often sensitive to channel state information and carry significant complexity. FB methods draw on expert features such as cyclostationary statistics and higher-order cumulants, offering lower complexity and near-optimal performance for lower-order schemes, yet they struggle in multipath, overlapping sources, and high-order modulations. [4–7]

Deep learning (DL) has boosted AMC in both supervised CNNs and newer foundation-model-style approaches [8]. CNNs exploit multi-scale structures and constellation geometry effectively (e.g., constellation-image CNNs; robust multi-scale designs under synthetic channel impairments). ResNet-style networks further mitigate vanishing gradients and improve feature reuse; lightweight/binarized ResNets demonstrate competitive accuracy–efficiency trade-offs for edge deployment. [9–13]

Transformers (TRN) provide an alternative by modeling long range dependencies via self-attention with efficient parallelism. Beyond their foundational success, time series surveys highlight their strengths for sequence tasks. In AMC, Transformer variants such as MCformer, CNN Transformer hybrids, and CNN Transformer GNN adaptively weight multi scale patterns and improve robustness and scalability in non-cooperative settings [14-18].

We propose the AMC-Transformer, a transformer-based model designed for time-series IQ samples to improve AMC accuracy. Key contributions include:

1. Learnable Embedding of RF Signal Patches: To represent the time-series nature of raw IQ samples, we design a learnable embedding strategy that combines patch and positional information. RF signals are segmented into fixed-size patches, mapped into the feature space via an MLP, and augmented with positional embeddings to preserve temporal dependencies, enabling effective attention-based modeling.

2. Attention Mechanism on Raw Time-Series IQ Data: We apply self-attention directly to raw IQ data, enabling the model to capture long-range dependencies. This enables the model to capture long-range dependencies and temporal–spatial correlations, extracting global representations without handcrafted features or convolutional operations, thus overcoming the locality limitations of CNN-based methods.

3. Enhanced Diversity and Robustness with Multi-Head Attention: To improve generalization under varying SNR conditions, we employ multi-head attention, allowing feature extraction in multiple subspaces. This enriches representation diversity and enhances robustness against noise and channel impairments, improving classification reliability in realistic wireless environments.

4. Competitive performance on public RadioML datasets: On RadioML2018.01A [19] we obtain 98.8 percent accuracy at SNR at least 10 dB and observe higher average accuracy than MCformer, CNN, and ResNet. Similar trends appear on RadioML2016.10a [19]. We also provide ablations on depth, patch size, and head count.

Department of Engineering, The University of Bristol, Bristol, UK
(E-mail: tu23081@bristol.ac.uk)

DOI: 10.36244/ICJ.2025.4.5

The proposed AMC-Transformer provides a robust attention-based solution for AMC, demonstrating competitive performance across multiple modulation scenarios. At the same time, the increased model complexity introduced by attention mechanisms highlights an inherent accuracy–complexity trade-off, which is particularly relevant for practical 6G deployments. Despite these advantages, transformer models face challenges such as quadratic complexity, higher data requirements, and limited invariance to signal distortions. To enhance practicality and scalability, future work will focus on efficient attention mechanisms, hybrid Conv–Attention architectures, and RF-specific data augmentation strategies.

The paper is structured as follows: Section 2 summarizes related research. Section 3 details the AMC-Transformer architecture. Section 4 presents evaluation results. Section 5 concludes the paper.

II. RELATED WORK

Deep learning–based AMC has advanced markedly in recent years. On the CNN side, one-dimensional residual networks for I/Q sequences and complex-valued convolutions can extract discriminative features while keeping parameter counts manageable; for example, ResNet-style variants tailored to wireless signals and complex depthwise-separable CNNs report strong results on RadioML benchmarks [20,21]. Meanwhile, MCNet—using asymmetric kernels and skip connections—achieves about 93% accuracy at high SNR (20 dB) on RadioML2018.01A, illustrating the upper bound of CNNs in high-SNR regimes [22]. RNN/CRNN and LSTM models have also been used to capture long-range dependencies, but their generalization to unseen channel conditions and modulation parameters remains limited [23–25]. RadioML datasets (e.g., RML2018.01A, RML2016.10a/10b) continue to be the standard benchmarks in this area.

Transformer-based AMC has recently evolved in three directions. First, sequence models operating directly on raw I/Q: Cai et al. apply a Transformer to AMC and report consistent gains over CNN/LSTM baselines—especially at low SNR—with fewer parameters; MCformer embeds each (I, Q) sample via a lightweight 1-D convolution and stacks Transformer encoders, with the notable observation that omitting positional encodings works better; it attains state-of-the-art accuracy on RML2016.10b with only ~10k–72k parameters [16]. Second, hybrid CNN–Transformer designs: CTGNet/CTRNet use convolutions for local invariances and self-attention for long-range dependencies, improving robustness under multiple impairments and non-idealities [26]. Third, ViT on 2-D signal representations: by converting signals to constellation images, MobileViT and related ViT variants improve robustness under noise without an explicit denoising pipeline (e.g., NMformer) [27,28].

In addition, scalability and label efficiency have been advanced via meta-learning and semi/self-supervision: Meta-Transformer provides a general few-shot adaptation framework for previously unseen modulations, and subsequent studies further validate meta-learning for cross-domain generalization [29,30]. Transformer-based contrastive semi-supervised learning and self-supervised RF representation learning (e.g., Self-Contrastive, NextG RF SSL) substantially reduce labeled-data requirements while maintaining accuracy in low-label regimes [31–33].

III. MODEL DESCRIPTION

The AMC-Transformer is tailored to 2-D in-phase/quadrature (IQ) signals and addresses two challenges that limit conventional CNN/ResNet models on RF data: (i) high-frequency noise and (ii) long-range temporal dependencies. As summarized in Fig. 1, the model converts a 2×1024 IQ sample into fixed-length tokens via patching, augments them with positional encodings, and processes the sequence using a Transformer encoder whose output feeds an MLP head for prediction.

A. Input Processing

We adopted a minimal, task-compatible preprocessing pipeline per sample $x \epsilon R^{2\times 1024}$: including: (i) per-channel DC offset removal, (ii) RMS normalization across I/Q channels, and (iii) channel-wise z-score standardization using statistics estimated from the training split only. The same normalization parameters are then applied to validation and test data to avoid information leakage. Explicit filtering or denoising is intentionally avoided to preserve modulation-discriminative spectral and phase characteristics.

$$x'_{c,t} = x_{c,t} - \text{mean}_t\left(x_{c,t}\right)$$
$$\tilde{x}_{c,t} = \frac{x'_{c,t}}{\sqrt{\frac{1}{2T}\sum_c \sum_t \left(x'_{c,t}\right)^2 + \varepsilon}} \quad (1)$$
$$\hat{x}_{c,t} = \frac{\tilde{x}_{c,t} - \mu_c}{\sigma_c + \varepsilon}$$

where (μ_c, σ_c) are the channel-wise mean and standard deviation estimated on the training split after steps (i)–(ii) and then fixed for validation and test sets, and ε is a small constant for numerical stability.

B. Overall Pipeline

The preprocessed signal $\hat{x}$ is treated as a two-channel 2-D array 2×1024. We tokenize it into $N = 1024$ non-overlapping $2 \times P$ patches (covering both I and Q to retain I/Q coherence), add learned positional embeddings, and process tokens with a stack of Transformer encoder blocks (multi-head self-attention and MLP, each preceded by layer normalization and followed by dropout). An MLP head produces the final logits. Fig. 1 is updated to include the preprocessing block.

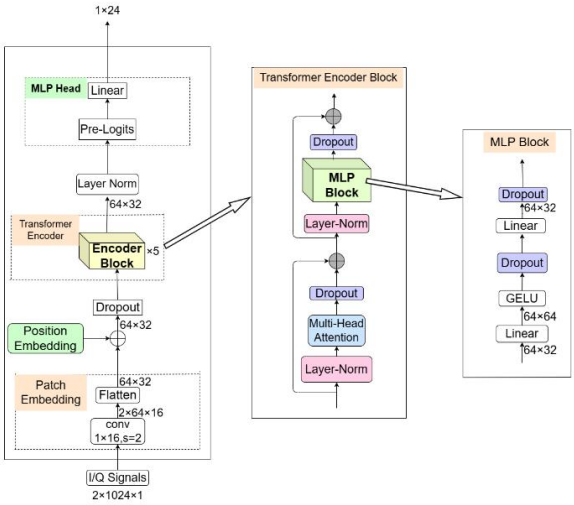

Fig. 1. Model Overview

C. Patch Embedding

To balance context coverage and efficiency, the input is partitioned along the time axis into non-overlapping patches of size 1×16, yielding 64 patches per sample. Each patch contains 2×16 values that are flattened to a 32-D vector and linearly projected to the model width d (we use $d = 32$), producing a token sequence of shape $64 \times d$. This tokenization lets the self-attention mechanism relate local structures (e.g., short-term amplitude/phase transitions) to global patterns influenced by modulation type and SNR. The choice 16 ensures sufficient local resolution under symbol-rate offsets, delays, and noise fluctuations while keeping compute tractable.

D. Positional Encoding

Because Transformer blocks process all tokens in parallel, explicit position information is required to capture temporal dynamics (amplitude, frequency, and phase evolutions). We use learnable positional embeddings of length 64 and dimension d; the positional vector for each patch is added to its token embedding, yielding $E \in R^{64\times d}$. This enables the model to distinguish early/late patches and to learn temporal patterns associated with different modulations and SNRs.

E. Self-Attention

Given the token matrix E, the encoder computes query, key, and value projections

$$Q = EW_Q$$
$$K = EW_K$$
$$V = EW_V$$

where $Q, K \in R^{64\times d_k}$ and $V \in R^{64\times d_v}$.

Scaled dot-product attention (Fig. 2) is

$$Attention(Q,K,V) = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V \qquad (2)$$

The output $Z \in R^{64\times d_v}$ preserves sequence length while re-weighting each token by its global correlations. For token i, the vector z_i aggregates values v_j according to the similarity between q_i and k_j, thereby encoding long-range dependencies across the entire 2×1024 signal.

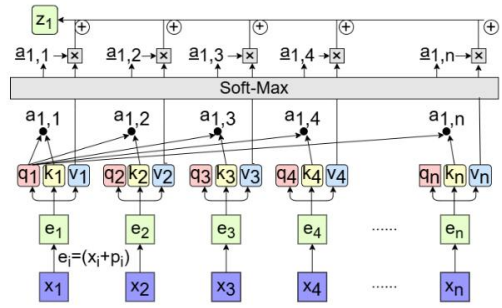

Fig. 2. Scaled Dot-Product Attention

F. Multi-Head Attention and Classifier

To learn complementary temporal and frequency relations. For head i,

$$Q_i = QW_i^Q, K_i = KW_i^K, V_i = VW_i^V, i = 1,2 \qquad (3)$$

with per-head dimension $d_k = d_v = 8$. Each head produces the heads are concatenated and projected:

$$MultiHead(Q,K,V) = Concat(head_1, head_2)W^O \qquad (4)$$

$$head_i = Attention\left(QW_i^Q, KW_i^K, VW_i^V\right) \qquad (5)$$

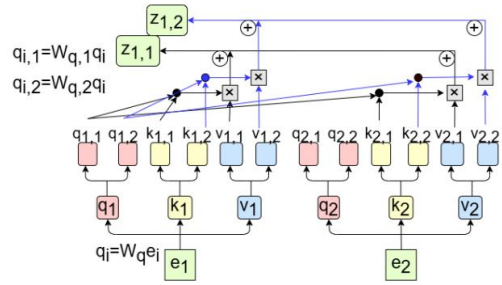

Fig.3. Multi-Head Attention

restoring the model width $d = 32$. Each encoder block applies LayerNorm, multi-head self-attention, dropout, and an MLP with residual connections (see Fig. 3). The encoder output is passed to an MLP head (linear layers with normalization/dropout) to obtain pre-logits and final predictions (e.g., via softmax for classification)

IV. EXPERIMENT AND RESULTS

A. Datasets and Split Protocol

We use the publicly available RadioML 2018.01A dataset, which synthesizes realistic channel impairments—including delay spread, carrier frequency offset, and thermal noise. It covers 24 modulation types across 26 SNR levels from −20 dB to 30 dB in 2 dB steps. Each SNR level contains 4,096 signal examples, yielding a total of 2,555,904 samples, each represented as complex IQ (in-

phase and quadrature) sequences. In addition to training and validation on 2018.01A, we perform cross-dataset evaluation on RadioML 2016 to assess the algorithm's robustness and generalization under distribution shifts and differing channel conditions.

To prevent sample leakage across SNR conditions, we perform a group-aware split of RadioML2018.01A. Let each waveform be indexed by its modulation $m \in M$, SNR $s \in S$, and within-class index $i \in \{0, \ldots, 4095\}$. We define a group as

$$G_{m,i} = \{(m, s, i)|\}$$

i.e., the same base waveform rendered at all SNRs for a given modulation. Splitting is conducted at the group level so that no group $G_{m,i}$ appears in more than one subset, eliminating leakage where the same underlying waveform at different SNRs would otherwise straddle train and evaluation sets.

We adopt a group-aware split to avoid cross-SNR leakage: within each modulation, base examples are grouped across all SNRs and treated as indivisible units. Groups are randomly assigned to 70% / 20% / 10% train/validation/test with a fixed seed (48), stratified by modulation to preserve class priors. Because each group spans the full SNR set, the SNR distribution is preserved across splits by construction. At the per-(modulation, SNR) level this yields approximately 2867 / 819 / 410 samples for train/val/test, respectively (rounded from 4096 per pair).

Class IDs are remapped to [0, 23] following the fixed 24-class list in Sec. A, ensuring a stable label order aligned with the classifier's output layer.

We apply lightweight, task-compatible preprocessing: (i) per-sample DC offset removal on I/Q channels; (ii) per-sample RMS normalization (AGC-style) to unit average power across I/Q; and (iii) channel-wise z-score standardization using training-split statistics (μ_c, σ_c) only. The same normalization parameters are then applied to validation and test data to avoid information leakage. Random seeds and the exact split indices are fixed and recorded to ensure reproducibility.

B. Baselines and training protocol

We re-train all baselines (CNN[34], ResNet[3], MCformer[16]) and our AMC-Transformer under the same preprocessing and training protocol. All models take identical inputs (I/Q, shape 2×1024), use the same loss (multiclass cross-entropy), optimizer and learning-rate schedule, and share the same group-aware data split. Specifically, within each modulation, base examples are grouped across all SNRs, and each group is assigned wholly to train/validation/test (70/20/10), which prevents cross-SNR leakage while preserving class priors. Early stopping and weight decay are applied to mitigate overfitting.

CNN: A 2D ConvNet consisting of four sequential stages including ABlock, BBlock, CBlock1, and CBlock2, followed by global average pooling and a 24-way classifier. The model contains 66,008 parameters.

ResNet: A 1D ResNet with residual connections, featuring an initial Conv1D layer followed by 5 residual blocks with progressive channel expansion from 32 to 64 to 128 channels. The architecture uses kernel size 7, batch normalization, and ReLU activations. The final layers consist of global average pooling followed by dropout and a dense classifier. The model contains 534,104 parameters.

MCformer: A hybrid architecture combining Conv1D with 8 channels and 4 lightweight encoder blocks, followed by temporal aggregation to 4 tokens. The output is processed through flattening, a 128-dimensional fully connected layer, and finally a 24-dimensional classification layer. The parameter counts increases from 10,050 for the original 10-class head to 11,856 for 24 classes, with the increase attributed to the expanded classifier.

AMC-Transformer (ours): The input is reshaped to dimensions 2×1024×1 and divided into 64 patches of 32 dimensions each. The architecture employs an embedding dimension of 96 with positional encoding, followed by 6 encoder layers with 8 attention heads each. The final process consists of flattening followed by a multi-layer perception with layer dimensions 6144, 2048, 1024, and 24. The model contains 15,834,680 parameters.

All models are trained using the AdamW optimizer with a learning rate of 1e-3, cosine decay scheduling with 5-epoch warm-up, weight decay of 1e-4, and gradient clipping at 1.0. The default batch size is 256, with additional results reported using batch size 800. Dropout of 0.1 is applied to MLP and classifier layers. Input data undergoes z-score normalization with no data augmentation applied. Hardware specifications, random seeds, and library versions are documented in the Appendix. All code and training scripts are provided to ensure reproducibility.

TABLE I
MODEL AND ARCHITECTURE OVERVIEW.

Model	Params	Tokens / Patch	Heads	Dim	Blocks
CNN	66,008	–	–	–	A/B/C ×4 → GAP → FC
ResNet-1D	534,104				Conv1D → 5×ResBlock → GAP → FC
MCformer-24 (reimpl.)	11,856	T-agg → 4	–	–	Conv1D → 4×Enc → FC
AMC-Trans (ours)	15,834,680	64 / 16	8	96	6×(MHA+FFN) → MLP

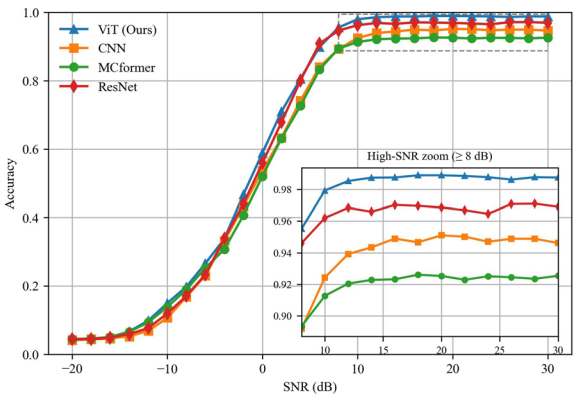

Fig. 4. Accuracy versus SNR on the RML2018.01A dataset

Over the full SNR range, AMC-Transformer achieves an average accuracy of 63.58%, higher than MCformer at 59.02%, CNN at 59.84%, and ResNet at 61.96% (Fig. 4.) In the high-SNR region (SNR of at least 10 dB), AMC-Transformer reaches 98.8% and exhibits a clear saturation plateau, higher than MCformer at 92.29%, CNN at 94.50%, and ResNet at 96.79%. In the low-SNR region (SNR at most -8 dB), AMC-Transformer attains 20.97%, higher than MCformer at 19.19%, CNN at 18.70%, and ResNet at 19.37%. In the mid-SNR range from 2 to 8 dB, AMC-Transformer averages 84.20%, higher than MCformer at 77.07% and CNN at 77.66%, and essentially on par with ResNet at 83.32%.

Overall, AMC-Transformer maintains robustness at low SNR and sustains a consistent performance margin as SNR increases, with a near-saturated accuracy around 98.8% on RML2018.01A in the high-SNR region.

C. Robustness Analyses

1. Accuracy Across SNRs on RML2016.10a

RML2016.10a (Fig. 5). Over the full SNR range, the average accuracies of AMC-Transformer and MCformer are essentially identical (63.48% and 63.48%). In the high-SNR region (SNR of at least 10 dB), AMC-Transformer attains an average accuracy of 93.51%, which is higher than MCformer by 0.82 percentage points and higher than CNN and ResNet by 3.00 and 6.37 percentage points, respectively. In the low-SNR region (SNR at most -10 dB), AMC-Transformer reaches 29.12%, comparable to MCformer at 29.06% and higher than CNN at 25.68% and ResNet at 23.70%

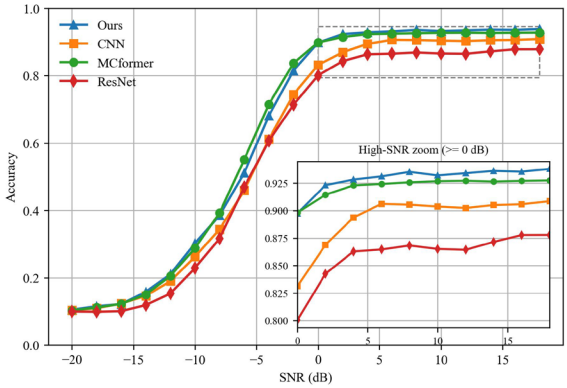

Fig. 5. Evaluation on RML2016.10a Across SNRs

2. Macro-F1 Stability Over Random Seeds

Having established performance trends across SNRs, we next test whether these gains persist under different random initializations. To address the concern that the proposed model only achieves high accuracy under favorable conditions, we further evaluate its robustness across the full SNR range. Fig. 6 shows the Macro-F1 scores from −20 dB to 30 dB, averaged over five independent runs with different random seeds. In addition to the high-SNR regime (20–30 dB), where AMC-Transformer attains near-saturation performance, the model maintains competitive robustness under mid and low SNR conditions. For example, at -10 dB and 0 dB, the Macro-F1 remains above 10.8% and 57.4%, respectively, with narrow confidence intervals, indicating stable generalization across noise levels. This result confirms that the performance of AMC-Transformer is not restricted to high SNRs but extends to more challenging communication environments as well.

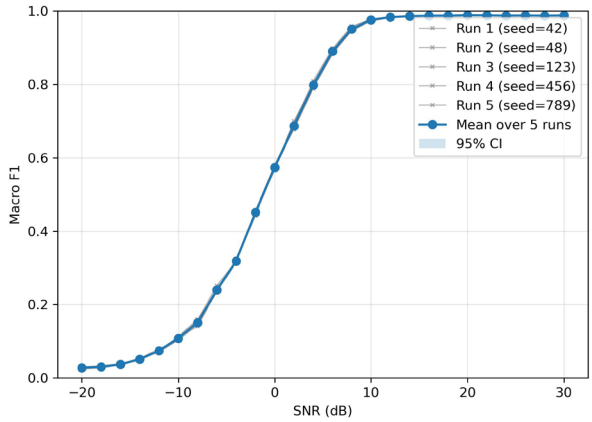

Fig. 6. Robustness to Random Initialization: Macro-F1 over Five Seeds.

3. Per-Class F1 Across SNRs

To further examine the robustness of the AMC-Transformer, we report per-class F1-scores across different SNR levels. The horizontal axis corresponds to SNR values (−20 dB to 30 dB), the vertical axis lists the 24 modulation types, and the color intensity indicates the F1-score.

Overall, fig.7 shows that F1-scores consistently increase with SNR. Low-order modulations such as BPSK and QPSK remain relatively robust even at low SNR (−10 dB), whereas higher-order QAM schemes suffer significant degradation under noise but quickly recover above 0 dB. Importantly, the model maintains competitive per-class F1 performance in the mid-SNR regime (0–10 dB), demonstrating that its effectiveness is not limited to high SNR conditions.

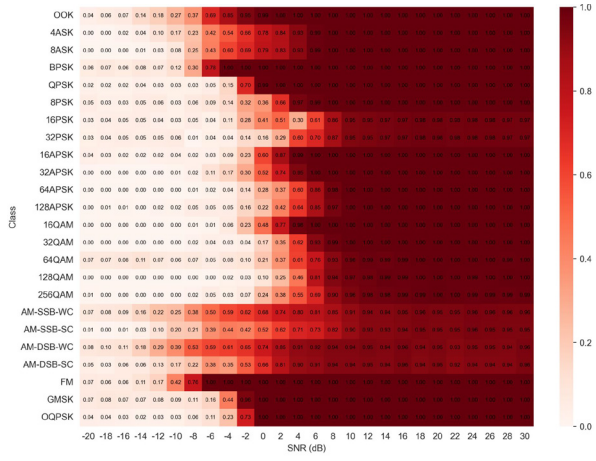

Fig. 7. Per-class F1-scores across SNR levels

D. Analysis of AMC-Transformer Model Parameter Tuning

To systematically evaluate the impact of key hyperparameters on AMC-Transformer performance, we conducted a parameter sensitivity study across four critical dimensions: batch size, transformer layer depth, patch size, and number of attention heads. The baseline configuration used a learning rate of 0.001, a batch size of 256, 100 training epochs, three transformer layers, a patch size of 32, and two attention heads, achieving an overall accuracy of 58.30% and an average accuracy of 89.31% at SNR above 10 dB.

During tuning, only one parameter was varied at a time, while the others were fixed at the baseline values. This single-factor analysis is adopted to provide interpretable sensitivity trends for each design choice under controlled conditions, while we acknowledge that hyperparameters may be coupled. Joint hyperparameter optimization (e.g., Bayesian optimization) could be explored in future work to more efficiently search the coupled space; however, the focus here is to characterize the main effects and practical ranges of key parameters. This single-factor analysis is adopted to provide interpretable sensitivity trends for each design choice under controlled conditions, while we acknowledge that hyperparameters may be coupled. Joint hyperparameter optimization (e.g., Bayesian optimization) could be explored in future work to more efficiently search the coupled space; however, the focus here is to characterize the main effects and practical ranges of key parameters.

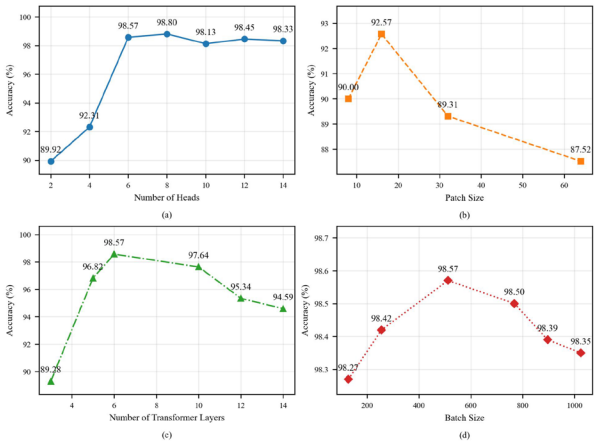

Fig. 8. Hyperparameter sensitivity analysis of AMC-Transformer. (a) Batch size vs. Accuracy; (b) Transformer layer depth vs. Accuracy; (c) Patch size vs. Accuracy; (d) Number of attention heads vs. accuracy

In addition to accuracy, varying the number of layers and attention heads directly changes model complexity (parameter count and compute), whereas batch size mainly affects optimization dynamics and patch size trades temporal resolution against sequence length. Therefore, the following results are discussed from both accuracy and complexity perspectives, which is particularly relevant for low-overhead 6G deployment scenarios.

1) Batch Size

Table 1 shows the relationship between batch size and classification accuracy. A baseline batch size of 256 yields 58.30 percent accuracy. Increasing the batch size gives only marginal gains, with peak performance observed around the 512–732 range; further increases lead to a slight degradation. This suggests the model benefits from more stable gradient updates, but overly large batches reduce helpful stochasticity.

2) Patch Size

Patch size strongly influences feature resolution. As shown in Fig. 8(c), reducing the patch size from 32 to 16 markedly improves accuracy to 92.57 percent, while a patch size of 8 produces 90.00 percent under baseline conditions. Larger patch sizes such as 64 degrade performance to 87.52 percent due to loss of fine-grained temporal features.

3) Transformer Layer Depth

Fig. 8(b) illustrates that accuracy improves steadily as the number of transformer layers increases, up to 10 layers where it reaches about 91.64 percent. Beyond this point, performance plateaus or slightly decreases, reflecting a trade-off between representational capacity and the risk of overfitting. Moreover, deeper stacks increase parameters and attention compute roughly linearly with depth, so the marginal accuracy gains beyond 6–10 layers should be weighed against the added complexity.

4) Multi-Head Attention

The number of attention heads has a pronounced effect. As seen in Fig. 8(d), accuracy rises quickly from 2 heads, where the baseline is 58.30 percent, to 6 heads, which achieves 92.57 percent. It then stabilizes around 8 to 14 heads near 90.8 percent and declines slightly thereafter, indicating that a moderate number of heads captures diverse signal dependencies without introducing redundancy. Since multi-head attention increases projection parameters and compute, the observed saturation beyond 8–14 heads indicate diminishing returns in accuracy relative to complexity.

Combining the best settings from each dimension — batch size 512, transformer layers 6, patch size 16, and attention heads 8 — yields an overall accuracy of 63.87 percent and an average accuracy of 98.80 percent at SNR greater than 10 dB (as shown in Table 2). This represents a notable improvement over the baseline, with an absolute overall gain of 5.57 percentage points and a gain of 9.49 points in the high-SNR regime.

TABLE II
EFFECT OF INDIVIDUAL PARAMETER OPTIMIZATION ON MODEL ACCURACY

Parameter	Baseline	Best Value	Accuracy (Overall)	Accuracy (SNR > 10 dB)
Batch Size	256	512	58.30 → 59.50	89.31 → 90.33
Layers	3	6	58.30 → 61.64	89.31 → 91.64
Patch Size	32	16	58.30 → 62.57	89.31 → 92.57
Heads	2	8	58.30 → 61.23	89.31 → 94.80
Combined	–	(6 layers, 8 heads, patch 16, batch 512)	63.87	98.80

To further illustrate the effect of hyperparameter optimization, Figure 9 presents the classification accuracy across the full SNR range for both the baseline and optimized configurations. While the baseline model saturates around 90% accuracy at high SNR levels, the optimized AMC-Transformer achieves up to 98.8% accuracy at SNR of at least 10 dB and shows consistent improvements in the mid-SNR range from 0 to 10 dB. This confirms that the performance gain is not restricted to very high SNR conditions, addressing concerns about robustness.

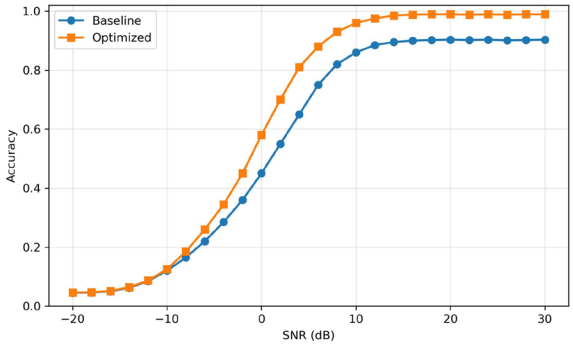

Fig. 9. Accuracy versus SNR for baseline and optimized AMC-Transformer.

To analyze the source of the performance gains, Fig. 10 and Fig. 11 contrasts class-wise confusion matrices before and after hyperparameter optimization across SNR ranges.

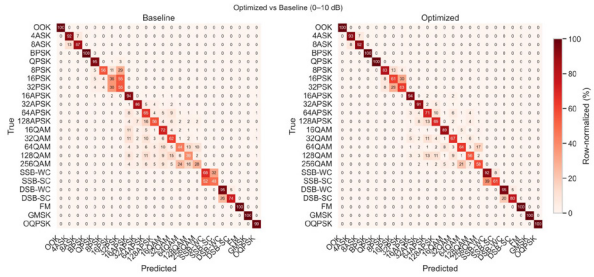

Fig. 10. Confusion matrices comparing baseline and optimized models at mid-to-low SNR range (5-15 dB).

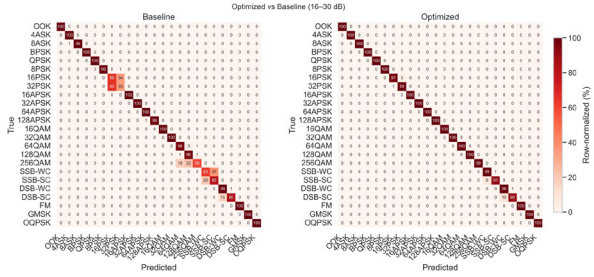

Fig. 11. Confusion matrices comparing baseline and optimized models at high SNR range (10-30 dB).

At higher SNRs (10–30 dB), the optimized model substantially reduces within-family confusion—most notably among high-order QAM constellations (e.g., 256-QAM), between adjacent PSK orders (16- vs. 32-PSK), and between AM subtypes (SSB vs. DSB). In the mid- to low-SNR regime (0–10 dB), the optimization primarily mitigates cross-family confusion, yielding sizable per-class recall gains. Overall, the confusion matrices indicate that the accuracy improvement arises from a systematic attenuation of characteristic misclassification patterns across the entire SNR spectrum, rather than from isolated gains at specific operating points.

Overall, the sensitivity results indicate that most of the achievable gains come from selecting an appropriate patch size and a moderate number of layers/heads, while very deep or heavily multi-headed configurations exhibit diminishing returns. Importantly, the optimized configuration improves accuracy across the full SNR range (Fig. 9) but does so with increased model complexity. This accuracy–complexity trade-off should be considered when targeting resource-constrained receivers and low-overhead 6G deployments.

E. Positional Encoding Strategy

To evaluate the impact of positional encoding methods on AMC-Transformer performance, we compared learnable positional embeddings against fixed sinusoidal encodings while keeping all other hyperparameters constant (6 layers, 8 heads, patch size 16, batch size 512).

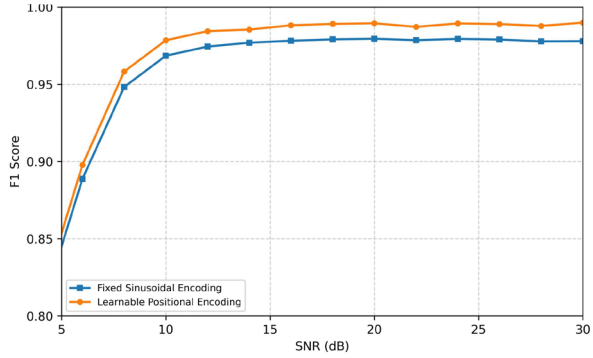

Fig. 12. Comparison of positional encoding strategies across SNR levels

Fig. 12 shows the F1 scores across the full SNR range for both encoding strategies. While both methods achieve comparable performance at low SNRs (below 0 dB), a clear divergence emerges in the mid-to-high SNR regime. Learnable positional encoding consistently outperforms fixed sinusoidal encoding above 10 dB SNR, maintaining an average F1 score of 98.5% compared to 97.3% for the fixed encoding—a relative improvement of 1.2 percentage points. The performance gap is most pronounced between 10-20 dB, suggesting that learnable embeddings better capture the position-dependent temporal patterns specific to modulated signals under favorable channel conditions.

This result indicates that allowing the model to learn task-specific positional representations provides measurable benefits for AMC, particularly when signal quality permits extraction of fine-grained temporal features. The learned embeddings likely adapt to the periodic structures and phase relationships inherent in different modulation schemes, which generic sinusoidal patterns cannot fully capture.

V. CONCLUSIONS

In this paper, we introduced AMC-Transformer, a transformer-based framework for automatic modulation classification that operates directly on raw I/Q time series. The model tokenizes I/Q sequences into fixed-length temporal patches, augments them with learnable positional embeddings, and applies multi-head self-attention to capture both short-range transitions and long-range dependencies in the waveform.

On RadioML2018.01A, our best configuration achieves 98.8% accuracy for SNR of 10 dB or higher and 63.9% on average across all SNRs, showing improved accuracy compared to reimplemented CNN/ResNet and MCformer baselines under the same data splits and training protocol. On RadioML2016.10a, AMC-Transformer maintains competitive accuracy across SNR levels and reaches 93.5% in the high-SNR regime, demonstrating robustness beyond a single dataset. Ablation studies indicate that model capacity and tokenization drive the accuracy–efficiency trade-off: patch size 16, 8 attention heads, and about six encoder layers offer a favorable balance across SNR conditions. These gains, however, come at the cost of increased model complexity, which should be carefully considered for low-overhead and resource-constrained 6G receivers.

Despite these gains, transformer attention scales quadratically and benefits from substantial data. Future work will explore efficient attention (for example, linear or clustered variants), hybrid Conv–Attention designs that inject local inductive bias, and RF-aware augmentation and self-supervision to improve robustness to channel non-idealities while reducing the need for labeled data.

REFERENCES

[1] A. Aboulfotouh, A. Eshaghbeigi and H. Abou-Zeid, "Building 6G Radio Foundation Models with Transformer Architectures," *ICC 2025 – IEEE International Conference on Communications*, Montreal, QC, Canada, 2025, pp. 1888–1893, DOI: 10.1109/ICC52391.2025.11161954.

[2] R. Ding, F. Zhou, H. Zhang, Q. Wu, and Z. Han, "Data- and knowledge dual-driven automatic modulation classification for 6G wireless communications," *IEEE Transactions on Wireless Communications*, 2024 (early access), DOI: 10.1109/TWC.2023.3316197.

[3] 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, Feb. 2018, DOI: 10.1109/JSTSP.2018.2797022.

[4] W. Wei and J. M. Mendel, "Maximum-likelihood classification for digital amplitude-phase modulations," *IEEE Transactions on Communications*, vol. 48, no. 2, pp. 189–193, Feb. 2000, DOI: 10.1109/26.823550.

[5] F. Hameed, O. A. Dobre, and D. C. Popescu, "On the likelihood-based approach to modulation classification," *IEEE Transactions on Wireless Communications*, vol. 8, no. 12, pp. 5884–5892, Dec. 2009, DOI: 10.1109/TWC.2009.12.080883.

[6] W. Gardner, "Spectral Correlation of Modulated Signals: Part I – Analog Modulation," in *IEEE Transactions on Communications*, vol. 35, no. 6, pp. 584–594, June 1987, DOI: 10.1109/TCOM.1987.1096820.

[7] W. A. Gardner, "Spectral correlation of modulated signals: Part I—Analog modulation," *IEEE Transactions on Communications*, vol. 35, no. 6, pp. 584–594, Jun. 1987, DOI: 10.1109/TCOM.1987.1096820.

[8] C. A. Harper, M. A. Thornton, and E. C. Larson, "Automatic Modulation Classification with Deep Neural Networks," *Electronics*, vol. 12, no. 18, 3962, Sep. 2023, DOI: 10.3390/electronics12183962.

[9] B. Jdid, K. Hassan, I. Dayoub, W. H. Lim, and M. Mokayef, "Machine learning based automatic modulation recognition for wireless communications: A comprehensive survey," *IEEE Access*, vol. 9, pp. 57 851–57 873, 2021, DOI: 10.1109/ACCESS.2021.3071801.

[10] S. Peng, H. Jiang, H. Wang, H. Alwageed, Y. Zhou, M. M. Sebdani, and Y.-D. Yao, "Modulation classification based on signal constellation diagrams and deep learning," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 30, no. 3, pp. 718–727, Mar. 2019, DOI: 10.1109/TNNLS.2018.2850703.

[11] T. Huynh-The, V. S. Doan, C.-H. Hua, Q.-V. Pham, and D.-S. Kim, "Chain-Net: Learning deep model for modulation classification under synthetic channel impairment," in *Proc. IEEE GLOBECOM*, 2020, pp. 1–6, DOI: 10.1109/globecom42002.2020.9322394.

[12] N. P. Shankar, D. Sadhukhan, N. Nayak, T. Tholeti and S. Kalyani, "Binarized ResNet: Enabling Robust Automatic Modulation Classification at the Resource-Constrained Edge," in *IEEE Transactions on Cognitive Communications and Networking*, vol. 10, no. 5, pp. 1913–1927, Oct. 2024, DOI: 10.1109/TCCN.2024.3391325.

[13] M. L. Xue, M. Huang, J. J. Yang, and J. D. Wu, "MLResNet: An efficient method for automatic modulation classification based on residual neural network," in *Proc. 2nd Int. Symp. on Computer Engineering and Intelligent Communications (ISCEIC)*, 2021, pp. 122–126, DOI: 10.1109/ISCEIC53685.2021.00032.

[14] A. Vaswani et al., "Attention is all you need," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2017, pp. 5998–6008, *arXiv:1706.03762*.

[15] Q. Wen, T. Zhou, C. Zhang, et al., "Transformers in time series: A survey," in *Proc. Int. Joint Conf. on Artificial Intelligence (IJCAI)*, 2023, DOI: 10.24963/ijcai.2023/759.

[16] S. Hamidi-Rad and S. Jain, "MCformer: A Transformer-based deep neural network for automatic modulation classification," in *Proc. IEEE GLOBECOM*, 2021, pp. 1–6, DOI: 10.1109/globecom46510.2021.9685815.

[17] W. Ma, Z. Cai, and C. Wang, "A Transformer- and convolution-based learning framework for automatic modulation classification," *IEEE Communications Letters*, vol. 28, no. 6, pp. 1392–1396, Jun. 2024, DOI: 10.1109/lcomm.2024.3380623.

[18] D. Wang, M. Lin, X. Zhang, Y. Huang, and Y. Zhu, "Automatic modulation classification based on CNN–Transformer–GNN (CTGNet)," *Sensors*, vol. 23, no. 16, 7281, Aug. 2023, DOI: 10.3390/s23167281.

[19] Radioml2018.01A. [Online]. Available: https://www.deepsig.ai/datasets

[20] A. Abbas, V. Pano, G. Mainland and K. Dandekar, "Radio Modulation Classification Using Deep Residual Neural Networks," *MILCOM 2022 – 2022 IEEE Military Communications Conference (MILCOM)*, Rockville, MD, USA, 2022, pp. 311–317, DOI: 10.1109/MILCOM55135.2022.10017640.

[21] C. Xiao, S. Yang, and Z. Feng, "Complex-Valued Depth-wise Separable CNN for AMC," *IEEE Transactions on Instrumentation and Measurement*, vol. 72, 2023, DOI: 10.1109/TIM.2023.3298657.

[22] T. Huynh-The, C.-H. Hua, Q.-V. Pham, and D.-S. Kim, "MCNet: An Efficient CNN Architecture for Robust AMC," *IEEE Communications Letters*, vol. 24, no. 4, pp. 811–815, Apr. 2020, DOI: 10.1109/LCOMM.2020.2968030.

[23] K. Liao, Y . Zhao, J. Gu, Y . Zhang, and Y . Zhong, "Sequential Convolutional Recurrent Neural Networks for Fast AMC," *IEEE Access*, vol. 9, pp. 27 182–27 188, 2021, DOI: 10.1109/ACCESS.2021.3053427.

[24] S. Ramjee *et al*., "Fast Deep Learning for Automatic Modulation Classification," *ArXiv, abs/1901.05850*.

[25] T. Wang, Z. Zhang, X. Wang, and Y. Li, "A Survey of Applications of Deep Learning in Radio Signal Modulation Recognition," *Applied Sciences*, vol. 12, no. 23, 12052, Nov. 2022, **DOI**: 10.3390/app122312052.

[26] W. Zhang *et al*., "CTRNet: An Automatic Modulation Recognition Based on CNN with Transformer," *Electronics*, vol. 13, no. 6, 1047, Mar. 2024, **DOI**: 10.3390/electronics13173408.

[27] Q. Zheng *et al*., "A Real-Time Constellation Image Classification Method of AMC Based on MobileViT," *Scientific Reports*, vol. 13, 18656, Oct. 2023, **DOI**: 10.1007/s11571-023-10015-7.

[28] A. Faysal, M. Rostami, R. G. Roshan, H. Wang and N. Muralidhar, "NMformer: A Transformer for Noisy Modulation Classification in Wireless Communication," *2024 33rd Wireless and Optical Communications Conference (WOCC)*, Hsinchu, Taiwan, 2024, pp. 103–108, **DOI**: 10.1109/WOCC61718.2024.10786062.

[29] J. Jang, J. Pyo, Y.-I. Yoon and J. Choi, "Meta-Transformer: A Meta-Learning Framework for Scalable Automatic Modulation Classification," in *IEEE Access*, vol. 12, pp. 9267–9276, 2024, **DOI**: 10.1109/ACCESS.2024.3352634.

[30] X. Hao, Z. Feng, S. Yang, M. Wang and L. Jiao, "Automatic Modulation Classification via Meta-Learning," in *IEEE Internet of Things Journal,* vol. 10, no. 14, pp. 12 276–12 292, 15 July15, 2023, **DOI**: 10.1109/JIOT.2023.3247162.

[31] W. Kong, X. Jiao, Y. Xu, B. Zhang and Q. Yang, "A Transformer-Based Contrastive Semi-Supervised Learning Framework for Automatic Modulation Recognition," in *IEEE Transactions on Cognitive Communications and Networking*, vol. 9, no. 4, pp. 950–962, Aug. 2023, **DOI**: 10.1109/TCCN.2023.3264908.

[32] D. Liu, P. Wang, T. Wang, and T. Abdelzaher, "Self- Contrastive Learning based Semi-Supervised Radio Modulation Classification," in *Proc. IEEE MILCOM*, 2021, pp. 777–782, **DOI**: 10.1109/MILCOM52596.2021.9652914.

[33] K. Davaslioglu, S. Boztaş, M. C. Ertem, Y. E. Sagduyu and E. Ayanoglu, "Self-Supervised RF Signal Representation Learning for NextG Signal Classification With Deep Learning," in *IEEE Wireless Communications Letters*, vol. 12, no. 1, pp. 65–69, Jan. 2023, **DOI**: 10.1109/LWC.2022.3217292.

[34] S.-H. Kim, J.-W. Kim, W.-P. Nwadiugwu and D.-S. Kim, "Deep Learning-Based Robust Automatic Modulation Classification for Cognitive Radio Networks," in *IEEE Access*, vol. 9, pp. 92 386–92 393, 2021, **DOI**: 10.1109/ACCESS.2021.3091421.

Yuewen Xu was born in 2001 in Zhuanglang County, Pingliang City, Gansu Province, China. He enrolled at Xi'an Jiaotong-Liverpool University in 2019 and obtained a bachelor's degree in Electrical Engineering and Automation in 2023. Currently, he is pursuing a master's degree in Communication Network and Signal Processing at the University of Bristol.

From June 2022 to December 2022, he participated in the development of an Artificial Intelligence-based Surgical Instrument Inventory System, where he was responsible for designing the user interface. Between September 2022 and May 2023, he completed a thesis on the design of electric propulsion systems for unmanned aerial vehicles. His research interests include Artificial Intelligence applications, Communication Networks, and Signal Processing. In his free time, he enjoys exploring technological advancements and cultural heritage, striving to apply his expertise to real-world challenges.