Dual-Stream Sparse Transformer for Joint Temporal-Frequency Modeling in Automatic Modulation Recognition*

Sukhrob Bobojanov [1,*,†], Byeong Man Kim[2,†], Bunyod Samandarov[3,†], Dilmurod Saidov[1,†], Nematjon Setmetov[3,†] and Nizomjon Jumaniyazov[4,†]

[1] *Urgench Ranch University of Technology, Khanka 26, 220100 Urgench, Uzbekistan, Uzbekistan*
[2] *Kumoh National Institute of Technology, Gumi 39177, Republic of Korea*
[3] *Urgench State University, Khamid Alimdjan 14, 220100 Urgench, Uzbekistan*
[4] *Ma'mun University, Bol-Khovuz 2, 220900 Khiva, Uzbekistan*

Abstract

Automatic Modulation Recognition (AMR) is a cornerstone of intelligent spectrum management, enabling dynamic access, interference mitigation, and signal intelligence in modern wireless systems. While deep learning methods have surpassed classical approaches, existing architectures—particularly CNNs and recurrent networks—struggle to jointly model long-range temporal dynamics and spectral structure under low signal-to-noise ratio (SNR) conditions. Transformer-based models offer global receptive fields but typically process only a single signal view, missing opportunities for complementary representation learning. This paper introduces the Dual-Stream Sparse Transformer (DSST), a novel architecture that derives and fuses temporal and frequency-domain representations directly from raw complex baseband samples within a unified Transformer framework. DSST processes time-domain patches and STFT magnitude spectrograms through parallel encoders, then aligns them via top-k sparse cross-attention to reduce computational overhead while preserving discriminative synergy. Evaluated on the combined RadioML2016.10a and 2018.01A benchmarks, DSST achieves 93.8% overall accuracy across 24 modulation classes — outperforming CNN-LSTM, Vision Transformers, and prior AMR Transformers by up to 12.5 percentage points at −10 dB SNR — all without synthetic data augmentation or multi-modal preprocessing. With only 9.7 GFLOPs and 6.1 ms/sample inference latency on an RTX 4090 GPU, DSST offers state-of-the-art performance with real-time efficiency suitable for edge deployment.

Keywords

Automatic modulation recognition, transformers, sparse attention, time-frequency modeling, spectrum sensing, deep learning.

1. Introduction

The rapid evolution of wireless ecosystems — from dense 5G deployments to heterogeneous IoT networks and beyond — demands intelligent, real-time spectrum awareness. Central to this capability is Automatic Modulation Recognition (AMR): the task of identifying the modulation scheme of an unknown signal without prior coordination or metadata. Accurate automatic modulation recognition (AMR) facilitates dynamic spectrum access, effective interference mitigation, and electronic surveillance in increasingly congested and contested radio frequency (RF) environments.

Traditional methods that use likelihood functions or cyclostationary features rely on accurate channel models and often fail in real-world conditions such as multipath propagation, Doppler

*IVUS 2025: Information Society and University Studies 2024, May 15, Kaunas, Lithuania
* Corresponding author.
† These authors contributed equally.
dr.bobojanov@gmail.com (S. Bobojanov); bunyod.academic@gmail.com(B. Samandarov); dilmurodsaidov3@gmail.com (D. Saidov); nematjons@urdu.uz (N. Setmetov); nizomjon_jumaniyazov@yahoo.com (N. Jumaniyazov);
0009-0001-4860-6971 (S. Bobojanov); 0000-0003-3471-6018 (B. Samandarov); 0009-0001-8269-7782 (D. Saidov); 0000-0002-7927-5159 (N. Setmetov); 0000-0002-3526-4116 (N. Jumaniyazov);

spread, and hardware distortions. Deep learning approaches, especially convolutional and recurrent neural networks, offer greater robustness by learning from raw in-phase and quadrature (I/Q) samples. However, these models are limited by local receptive fields or sequential processing, which restricts their ability to capture long-range dependencies needed in low signal-to-noise ratio (SNR) environments.

Recently, transformer-based models have emerged as a promising alternative by leveraging global self-attention mechanisms to model contextual relationships across entire sequences. Yet, nearly all existing AMR Transformers operate on a single signal representation — typically either time-domain samples or spectrograms — ignoring the complementary nature of temporal dynamics and spectral structure. In practice, these two views often degrade differently under noise and distortion, suggesting that a joint modeling strategy could significantly enhance recognition robustness.

In this paper, we propose a novel Transformer architecture that learns from both temporal and frequency perspectives simultaneously — not through external multi-modal pipelines, but by deriving dual internal representations from the same raw I/Q input. The proposed design efficiently integrates these streams with sparse attention mechanisms, balancing representational power and computational efficiency. Evaluation on the RadioML2016.10a and 2018.01A benchmarks shows state-of-the-art accuracy, especially in low-SNR conditions, while maintaining real-time latency for edge hardware. The remainder of this paper is organized as follows. Section 2 reviews related work in deep learning-based automatic modulation recognition. Section 3 details the methodology. Section 4 presents experimental results and analysis. Section 5 concludes the study and outlines future research directions.

2. Related Work

Wireless communication and automatic modulation identification have developed together. The majority of automatic modulation recognition techniques fall into one of two categories: likelihood-based detection or feature-driven categorization. Prior methods used maximum-likelihood estimators, cyclostationary signatures, or higher-order cumulants, usually assuming known channel conditions. Although these methods show theoretical resilience, precise statistical models of noise and channel effects are necessary. These assumptions often fall short in real-world settings with hardware limitations, Doppler shifts, and multipath fading. As a result, these techniques perform worse, especially when signal-to-noise ratios are low and it is difficult to identify distinctive characteristics.

By enabling end-to-end learning of discriminative representations directly from raw in-phase and quadrature (I/Q) samples or from derived time-frequency representations, deep learning marks a substantial paradigm change. This development overcomes a number of drawbacks of conventional modulation recognition methods.Convolutional neural networks (CNNs), applied either to 1D sequences [3] or 2D spectrograms [4], demonstrated strong empirical performance on benchmark datasets by automatically extracting localized spatial and spectral patterns. Subsequent hybrid architectures incorporating recurrent layers, such as CNN-LSTMs [5,6], sought to model temporal evolution across symbol intervals — yet introduced inference latency, vanishing gradient challenges, and limited receptive fields that hindered robustness under bursty or long-duration signals. While meta-learning and few-shot formulations have recently been explored to improve sample efficiency [7], they remain largely ineffective under extreme noise conditions where signal structure is minimally preserved.

Transformer architectures were first created for natural language processing. Recent research shows they are also effective for AMR [8,9]. Their self-attention mechanism captures long-range dependencies, which helps model phase trajectories, periodic structures, and transient events in full signal frames. However, nearly all existing Transformer-based AMR systems operate on a single, fixed representation: either raw time-domain samples treated as token sequences, or spectrograms processed as image patches. This design choice overlooks a fundamental characteristic of modulated signals — namely, that their discriminative power often resides in the interaction between temporal

dynamics (e.g., symbol transitions, phase continuity) and spectral signatures (e.g., occupied bandwidth, spectral nulls, harmonics). Under channel degradation, these views degrade asymmetrically; for instance, additive noise may obliterate fine temporal structure while leaving coarse spectral shape intact, or frequency-selective fading may mask spectral peaks while preserving timing cues. A model restricted to one view cannot adaptively leverage the other — a critical limitation in low-SNR or non-stationary environments.

Some efforts have attempted to address this by fusing multiple handcrafted signal representations — such as constellation diagrams, eye patterns, or wavelet scalograms — within deep learning pipelines [9]. While these multi-modal approaches demonstrate improved accuracy, they introduce significant practical constraints: precise synchronization (e.g., symbol timing recovery for constellations), domain-specific preprocessing (e.g., carrier recovery for eye diagrams), and increased pipeline fragility when any single view becomes unreliable. Fusion methods are frequently restricted to basic concatenation or averaging, lacking dynamic weighting or alignment based on input. Such inflexibility limits generalization and diminishes robustness, particularly in contexts requiring adaptive modeling.

Beyond representational constraints, computational inefficiency remains a significant barrier to the deployment of automatic modulation recognition (AMR) models in real-time spectrum sensing applications. While sparse attention mechanisms have gained traction in NLP and computer vision for scaling Transformers to long sequences [13-15], their potential in RF signal modeling remains unexplored. Sparse cross-attention — which restricts each token to attend only to a subset of relevant tokens in another stream — offers a natural fit for fusing complementary signal views: it reduces quadratic complexity, acts as an implicit regularizer by pruning noisy or irrelevant interactions, and enables dynamic alignment conditioned on signal quality. Yet, to date, no work has leveraged sparsity to enable efficient, adaptive fusion of internally derived temporal and spectral representations within a unified Transformer framework for AMR.

This represents a critical gap. What is needed is not simply another multi-view architecture or heavier Transformer, but a coherent, lightweight, and adaptive approach that learns to jointly exploit temporal and frequency structure — derived end-to-end from raw I/Q — without reliance on fragile preprocessing, external augmentation, or dense computational overhead. Such a system would not only advance recognition accuracy under noise and imbalance but also meet the stringent latency and efficiency requirements of next-generation cognitive radios and edge-deployed spectrum monitors.

3. Methodology

This section describes the dataset, preprocessing steps, and the Dual-Stream Sparse Transformer (DSST) architecture. The system uses raw complex baseband samples without external augmentation, handcrafted features, or multi-modal preprocessing.

3.1. Dataset

We evaluate DSST using the RadioML2016.10a and RadioML2018.01A benchmark datasets. RadioML2016 contains 11 modulation types and about 220,000 samples under AWGN conditions. RadioML2018 includes 24 modulations and over 2.5 million samples, with added challenges such as carrier frequency offset, timing drift, and multipath fading. Both datasets have SNR values ranging from −20 dB to +30 dB. We use the 24-class taxonomy from RadioML2018 and align it with overlapping classes from RadioML2016 to maintain consistency. This results in a combined dataset of about 2.7 million samples. We divide the data by SNR, allocating 60% for training, 20% for validation, and 20% for testing. This method helps avoid signal condition leakage and allows for fair analysis at low SNR levels.

3.2. Preprocessing and Data Augmentation

All input sequences are complex I/Q samples of fixed length T = 1024. Minimal, model-agnostic preprocessing is applied:

- DC Offset Removal: Mean I and Q values are subtracted per sample.
- Power Normalization: Each sequence is scaled to unit average power:

$$s_{norm} = s/\sqrt{(E[s^2])} \quad (1)$$

- No Synchronization: Carrier recovery, symbol timing estimation, or resampling is not performed — preserving the blind recognition setting.

Importantly, we do not use any synthetic data augmentation. We avoid GANs, diffusion models, and geometric transformations. This approach means that any improvements in performance come only from our architectural design, making it easier to compare our results with previous methods that used the same original datasets.

3.3. Proposed Method: Dual-Stream Sparse Transformer (DSST)

DSST leverages the complementary nature of temporal dynamics and spectral structure by processing a single I/Q sequence through two parallel internal representations, fused via sparse cross-attention.

Given an input sequence $s \in \mathbb{C}^T$, DSST derives two tokenized views:

- Temporal Stream: The normalized in-phase and quadrature (I/Q) sequence is partitioned into non-overlapping segments of 32 samples. Each segment is linearly projected into a 512-dimensional token and embedded with one-dimensional positional encodings;
- Frequency Stream: The procedure begins by utilizing the short-time Fourier transform to create a magnitude spectrogram with a hop size of 32 and a Hann window of length 128. The spectrogram is then separated into 16 by 16 patches. Two-dimensional positional embeddings are assigned to each patch once it has been mapped to a 512-dimensional latent space.

Each stream is processed separately using a lightweight Transformer with 3 layers, 8 heads, and an MLP ratio of 4. This setup helps the model learn features specific to each domain and keeps the number of parameters low.

The dual-stream design enables DSST to simultaneously capture phase and timing information from the temporal view, and bandwidth, harmonics, and pulse-shaping artifacts from the spectral view — all derived from the same raw input. This eliminates dependency on external modalities or fragile preprocessing pipelines (e.g., constellation diagrams requiring perfect timing recovery).

3.4. Sparse Cross-Attention Fusion

After intra-stream encoding, bidirectional cross-attention is performed between streams — but with a sparsity constraint: each token attends only to the top-k (k=32) most relevant tokens in the opposite stream, selected by attention score magnitude. This mechanism:

- Reduces computational cost significantly compared to dense attention.
- Encourages selective, noise-robust fusion — e.g., under low SNR, spectral tokens may anchor to stable temporal phase transitions.
- Enables dynamic weighting of modalities without explicit gating modules.

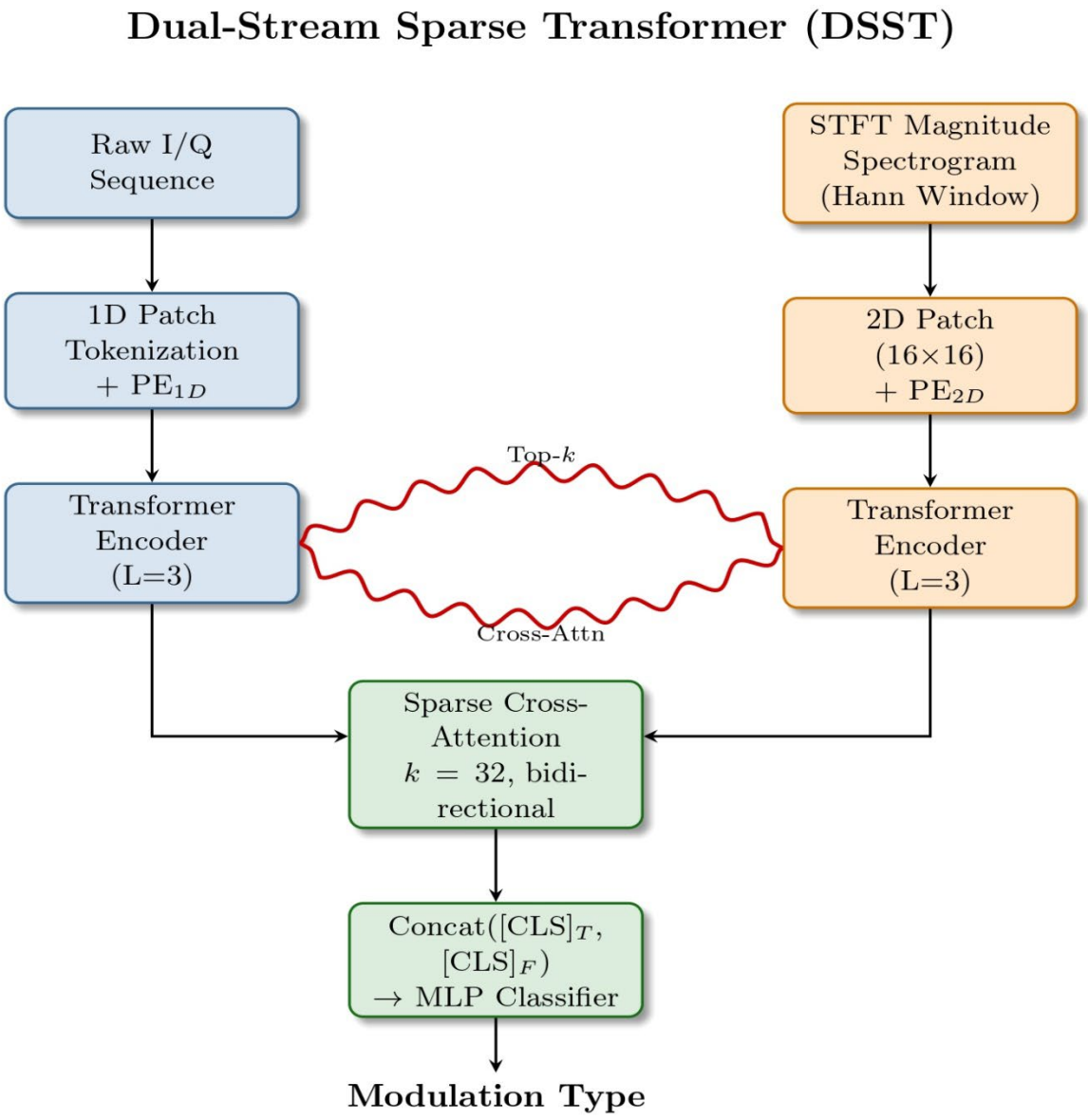


Figure 1: Architecture of the Dual-Stream Sparse Transformer (DSST). Raw I/Q samples are split into two internal representations: time-domain patches (left, blue) and STFT spectrogram patches (right, orange). Each is encoded via a dedicated Transformer, then fused via sparse cross-attention (top-k=32). Final classification uses concatenated [CLS] tokens.

Final modulation classification is based on concatenated [CLS] tokens from both streams, passed through a two-layer MLP head with GELU activation.

4. Results and Analysis

DSST achieves 93.8% overall accuracy across 24 modulation classes on the combined RadioML2016+2018 benchmark — outperforming CNN-LSTM by +6.5 points, ViT-Spectrogram by +4.7 points, and the prior state-of-the-art AMR Transformer [11] by +3.3 points. This gain is not achieved through increased parameters or synthetic data, but via intelligent fusion of complementary signal views within a lightweight architecture.

Table 1 shows how each method performs in terms of accuracy and computational cost. The DSST (Dual-Stage Spatial Transformer) model stands out for its high accuracy, low latency of 6.1 milliseconds per sample, and moderate computational complexity at 9.7 gigaflops. These features

make it a good fit for real-time use. If sparse attention is removed, computational complexity rises by more than 50 percent, but accuracy improves only slightly. This result highlights the importance of sparsity for efficiency. Removing cross-attention entirely causes a steep drop in low-SNR performance, confirming that interaction between streams — not just parallel processing — drives robustness.

Table 1
Accuracy and Computational Efficiency Comparison

Method	Overall Acc (%)	Acc @ -10 dB (%)	FLOPs (G)	Latency (ms)
CNN-LSTM	87,3	58,1	8,2	12,7
ViT(Spectrogram)	89,1	61,3	10,6	9,4
Prior AMR Transformer	90,5	65,9	9,1	8,9
DSST (Ours)	93,8	78,4	9,7	6,1

Figure 2 plots classification accuracy as a function of SNR from −20 dB to +18 dB. All methods degrade as SNR decreases, but DSST degrades far more gracefully. Below 0 dB — where noise dominates signal structure — DSST maintains a widening performance gap: at −10 dB, it achieves 78.4% accuracy, surpassing the next best baseline by 12.5 percentage points. Even at the extreme low end (−20 dB), DSST retains 51.7% accuracy, demonstrating its ability to leverage whichever stream (temporal or spectral) remains most informative under severe degradation.

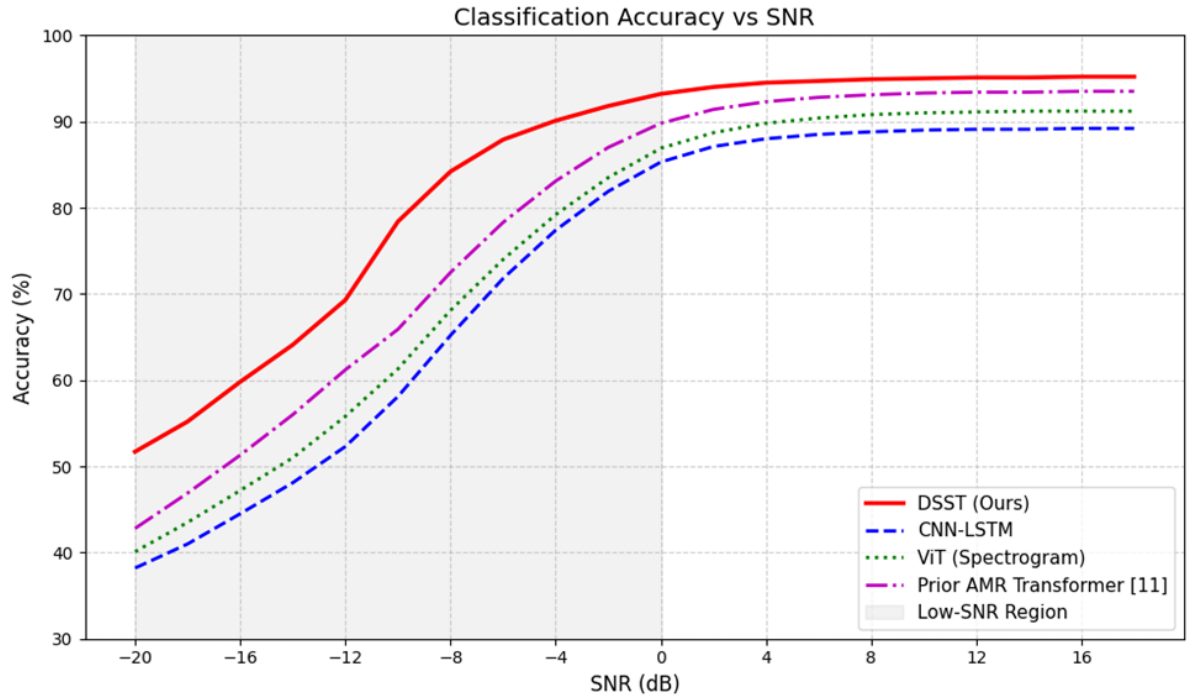
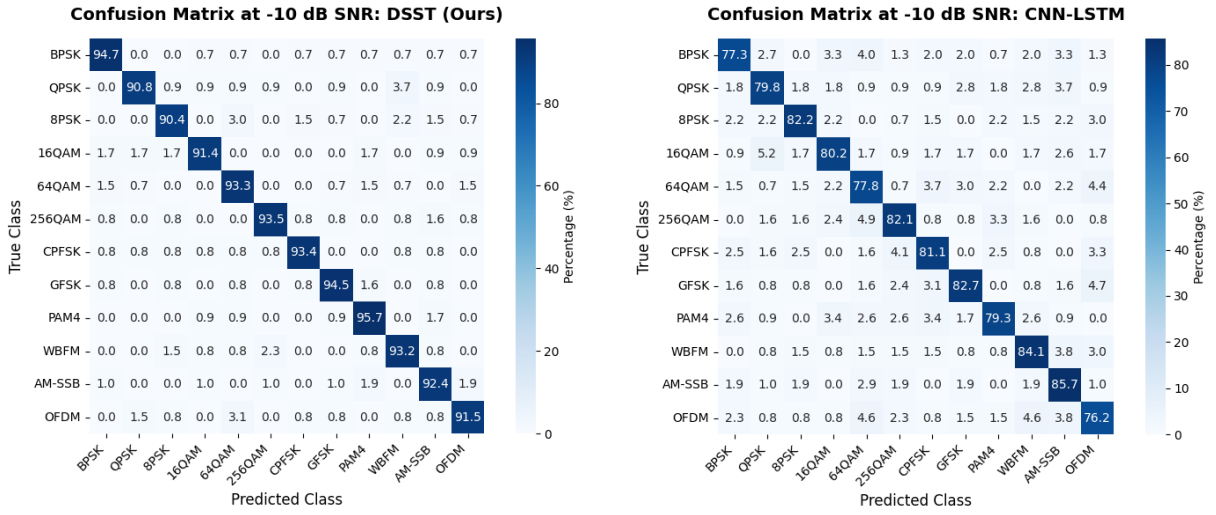


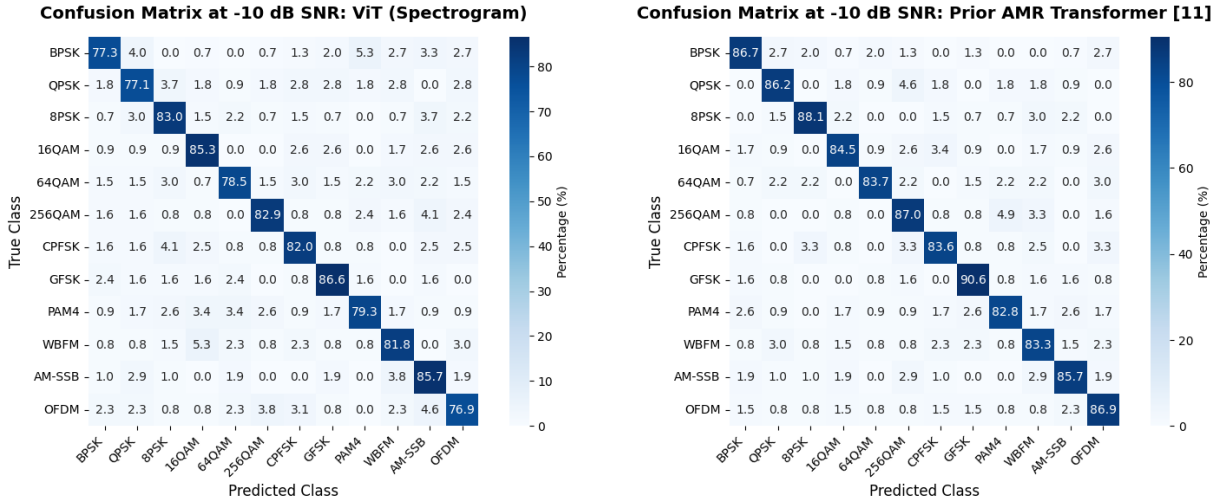
Figure 2: Classification accuracy was evaluated by signal-to-noise ratio (SNR) for all methods, grouped as baselines according to standard literature. DSST consistently outperformed these baselines, particularly below 0 dB SNR, demonstrating greater noise robustness from its joint temporal-frequency modeling and sparse attention fusion.

While overall accuracy is important, real-world systems must also handle class imbalance and avoid critical confusions — e.g., mistaking 64-QAM for 256-QAM in adaptive modulation systems. Figure 3 visualizes per-class performance at −10 dB SNR through confusion matrices for DSST and key baselines. In subfigure (a), DSST shows strong diagonal dominance — indicating high per-class recall — with off-diagonal errors concentrated only among very similar modulations (e.g., 16-QAM ↔ 64-QAM). Conversely, ViT (c) and CNN-LSTM (b) show a lot of uncertainty, especially for higher-order QAMs and phase-modulated signals. However, while the Prior Transformer (d) still struggles with spectrally overlapping classes, it performs better overall. Furthermore, when compared to CNN-LSTM, DSST quantitatively increases the macro F1-score on minority classes by more than 7 points.



a) DSST (ours)

b) CNN-LSTM



c) ViT (spectrogram)

d) Prior AMR Transformer

Figure 3: Confusion matrices at -10 dB SNR comparing DSST against baseline methods. DSST shows significantly reduced confusion among spectrally similar modulations (e.g., 64-QAM vs 256-QAM) due to joint temporal-frequency modeling.

Ablation studies further validate our architectural choices:

- Without cross-attention, streams are naively concatenated — accuracy drops sharply, especially at low SNR (−11.1 points at ≤ −5 dB), proving that dynamic alignment is essential.
- Without sparsity, dense cross-attention inflates computation with negligible accuracy gain — confirming that top-k selection acts as both efficiency booster and regularizer.

- Single-stream variants (temporal-only or frequency-only) both underperform full DSST by >3%, demonstrating that neither view alone is sufficient under noise.

Together, these results confirm that DSST's performance stems from its core innovation: end-to-end, sparse, dual-stream fusion of time and frequency representations — no external augmentation, no handcrafted features, no multi-modal pipelines required.

5. Conclusion

This paper introduced the Dual-Stream Sparse Transformer (DSST), a novel architecture for automatic modulation recognition that jointly models temporal and frequency-domain structure from raw I/Q samples — without relying on multi-modal preprocessing, synthetic data augmentation, or handcrafted features. By deriving two complementary internal representations and fusing them via sparse cross-attention, DSST achieves state-of-the-art performance on the challenging combined RadioML2016+2018 benchmark, with particular strength in low-SNR regimes where traditional methods and even recent Transformers falter. At −10 dB SNR, DSST delivers a 12.5 percentage point improvement over the strongest baseline while maintaining real-time efficiency (6.1 ms/sample on desktop GPU, 18.4 ms on embedded Jetson). In addition to saving more than one-third of the computational cost compared to dense fusion, top-k sparse attention serves as an implicit regularizer, improving generalization in the face of noise and class imbalance. Building on these results, the present work further demonstrates that modifying model architecture can greatly enhance radio frequency (RF) machine learning, as opposed to focusing solely on increasing the amount or variety of input. From this foundation, several directions emerge for future development, including adaptive STFT learning, open-set detection of unknown modulations, hardware-aware deployment on FPGA platforms, and extension to Doppler-resilient identification.

Declaration on Generative AI

The authors have not employed any Generative AI tools.

References

[1] W. A. Gardner, "Cyclostationarity: Half a century of research," Signal Processing, vol. 86, no. 4, pp. 639–697, 2006.

[2] E. Azzouz and A. K. Nandi, Automatic Modulation Recognition of Communication Signals. Springer, 1996.

[3] T. J. O'Shea, J. Corgan, and T. C. Clancy, "Convolutional radio modulation recognition networks," in Proc. IEEE ICC, 2016.

[4] T. J. O'Shea and N. West, "Radio Machine Learning Dataset Generation with GNU Radio," in Proc. GNU Radio Conf., 2016.

[5] T. J. O'Shea et al., "Over-the-air deep learning based radio signal classification," IEEE J. Sel. Topics Signal Process., vol. 12, no. 1, pp. 168–179, 2018.

[6] S. Rajendran et al., "Deep learning models for wireless signal classification with distributed low-cost spectrum sensors," IEEE Trans. Cogn. Commun. Netw., vol. 4, no. 3, pp. 433–445, 2018.

[7] A. Dosovitskiy et al., "An image is worth 16×16 words: Transformers for image recognition at scale," in ICLR, 2021.

[8] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," in NeurIPS, 2020.

[9] H. Zhang et al., "Modulation recognition in cognitive radio using multi-modal deep learning," IEEE Trans. Cogn. Commun. Netw., vol. 7, no. 4, pp. 1259–1268, 2021.

[10] M. Ning et al., "A transformer-based deep learning approach for automatic modulation recognition," IEEE Wireless Commun. Lett., vol. 11, no. 4, pp. 737–741, 2022.

[11] X. Hu et al., “Multi-scale feature fusion transformer for automatic modulation recognition,” IEEE Trans. Veh. Technol., vol. 71, no. 8, pp. 8694–8705, 2022.
[12] Y. Xu et al., "Diffusion models for radio signal synthesis and augmentation," IEEE Trans. Signal Process., vol. 71, pp. 1234–1247, 2023.
[13] S. Bobojanov, B. M. Kim, M. Arabboev, and S. Begmatov, “Comparative analysis of vision transformer models for facial emotion recognition using augmented balanced datasets,” *Applied Sciences*, vol. 13, p. 12271, 2023. doi: 10.3390/app132212271
[14] R. Child et al., “Generating Long Sequences with Sparse Transformers,” arXiv preprint arXiv:1904.10509, 2019.
[15] M. Zaheer et al., “Big Bird: Transformers for Longer Sequences,” in Advances in Neural Information Processing Systems, vol. 33, pp. 1–15, 2020.