Proceedings of CONF-MPCS 2025 Symposium: Mastering Optimization: Strategies for Maximum Efficiency
DOI: 10.54254/2753-8818/101/2025.CH22300

Multi-Frame Dual-Stream 2DCNN-LSTM Model for Automatic Modulation Recognition

Zihan Zhou

Xi'an University, Xi'an, China
22009100769@stu.xidian.edu.cn

Abstract: This paper proposes a multi-frame dual-stream 2DCNN-LSTM model (MF-DS-2DCNN-LSTM) for automatic modulation recognition. The model discretizes long sequences into two-dimensional frame structures and uses 2D CNN and LSTM together to model the spatiotemporal features of multi-channel IQ/AP signals. By employing a frame-based strategy, the original signal is reshaped into" small images," with the 2D CNN extracting intra-frame spatial structures and inter-channel interaction features, while the LSTM captures the temporal evolution between frames. This approach integrates hierarchical modeling concepts from image processing and video analysis, and utilizes the Crested Porcupine Optimizer for hyperparameter tuning. Simulations show that, when recognizing nine modulation types, the model significantly outperforms methods such as CLDNN, achieving an average accuracy of 91.4% under high-SNR conditions (SNR above 2 dB). Moreover, the model maintains an accuracy of over 90% in small-sample training scenarios for SNRs above 4 dB. After optimization with the Crested Porcupine Optimizer, the model's performance improved by 2.2%, and a 20.7% reduction in parameters was achieved.

Keywords: Automatic Modulation Recognition, CNN, LSTM, Crested Porcupine Optimizer

1. Introduction

Automatic modulation recognition (AMR) is a core technology in cognitive radio, enabling modulation recognition through signal time-frequency analysis. Its applications span military electronic warfare and civilian sectors. However, challenges such as low signal-to-noise ratios (SNR), multipath fading, and novel modulation techniques necessitate overcoming the limitations of conventional AMR methods. Traditional modulation recognition techniques can be categorized into likelihood-based (LB) and feature-based (FB) methods. LB methods, reliant on likelihood function modeling, demand detailed prior knowledge and precise model assumptions. They suffer from high computational complexity, threshold dependence, and susceptibility to parameter variations, compromising their practical robustness [1-4]. Conversely, FB methods involve feature extraction and classifier utilization for modulation type determination. While reducing computational load, they require meticulous feature design and deep domain expertise [5-7]. Mismatched feature selection and classification models can severely impact recognition performance.

To overcome these limitations, deep learning-based modulation recognition has emerged as a promising solution. In AMR, deep learning models have evolved toward multimodal fusion and spatiotemporal feature optimization. O'Shea et al. pioneered the use of convolutional neural networks (CNNs) in modulation recognition, demonstrating superior performance over traditional methods [8].

DOI: 10.54254/2753-8818/101/2025.CH22300

Qi et al. enhanced classification robustness by integrating waveform and spectrum dual-modal features within ResNet, employing a cross-modal feature enhancement strategy [9]. Wu's team addressed temporal correlation modeling with a CNN-LSTM cascaded model, enabling serial spatiotemporal feature extraction [10]. Liu et al. developed a CNN-GRU architecture with parallel processing for low-SNR scenarios, employing dual-path collaborative learning [11]. Chen's team constructed a multi-stream CLDNN model, processing I/Q sequences and amplitude/phase information in parallel, achieving a 9% accuracy boost in dynamic channels [12]. To address feature weighting and sample efficiency, Wang et al. introduced an attention mechanism [13]. Their CNN-LSTM-Attention hybrid model dynamically weights spatiotemporal features, achieving a 59.8% recognition accuracy in few-sample scenarios, outperforming others by 0.3% to 7%. R. Liang et al. introduced a three-stream CNN-LSTM network that leverages interactive learning between phase, amplitude, and frequency feature streams, significantly enhancing performance in non-stationary channels [14].

The main contributions of the model presented in this paper are: By discretizing long sequence data into frames and reshaping them into small 2D images, the author employed a 2D CNN to fully exploit multi-channel IQ and AP signal features, achieving higher recognition accuracy across all SNRs for nine modulation types compared to existing models. Besides, the use of the CPO optimization algorithm enabled fine-tuning of model hyperparameters, resulting in a 2.2% accuracy improvement and a 20.7% parameter tailoring.

2. Signal Model and Data Preprocessing

This section delineates the signal model and data preprocessing procedures used in the study of AMR. The system under consideration is a single-carrier transmission where the complex-valued transmitted signal $s(t)$ propagates through a frequency-selective, time-varying fading channel with a delay spread τ . The received signal $y(t)$ can be expressed as:

$$y(t) = \exp\left(j\,\Delta f(t)\right) \int_0^{\tau} s\left(t - \Delta c(t) - \tau'\right) h\left(t, \tau'\right) d\tau' + n(t) \tag{1}$$

where $j = \sqrt{-1}$ is the imaginary unit, $\Delta c(t)$ denotes the instantaneous carrier frequency offset, $\Delta c(t)$ represents the sampling time offset, $h(t, \tau')$ is the time-varying channel impulse response, and n(t) is the additive white Gaussian noise (AWGN) with zero mean and variance σ_n^2. After discretizing the received signal $y(t)$ at a sampling rate $f_s = 1/T_s$, the discrete-time sequence $\{y[n]\}$ is obtained:

$$y[n] = y^I[n] + j\,y^Q[n], \quad n = 0,1,\ldots,N-1, \tag{2}$$

where $y^I[n]$ and $y^Q[n] \in R$ are the in-phase (I) and quadrature (Q) components, respectively. Within a time window, N discrete samples form a data vector:

$$y_j = \left[y_j[0],\ y_j[1],\ \ldots,\ y_j[N-1]\right]^T \in \mathbb{C}^N \tag{3}$$

For subsequent analysis and processing, real-valued feature representations are often preferred. Two common transformations are considered: First, the complex-valued data is decomposed into its in-phase and quadrature components, which are then stacked to form a real-valued vector:

$$X_j^{I/Q} = \begin{bmatrix} y_j^I[0] & \cdots & y_j^I[N-1] \\ y_j^Q[0] & \cdots & y_j^Q[N-1] \end{bmatrix} \in \mathbb{R}^{2N} \tag{4}$$

DOI: 10.54254/2753-8818/101/2025.CH22300

Alternatively, using the polar coordinate representation of complex numbers, where amplitude and phase serve as distinct features:

$$X_j^{\{A\}\{P\}} = \begin{bmatrix} |y_j[0]| & \cdots & |y_j[N-1]| \\ \arg\left(y_j[0]\right) & \cdots & \arg\left(y_j[N-1]\right) \end{bmatrix} \in \mathbb{R}^{2N} \tag{5}$$

Given the in-phase and quadrature components $y_j^I[n]$ and $y_j^Q[n]$, the amplitude and phase are computed as:

$$|y_j[n]| = \sqrt{(y_j^I[n])^2 + (y_j^Q[n])^2}\ ,\ \arg(y_j[n]) = \arctan(\frac{y_j^Q[n]}{y_j^I[n]}) \tag{6}$$

To resolve phase ambiguity, the $\text{atan2}(y_j^Q[n], y_j^I[n])$ function is employed in practice.

Effective data preprocessing ensures the quality and consistency of the inputs to neural network models. Both I/Q and A/P channels are normalized to have zero mean and unit variance, which mitigates the effects of varying signal power levels and presents the training process from being biased by differing amplitude scales. This normalization is mathematically represented as:

$$Y_{\text{norm}} = \frac{Y - \mu}{\sigma} \tag{7}$$

where X is the original data vector, μ is the mean of the data, and σ is the standard deviation. The I/Q representation preserves the original complex structure, supporting traditional linear filtering and estimation. Meanwhile, the A/P representation offers intuitive features for channel characteristic analysis and phase-sensitive post-processing. These real-valued, feature-rich representations lay the foundation for effective signal processing and deep learning modeling in the AMC framework.

3. MultiFrame-DualStream-2DCNN-LSTM Model Design

To capture both the spatiotemporal structure and inter-channel coupling in multi-channel IQ and AP signals, this paper segment the long sequential data into frames and reshape each frame into a 2D ”small image.” This transformation enables a 2D CNN to extract local spatial structures and model inter-channel interactions, while a Long Short-Term Memory (LSTM) network captures the dynamic evolution of frame-level sequences. Specifically, the original sequential data is split into overlapping frames, each assigned repeated labels to enhance sample diversity and provide frequent supervision during training. The 2D CNN applies convolution, batch normalization, and downsampling to each frame’s ”width × height × channels” tensor, compressing it into a feature representation while preserving local receptive field characteristics. These frame-level features, along with the repeated labels, are then passed sequentially into the LSTM, which captures long-term dependencies and further mines temporal information.

This approach combines spatial feature extraction via 2D CNNs (commonly used in image processing) and the” CNN + LSTM” hierarchical modeling used in video analysis and action recognition. By processing frames, the model strengthens local structure recognition within each frame while effectively modeling temporal dependencies across frames. This model is referred to as MultiFrame-DualStream-2DCNN-LSTM (MF-DS-2DCNN-LSTM), as shown in Figure 1.

DOI: 10.54254/2753-8818/101/2025.CH22300

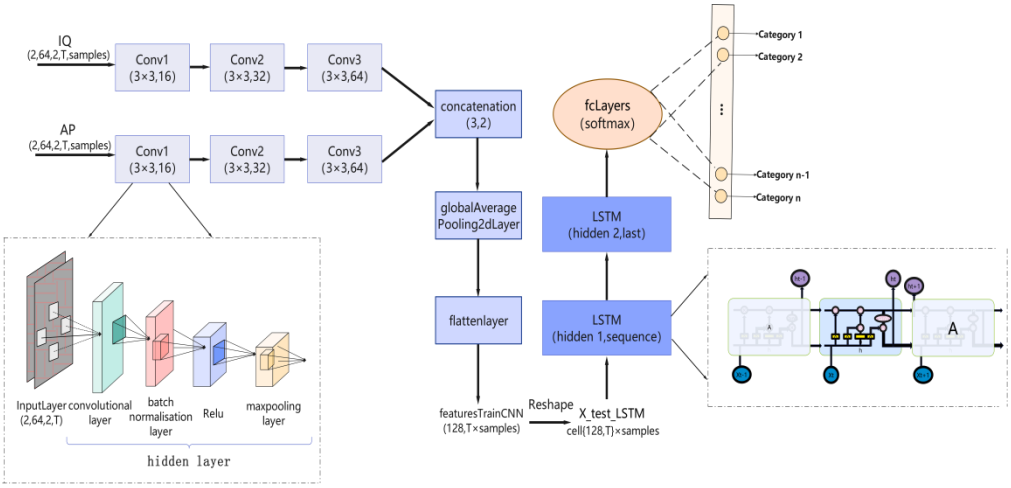

Figure 1: DualSteam-2DCNN-LSTM Model Diagrams

4. Crowded Porcupine Optimization Algorithm in Model Optimization

This study employs Crested Porcupine Optimizer (CPO) for hyperparameter optimization, a novel nature-inspired metaheuristic algorithm [15]. CPO is based on the defensive mechanisms of crested porcupines, effectively balancing global exploration and local exploitation in complex optimization landscapes.

CPO mimics the defensive strategies of crested porcupines, incorporating sight, sound, odor, and physical attack mechanisms into a mathematical framework to enhance the optimization process. These mechanisms are designed to facilitate transitions between exploration and exploitation phases. The search begins by initializing a population X of candidate solutions:

$$X_i = L + r \cdot (U - L), \ i = 1,2, \dots, N \tag{8}$$

where L and U represent the lower and upper bounds of the search space, r is a uniformly distributed random number in $[0, 1]$, and N is the population size. To simulate the natural behavior of crested porcupines, CPO integrates a Cyclic Population Reduction (CPR) strategy to dynamically adjusts the population size N_t during optimization:

$$N_t = N_{min} + \left(N' - N_{min}\right) \cdot \left(1 - \frac{t\%T_{max}}{T}\right) \tag{9}$$

Here, N_{min} and $N'$are the minimum and maximum population sizes, t is the current iteration, T is the reduction cycle, and T_{max} is the total number of iterations. The algorithm alternates between exploration and exploitation phases. In the exploration phase, sight and sound strategies are employed. For instance, the sight mechanism updates a candidate solution as follows:

$$X_i^{t+1} = X_{best}^t + \delta \cdot r \cdot \left(X_{rand}^t - X_i^t\right) \tag{10}$$

where X_{best}^t is the best solution so far, X_{rand}^t is a randomly chosen solution, δ is an exploration factor, and r is a random number. In the exploitation phase, odor and physical attack mechanisms refine the search. The physical attack mechanism, for example, is defined as:

$$X_i^{t+1} = X_{best}^t + \alpha \cdot r \cdot \left(X_{best}^t - X_{worst}^t\right) \tag{11}$$

DOI: 10.54254/2753-8818/101/2025.CH22300

where alpha is the attack intensity factor, X_{worst}^{t} is the worst solution, and r is a random number. The solutions are iteratively evaluated using a fitness function f(X), and the best solution X_{best} is returned as the optimal result after termination. The algorithm process is shown in the appendix.

5. Simulation Experiments and Results

5.1. Dataset and Simulation Parameter Settings

The RadioML 2018.01A dataset, provided by DeepSig, consists of 24 modulation types, including both analog and digital modulation schemes. Each signal sample contains 1024 IQ data points, with Signal-to-Noise Ratio (SNR) ranging from -20 dB to 30 dB in 2 dB increments, covering 26 distinct SNR levels. The dataset contains 210,000 samples, making it suitable for wireless signal modulation classification tasks and ideal for training and evaluating machine learning and deep learning models.

For this study, representative modulation schemes were selected from this dataset for classifcation, including amplitude modulation (OOK, 4ASK, AM-DSB-SC), phase modulation (BPSK, QPSK, 8PSK), combined amplitude and phase modulation (32QAM, 64QAM), and frequency modulation (FM). These modulation types cover the primary signal categories in wireless communication systems.

During training, samples with SNR values ranging from -10 dB to 20 dB (in 2 dB increments) were selected. The dataset was split into training and testing sets in an 80:20 ratio, with 1024 frames per modulation type allocated for training. The experimental platform used was Matlab R2023b, running on a hardware environment consisting of an RTX 2050 GPU and 16 GB of memory.

For the MF-DS-2DCNN-LSTM model, the parameter settings are provided in Table 1. Under these conditions, the accuracy training curve of the MF-DS-2DCNN-LSTM model was obtained, as depicted in Figure 2.

Table 1: Model Parameters (MF-DS-2DCNN-LSTM)

Component	Parameters and Configuration
Data Preparation (CNN)	Time frames: T=8; Input shape: 5D array [2, 64, 2(IQ/AP), T, samples]
CNN Architecture	Convolutional layers: conv1/4: 16 filters (3×3 kernel); conv2/5: 32 flters (3×3 kernel); conv3/6: 64 filters (3×3 kernel) Feature fusion: 3D concatenation layer
CNN Training	Max epochs: 20; Initial learning rate: 1e-3; L2 regularization: 1e-5 Learning schedule: Drop factor 0.5 every 5 epochs; Batch size: 128
Data Preparation (LSTM)	Time frames: T=8 Input format: [2×64, T] cell array from CNN fatten layer
LSTM Architecture	Stacked LSTM layers: First layer: 64 hidden units; Second layer: 64 hidden units
LSTM Training	Max epochs: 30; Initial learning rate: 2e-3; L2 regularization: 2e-5 Learning schedule: Drop factor 0.5 every 5 epochs; Batch size: 128

DOI: 10.54254/2753-8818/101/2025.CH22300

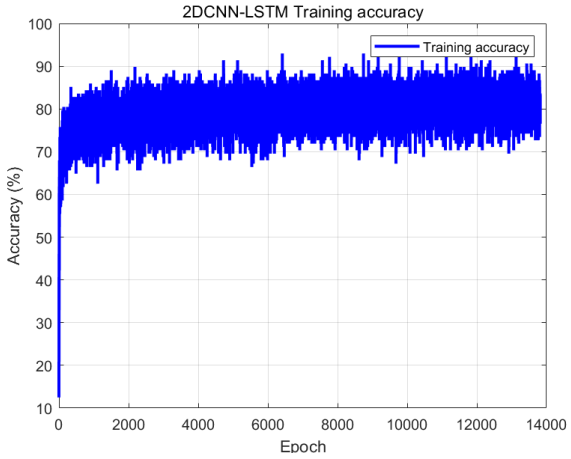

Figure 2: Training Curve

5.2. Results

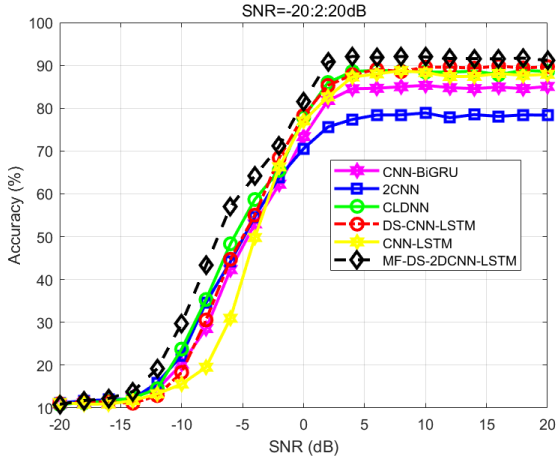

Figure 3: Comparison of MF-DS-2DCNN-LSTM model with other models at different SNRs

Figure 3 compares five CNN and LSTM variant models with the MF-DS-2DCNN-LSTM model at different SNR levels, including CNN-BiGRU, two CNNs (2CNN), CLDNN, Dual-Stream CNN-LSTM (DS-CNN-LSTM), and CNN-LSTM. The results indicate that at low SNRs (-10 dB to 0 dB), the MF-DS-2DCNN-LSTM model significantly outperforms all other models. At higher SNRs (above 2 dB), the model achieves an average accuracy of 94.0%. Overall,the MF-DS-2DCNN-LSTM model demonstrates a clear performance improvement compared to other models.

Figure 4 displays the confusion matrix showing the recognition accuracy of nine modulation schemes by the MF-DS-2DCNN-LSTM model at an SNR of 2 dB. Most schemes achieving recognition rates above 90%. However, for 32QAM and 64QAM, despite reduced noise under high SNR conditions, the subtle differences between symbol points remained difficult to capture, preventing complete recognition.

DOI: 10.54254/2753-8818/101/2025.CH22300

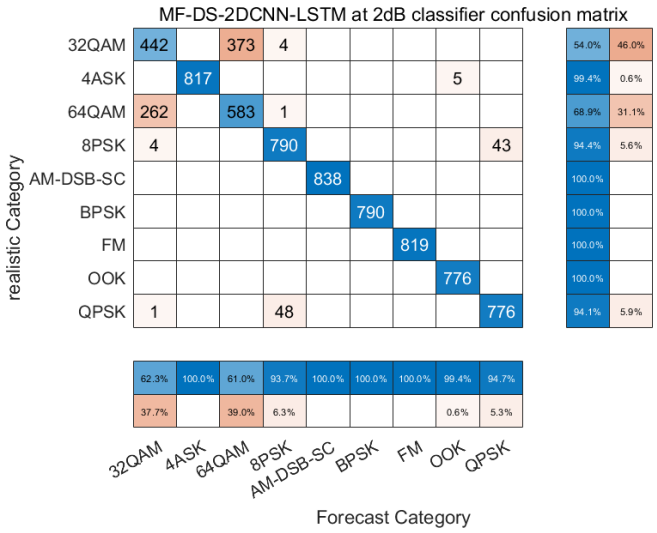

Figure 4: Confusion matrix for nine modulation schemes at 2 dB SNR

To assess the MF-DS-2DCNN-LSTM model's performance with small sample sizes, a small-sample training experiment was conducted. In this experiment, only 128 training samples were used (compared to 1024 in previous simulations). Accuracy curves for each model were analyzed across an SNR range from -20 to 20 dB. As shown in Figure 5, the MF-DS-2DCNN-LSTM model continues to outperform other models, maintaining accuracy above 90% at SNRs above 4 dB.

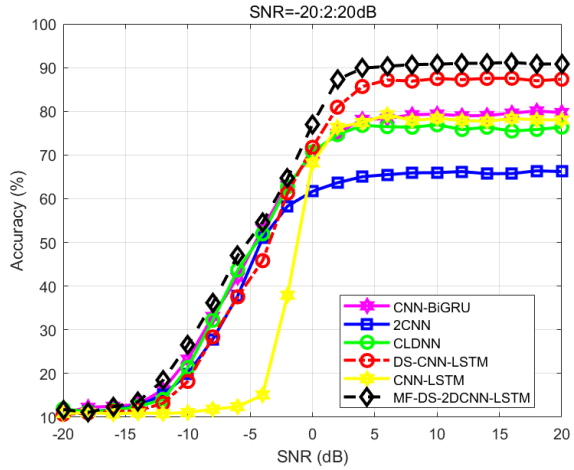

Figure 5: Accuracy curves for various models with 128 samples across SNR range of -20 to 20 dB

This study also uses the CPO algorithm to optimize hyperparameters of a deep learning model, specifically targeting the learning rate of the LSTM component, the node numbers in two hidden layers, and the L2 regularization coefficient. The algorithm is initialized with six search agents to explore the hyperparameter space in parallel. Each agent represents a unique combination of hyperparameters evaluated through a predefined objective function. With a maximum of six cycles, the agents iteratively refine their hyperparameter values based on real-time performance feedback. The search space is bounded by lower (lb) and upper (ub) limits: the learning rate spans 1×10^{-5} to 1×10^{-2}, the L2 regularization coefcient ranges from 1×10^{-5} to 1×10^{-3}, and both hidden layers' node counts are restricted to [32, 128].

DOI: 10.54254/2753-8818/101/2025.CH22300

Table 2: Comparison of Model Parameters

Model Parameter	MF-DS-CNN-LSTM	MF-DS-CNN-LSTM (CPO)
Learning Rate	2e-3	0.0018588
l2 Regularization Factor	1e-4	0.00013821
LSTM Hidden Layer 1	64	43
LSTM Hidden Layer 2	64	74
Model Parameters	87.1k	69k
Flops	3.682M	2.497M

As shown in Table 2, the CPO algorithm successfully identifies the optimal hyperparameter configuration through systematic iteration and evaluation. As illustrated in Figure 6, the CPO-optimized MF-DS-CNN-LSTM model achieves an accuracy of 93.1% under high SNR conditions, improving overall performance by 2.2% across all SNR levels. In addition, with optimal parameters, the model is more lightweight, achieving a 20.7% reduction in the number of parameters.

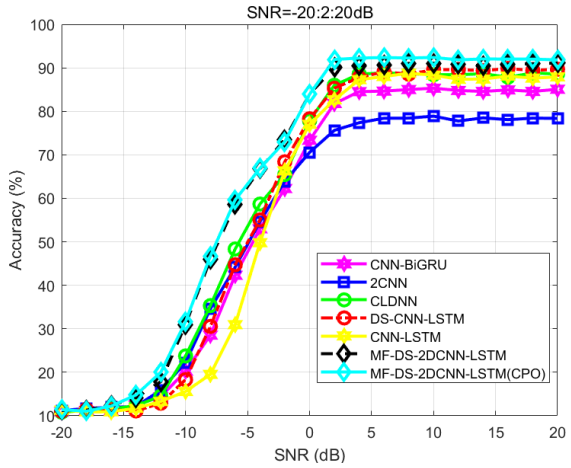

Figure 6: Comparison of CPO-optimized MF-DS-2DCNN-LSTM model with other models at different SNRs

6. Conclusion

This paper introduces a novel MultiFrame-DualStream-2DCNN-LSTM (MF-DS-2DCNN-LSTM) model for automatic modulation recognition in complex wireless channel environments. The proposed model enhances the traditional approach by dividing the original signal sequence into two-dimensional frames, which are processed using 2D Convolutional Neural Networks (CNNs) with Long Short-Term Memory (LSTM) networks. This hybrid architecture effectively captures both spatial and temporal features. Specifically, the model transforms the signal into small 2D images, where CNNs extract spatial information and inter-channel relationships, while LSTMs capture temporal dependencies between the frames. Inspired by hierarchical modeling techniques used in image and video analysis, this method significantly enhances the joint learning of spatiotemporal features, offering an advanced solution for automatic modulation recognition. To further optimize model performance, the Collaborative Parameter Optimization (CPO) algorithm was employed for hyperparameter tuning, The optimized model showed a notable improvement in performance, achieving 94.0% accuracy under high Signal-to-Noise Ratio (SNR) conditions, outperforming other existing models such as CLDNN. Additionally, the MF-DS-2DCNN-LSTM model demonstrated strong robustness and generalization, maintaining over 90% accuracy for SNR levels above 4 dB even with limited training data. After applying CPO optimization, the accuracy of the entire model

DOI: 10.54254/2753-8818/101/2025.CH22300

improved by 2.2%, and the number of parameters was reduced by 20.7%, making it more efficient and lightweight.

In conclusion, the MF-DS-2DCNN-LSTM model presents a novel and effective approach to automatic modulation recognition across varying SNR conditions. The results not only offer valuable insights but also provide a solid foundation for future research and real-world applications in automatic modulation classification, particularly in wireless classification systems where environmental conditions are often challenging.

References

[1] *Ramezani-Kebrya, A., Kim, I. M., Kim, D. I., Chan, F., & Inkol, R. (2013). Likelihood-based modulation cla ssification for multiple-antenna receiver. IEEE Transactions on Communications, 61(9), 3816-3829.*

[2] *Abdi, A., Dobre, O. A., Choudhry, R., Bar-Ness, Y., & Su, W. (2004, October). Modulation classification in fading channels using antenna arrays. In IEEE MILCOM 2004. Military Communications Conference, 2004. (Vol. 1, pp. 211-217). IEEE.*

[3] *El-Mahdy, A. E., & Namazi, N. M. (2002). Classification of multiple M-ary frequency-shift keying signals over a Rayleigh fading channel. IEEE Transactions on Communications, 50(6), 967-974.*

[4] *Wei, W., & Mendel, J. M. (2000). Maximum-likelihood classification for digital amplitude-phase modulations. IEEE transactions on Communications, 48(2), 189-193.*

[5] *Swami, A., & Sadler, B. M. (2000). Hierarchical digital modulation classification using cumulants. IEEE Transactions on communications, 48(3), 416-429.*

[6] *Gardner, W. A., & Spooner, C. M. (1988, October). Cyclic spectral analysis for signal detection and modulation recognition. In MILCOM 88, 21st Century Military Communications-What's Possible?'. Conference record. Military Communications Conference (pp. 419-424). IEEE.*

[7] *Marchand, P., Le Martret, C., & Lacoume, J. L. (1997, July). Classification of linear modulations by a combination of different orders cyclic cumulants. In Proceedings of the IEEE Signal Processing Workshop on Higher-Order Statistics (pp. 47-51). IEEE.*

[8] *O'Shea, T. J., Corgan, J., & Clancy, T. C. (2016). Convolutional radio modulation recognition networks. In Engineering Applications of Neural Networks: 17th International Conference, EANN 2016, Aberdeen, UK, September 2-5, 2016, Proceedings 17 (pp. 213-226). Springer International Publishing.*

[9] *Qi, P., Zhou, X., Zheng, S., & Li, Z. (2020). Automatic modulation classification based on deep residual networks with multimodal information. IEEE Transactions on Cognitive Communications and Networking, 7(1), 21-33.*

[10] *Wu, Y., Li, X., & Fang, J. (2018, June). A deep learning approach for modulation recognition via exploiting temporal correlations. In 2018 IEEE 19th international workshop on signal processing advances in wireless communications (SPAWC) (pp. 1-5). IEEE.*

[11] *Liu, F., Zhang, Z., & Zhou, R. (2021). Automatic modulation recognition based on CNN and GRU. Tsinghua Science and Technology, 27(2), 422-431.*

[12] *Chen, C., Zhang, N., Yang, C., Gao, Y., & Sheng, Y. (2023, May). Multiple-Stream Model with CNN-LSTM for Automatic Modulation Recognition. In 2023 International Conference on Microwave and Millimeter Wave Technology (ICMMT) (pp. 1-3). IEEE.*

[13] *Wang, Z., Wang, P., & Lan, P. (2022, December). Automatic modulation classification based on CNN, LSTM and attention mechanism. In 2022 IEEE 8th International Conference on Computer and Communications (ICCC) (pp. 105-110). IEEE.*

[14] *Liang, R., Yang, L., Wu, S., Li, H., & Jiang, C. (2021, October). A three-stream CNN-LSTM network for automatic modulation classification. In 2021 13th International Conference on Wireless Communications and Signal Processing (WCSP) (pp. 1-5). IEEE.*

[15] *Abdel-Basset, M., Mohamed, R., & Abouhawwash, M. (2024). Crested Porcupine Optimizer: A new nature-inspired metaheuristic. Knowledge-Based Systems, 284, 111257.*

Appendix

Algorithm: Crested Porcupine Optimizer (CPO)
Input: $N, T_{max}, N_{min}, T, \alpha, F, S, \delta$
Output: X_{best}
Initialize X_i randomly
Evaluate $f(X_i)$ and set X_{best}

DOI: 10.54254/2753-8818/101/2025.CH22300

$t \leftarrow 0$
while $t < T_{max}$ do:
Compute $N_t = N_{min} + (N - N_{min}) \cdot \left(1 - \frac{t\%T_{max}}{T}\right)$
for $i = 1$ to N_t do:
Generate $\tau_1, \tau_2, \tau_3 \in [0,1]$
if $\tau_1 < 0.5$:
if $\tau_2 < 0.5$: $X_i^{t+1} = X_{best} + \delta \cdot rand() \cdot (X_{rand} - X_i)$
else: $X_i^{t+1} = X_i + S \cdot sign(rand()\text{-}rand()) \cdot (X_{mean} - X_i)$
else:
if $\tau_3 < 0.5$: $X_i^{t+1} = X_{best} + F \cdot (X_{mean} - X_i)$
else: $X_i^{t+1} = X_{best} + \alpha \cdot rand() \cdot (X_{best} - X_{worst})$
Evaluate $f(X_i^{t+1})$
if $f(X_i^{t+1}) < f(X_i)$:$X_i = X_i^{t+1}$
Update X_{best}
$t \leftarrow t + 1$
return X_{best}