MM-Net: A Multi-Modal approach towards Automatic Modulation Classification

Konstantinos Triaridis, Constantine Doumanidis, Nestor D. Chatzidiamantis, *Member, IEEE*, and George K. Karagiannidis, *Fellow, IEEE*

***Abstract*—Automatic Modulation Classification (AMC) has become an important component in communication systems for both civil and defense applications. The shortcomings of traditional approaches to AMC have led researchers to develop complex machine learning (ML)-based approaches. In this work, inspired by multi-modal approaches for general Computer Vision tasks like Semantic Segmentation, we propose MM-Net, a multimodal approach to AMC that uses domain-specific features in the form of Higher Order Cumulants (HOCs) to improve classification performance. Furthermore, we explore the usage of HOCs in existing Deep Learning (DL)-based applications for AMC. Simulation results show that for eight modulation classification, MM-Net achieves high classification accuracy even at low SNRs, demonstrating the robustness of the multimodal approach even under challenging channel conditions, while existing methods are improved by utilizing HOCs, especially at low SNR values.**

***Index Terms*—automatic modulation classification, machine learning, deep learning, cumulants, convolutional neural networks, transfer learning**

I. Introduction

Automatic Modulation Classification (AMC) refers to the process of recognizing the modulation of a radio signal and is inherently a multi-class classification problem. It is the immidiately preceding step of signal demodulation [1], and is considered to be a problem of considerable importance for several communication systems. AMC can be an invaluable tool for both civil and military applications. It can provide the tools for highly efficient spectrum management, which is very important for modern wireless communications systems such as fifth generation (5G) infrastructure [2]. Software Defined Radio (SDR) also interacts with a wide range of telecom systems, so the use of AMC is essential. In addition, many military applications use AMC for advanced tasks such as signal detection, identification, processing and jamming [3].

The traditional approach to solving the problem of AMC has been to use signal processing methods that are either likelihood-based [3, 4] or feature-based [5–7], in particular using Higher Order Cumulants (HOCs) as inputs. Recently, researchers have adopted ML methods for AMC due to significant drawbacks of both traditional methods such as the very high computational cost of the likelihood-based algorithms and the challenging feature engineering for feature-based methods, as well as the limited accuracy of both. Today, as in many classification problems, DL-based approaches with complex architectures have established themselves as the state-of-the-art, while HOCs are mostly used as input for simple ML-based models that serve as baselines for researchers to test the performance of their new and increasingly complex DL-based methods [8–10].

While many different DL-based architectures have been utilized, the general robustness of convolutional neural networks (CNNs) in visual recognition tasks has led to their widespread adoption for the task of AMC. In their work, Meng et al [10] use a time series of complex IQ samples fed into a novel CNN for classification. In later works, researchers, valuing the robustness of CNNs for image classification tasks, convert IQ samples into constellation map images and achieve higher accuracy for a wide range of signal-to-noise ratio (SNR) values [1, 8, 9, 11, 12]. This approach to data representation is by far the most popular in modern AMC methods.

Several techniques using CNN configurations have been proposed, including VGG-based [8] and AlexNet-based [12] models, which contain a series of narrowing convolutional blocks followed by a fully connected classifier. However, the state of the art in AMC is represented by novel architectures that do not use large pre-trained backbones [1, 2, 11, 13]. The solutions that use pretrained models (AlexNet, GoogleNet[9], MobileNetV2[14]) as feature extractors generally lag behind in performance compared to novel architectures for AMC.

Furthermore, recent research trends in DL show that the choice of the correct data representation can significantly improve the robustness of a DL model. In this direction, several recent works have explored different ideas. Zeng et al [15] use the spectrogram of the data, while Zhang et al [14], Huang et al [13], and Peng et al [9] all propose different ways of preprocessing the constellation images. However, methods that only use visual representations of the data to address AMC often suffer in performance at lower SNR levels. Using multi-modal inputs is also seemingly helpful for a wide variety of general Computer Vision tasks such as semantic segmentation, object detection, and even facial recognition. In the case of AMC, HOCs display several useful properties, such as robustness to Gaussian Noise, so we theorize that their inclusion into networks with a CNN encoder and classifier is a space and time-efficient method to increase performance. To test this theory, we create MM-Net, a dual-encoder, single-classifier network that uses both images and HOCs. We observe that this network displayed increased performance compared to the CNN encoder baseline. Based on this observation, we further apply this approach to more efficient AMC networks that use a CNN encoder and a classifier: FiF-Net[1] and CCNN[13]. We demonstrate the effectiveness of networks using the MM-Net

K. Triaridis, C. Doumanidis and N. D. Chatzidiamantis are with the Department of Electrical and Computer Engineering, Aristotle University of Thessaloniki, 54124 Thessaloniki, Greece (e-mails: triaridis@ece.auth.gr, kdoumani@ece.auth.gr, nestoras@auth.gr).

G. K. Karagiannidis is with Department of Electrical and Computer Engineering, Aristotle University of Thessaloniki, Greece and also with Artificial Intelligence & Cyber Systems Research Center, Lebanese American University (LAU), Lebanon (geokarag@auth.gr)

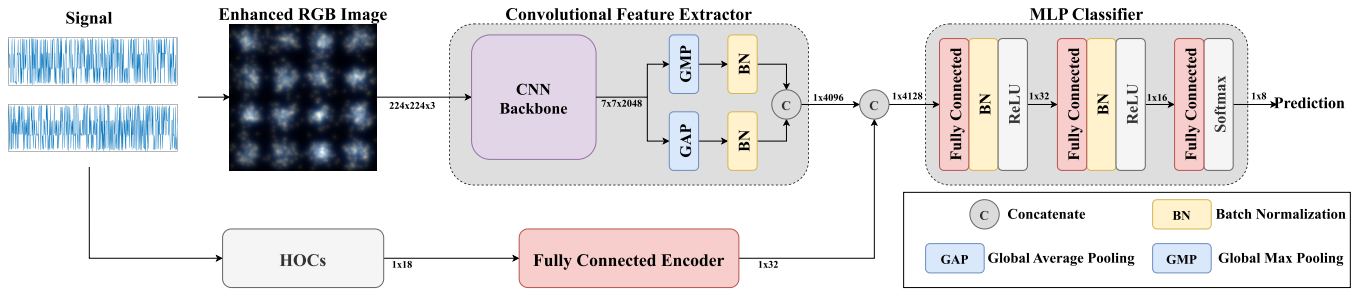

Fig. 1. High-level overview of MM-Net.

method as compared to the baseline networks while consuming the same network parameters. We also show the extensibility of the proposed approach by substituting the base structure in state-of-the-art AMC networks with the proposed architecture and showcase that it is applicable to any AMC network with an encoder-decoder architecture. Our contribution can be summarized as follows:

- To the best of our knowledge, we are the first to investigate combining heterogeneous features in the form of HOCs and constellation images as inputs of neural networks and their effectiveness for AMC.
- In this spirit, we propose MM-Net, a simple dual-encoder single-decoder paradigm for AMC that utilises both enhanced constellation images and HOCs. Experimental results validate that the proposed approach successfully guides the network to learn with complementary features and improves the classification performance for a wide range of SNR values.
- The proposed MM-Net paradigm can be implemented in any existing state-of-the-art AMC network that is based on an encoder-decoder architecture without increasing the model size in any significant capacity, and extensive experimental results showcase increased performance of those networks.

Code and models are available at https://github.com/CedArctic/AutoModClass. Data generation tool is available at https://github.com/kostino/AMCDataGen.

II. Methodology

We propose a multi-modal approach by designing MM-Net: a dual-encoder paradigm that utilizes both hand-crafted expert features (in the form of HOCs) and constellation images. MM-Net consists of two branches; one that utilizes a Convolutional Neural Network (CNN) to extract features from the constellation diagrams, and one that uses the HOCs as input, aiming to achieve high accuracy in a wide range of SNR levels for the task of 8-class Modulation Classification. The complete architecture is shown in Fig. 1.

First, as done in previously discussed work, we window N incoming samples $y \in \mathbb{C}^N$ and represent them in an IQ diagram $y_{iq} = IQ(y) \in \mathbb{R}^{H \times W}$, a diagram of height H and width W. Afterward, for each IQ diagram, we construct an enhanced 3-channel RGB image $y_{enh} = F_{enh}(y) \in \mathbb{R}^{3 \times H \times W}$ using the exponential decay method proposed by Peng et al [9]. With this transformation, we create a pixel representation that is more information-rich for the Deep Network to exploit. Furthermore, we propose a statistical method for creating the images by dynamically calculating the capture window size in the IQ space instead of choosing a static size as in previous work [9]. Our method aims at capturing the maximum amount of information whilst utilizing as much of the image resolution as possible : Given a padding ratio p, and $m = max(y)$ a padding offset $o = \frac{p \cdot m}{100 - p}$ is added resulting in IQ space windows with size of $(2(m+o)) \times (2(m+o))$.

Inspired by previous approaches[5, 6], we calculate the HOCs C20 - C63 and use their real and imaginary parts as input $y_{cum} \in \mathbb{R}^{2 \times 9}$ for the cumulant encoder. The HOCs are calculated through the higher-order moments as follows:

$$M_{pq} = E[y^{p-q}(y^*)^q] \tag{1}$$

$$C_{pq} = cum(\underbrace{y(n), \ldots, y(n)}_{(p-q) \text{ times}}, \underbrace{y^*(n), \ldots, y^*(n)}_{q \text{ times}}) \tag{2}$$

For instance:

$$\begin{aligned} C_{42} &= cum(y(n), y(n), y^*(n), y^*(n)) = \\ &= M_{42} - |M_{20}|^2 - 2 \cdot M_{21}^2 \end{aligned} \tag{3}$$

The enhanced images y_{enh} are fed into the convolutional feature extractor to extract the feature vector $f_{conv} = Y_{conv}(y_{enh})$, while the cumulants y_{cum} are fed into a fully connected encoder to produce the feature vector $f_{cum} = Y_{fc}(y_{cum})$. Both sets of features are then utilized by our MLP Classifier to produce a prediction score $z = MLP(f_{cum}, f_{enh})$.

A. Cumulant Feature Extractor

As shown in Fig 1, the first branch of MM-Net implements a simple MLP encoder that maps the input to a latent space of a higher dimension for feature enrichment. The encoder consists of two hidden layers of 32 neurons and batch normalization layers: the 1×18 input vector produced by the 9 complex HOCs is mapped to a 1×32 output.

B. Convolutional Feature Extractor

The second branch, displayed in Fig 1, feeds the $224 \times 224 \times 3$ enhanced constellation image to a convolutional feature extractor module. The module consists of a CNN backbone whose output feature maps are fed into two parallel processing flows which contain a Global Pooling layer (Global Average Pooling and Global Max Pooling respectively), followed by

a batch normalization layer. This approach is inspired by the Convolutional Block Attention Module (CBAM) [16], as Average Pooling effectively captures and models spatial information, while Max Pooling gathers another important clue about distinctive object features and channel-wise characteristics. The outputs are then concatenated as the output of the convolutional feature extractor. The backbone used for our Convolutional feature extractor is a ResNet152V2 pre-trained on the ImageNet dataset. We elect to use a pre-trained CNN due to its simplicity of use as a plug-and-play module. Our focus is on showcasing the efficacy of utilizing the HOCs in conjunction with the RGB image compared to a baseline, Image-only based approach and not to design an overly specialized AMC architecture.

C. Classifier

The network is completed by feeding the concatenated output of the two branches into a dense classifier block with the following structure: 32-neuron dense layer $\rightarrow$ batch normalization (BN) layer$\rightarrow$ Rectified Linear Unit (ReLU) $\rightarrow$ 16-neuron dense layer $\rightarrow$ BN layer$\rightarrow$ ReLU $\rightarrow$ 8-neuron dense layer $\rightarrow$ softmax layer. The BN layers are used to prevent overfitting, and improve generalization performance.

D. Extension to existing AMC networks

MM-Net's approach for utilizing cumulants can be applied to any conventional encoder-decoder networks for AMC with a convolutional feature extractor $\mathcal{E}$ and a dense classifier $\mathcal{C}$, such as [1] and [13]. We increase the input size of their classifier by 32 and concatenate the output of their feature encoder with the output of our cumulant encoder. Like in our network, the images x are fed into the convolutional feature extractor $\mathcal{C}$ to extract the feature vector $f_{conv} = \mathcal{E}(x)$, while the cumulants x_{cum} are fed into a fully connected encoder to produce the feature vector $f_{cum} = Y_{fc}(x_{cum})$. Both sets of features are then utilized by the dense Classifier to produce a prediction score $z = \mathcal{C}(f_{cum}, f_{enh})$. In the following section, we compare the MM-Net approach of using cumulants to the baseline image-only based methods using our network, FiF-Net[1] and CCNN[13] to confirm the efficiency and effectiveness of the MM-Net method.

III. Results

A. Experimental Setup

For the purposes of this work, a synthetic dataset was generated using our data generation tool. The dataset contains samples for 8 different representative modulations: QPSK, 8-PSK, 16-QAM, 64-QAM, 16-APSK, 64-APSK, 4PAM, 16PAM, and 4 different SNR values: 0, 5, 10, 15 dB. We use 15,000 samples for each Modulation for each SNR value for training, and 1,000 for testing, resulting in a training dataset containing 480,000 samples, and a testing dataset of 32,000 samples. We choose this SNR range as we consider the problem of AMC to be trivial for very high SNR values (above 20 dB), while simultaneously being less useful for SNR values less than 0 dB where accurate demodulation will be challenging.

MM-Net is implemented in Tensorflow and trained on an NVIDIA GeForce RTX 2080 GPU. The input size of the enhanced RGB images is 224×224. The selected ResNet-152V2 backbone is initialized with ImageNet-pretrained weights and kept frozen during training. For training, an SGD optimizer is used with a 10^{-3} linearly decaying learning rate, momentum equal to 0.9, and a mini-batch size of 100.

B. Results

In this section, we demonstrate the results from our experiments. First, we train all methods using our synthetic dataset and the baseline approach. Then, we extend all networks with the MM-Net approach and jointly train all parts (feature extractor, cumulant encoder, classifier) on our dataset. We observe that all methods (ours, FifNet, CCNN) showcase increased performance while using both HOCs and constellation images compared to their baseline approach of images only(Table I). The proposed approach consistently recorded performance improvement in terms of accuracy as compared with the baseline networks.

TABLE I
COMPARISON OF EXISTING METHODS

Model	Param. Count (M)	Accuracy
FiFNet	0.161	97.43%
FiFNet + HOCs	0.165(+2.4%)	98.56% (+1.13%)
CCNN	0.877	98.27%
CCNN + HOCs	0.894(+1.9%)	99.20%(+0.93%)
MM-Net	20.265	98.40%
MM-Net + HOCs	20.267	99.23%(+0.83%)

C. Complexity analysis

For all methods, we observe that our approach does not incur a significant complexity cost, as it adds less than 20K parameters for all networks that we experimented with. This is expected as the only two changes are adding the cumulant encoder (<1K params) and increasing the input size of the network classifier by 32 to incorporate the features produced by the cumulant encoder. This is showcased clearly in Table I where we present the size of all networks (in millions of parameters) before and after applying our method

D. Robustness Analysis

In this section, we report the classification accuracy of MM-Net for each of the eight different modulations under different channel conditions. As can be understood from Figure 2, MM-Net achieves accuracy values of over 90% for all modulation types for SNR values over 5dB. We also observe that the difference in classification accuracy between higher order modulations such as 64-QAM, and 64-APSK with their lower order counterparts (e.g. 16-QAM, 16-APSK) is not as extreme as that reported in other works [1], which further showcases the efficacy of our approach in more challenging conditions. For qualitative analysis, we also report the incorrect predictions in the confusion matrix for MM-Net in Figure 2.

We compare all methods across SNR values and display the results in Figure 2. We see that adding HOCs increases

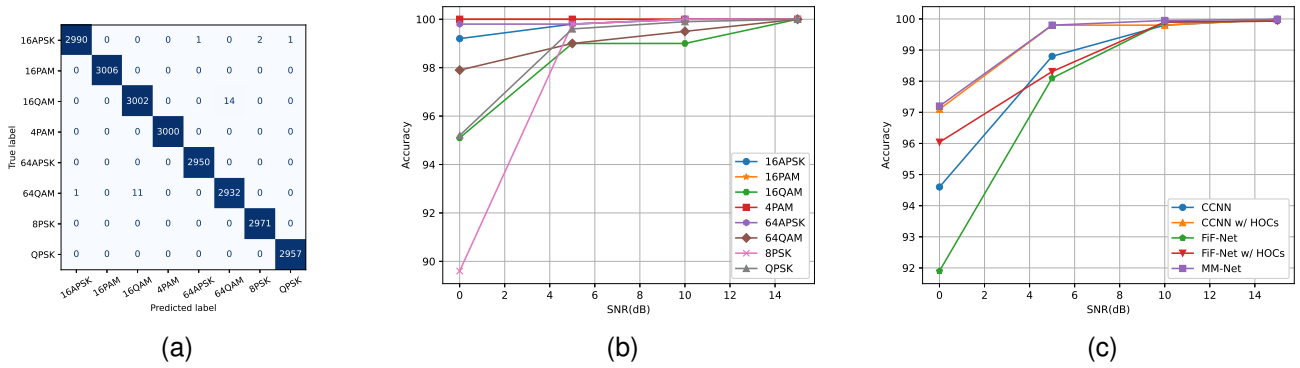

Fig. 2. (a) Confusion matrix for model predictions at SNR=5dB. (b) Classification performance of MM-Net for different SNR Values for different classes. (c) Classification performance of different methods

the accuracy of all methods and can specifically observe the robustness of the multi-modal approach proposed under harsh channel conditions, as the accuracy of FiF-Net is increased by 4.1% and CCNN by 2.8% for SNR = 0 dB. MM-Net outperforms other methods at a wide range of SNR values, as seen in Figure 2.

IV. Conclusions

In this paper, we have proposed utilizing both HOCs and constellation images in tandem for DL-based AMC methods. Our proposed network, MM-Net utilizes both a deep CNN to efficiently learn deep visual features from constellation images and a higher-order cumulant encoder to learn a useful representation of complex statistical features. Based on classification performance on eight modulation classification, MM-Net achieves classification accuracy of over 97% at 0dB SNR. This showcases the robustness of the multi-modal approach when facing challenging channel impairments and a total classification accuracy of 99.23%, showing its overall efficacy. We also apply our method on two existing AMC networks, FiF-Net, and CCNN and observe an accuracy increase of 1.13% and 0.93%, respectively. We especially observe a big accuracy increase of 4.1% for FiF-Net and 2.8% for CCNN in SNR = 0dB when using HOCs, further showcasing the robustness of our approach. Our approach for incorporating HOCs in DL-based methods is also computationally efficient as it only adds a small number of parameters to the models.

References

[1] V.-S. Doan, T. Huynh-The, C.-H. Hua, Q.-V. Pham, and D.-S. Kim, "Learning constellation map with deep cnn for accurate modulation recognition," in *GLOBECOM 2020 - 2020 IEEE Global Communications Conference*, 2020, pp. 1–6.

[2] T. Huynh-The, C.-H. Hua, Q.-V. Pham, and D.-S. Kim, "Mcnet: An efficient cnn architecture for robust automatic modulation classification," *IEEE Communications Letters*, vol. 24, no. 4, pp. 811–815, 2020.

[3] O. Dobre, A. Abdi, Y. Bar-Ness, and W. Su, "A survey of automatic modulation classification techniques: Classical approaches and new trends," vol. 1, pp. 137–156(19), April 2007.

[4] B. G. Mobasseri, "Digital modulation classification using constellation shape," *Signal Processing*, vol. 80, no. 2, pp. 251–277, 2000.

[5] M. W. Aslam, Z. Zhu, and A. K. Nandi, "Automatic modulation classification using combination of genetic programming and knn," *IEEE Transactions on Wireless Communications*, vol. 11, no. 8, pp. 2742–2750, 2012.

[6] M. Mirarab and M. Sobhani, "Robust modulation classification for psk /qam/ask using higher-order cumulants," in *2007 6th International Conference on Information, Communications Signal Processing*, 2007, pp. 1–4.

[7] A. Nandi and E. Azzouz, "Algorithms for automatic modulation recognition of communication signals," *IEEE Transactions on Communications*, vol. 46, no. 4, pp. 431–436, 1998.

[8] T. J. O'Shea, T. Roy, and T. C. Clancy, "Over-the-air deep learning based radio signal classification," *IEEE Journal of Selected Topics in Signal Processing*, vol. 12, no. 1, pp. 168–179, 2018.

[9] 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, 2019.

[10] F. Meng, P. Chen, L. Wu, and X. Wang, "Automatic modulation classification: A deep learning enabled approach," *IEEE Transactions on Vehicular Technology*, vol. 67, no. 11, pp. 10 760–10 772, 2018.

[11] T. Huynh-The, C.-H. Hua, V.-S. Doan, Q.-V. Pham, T.-V. Nguyen, and D.-S. Kim, "Deep learning for constellation-based modulation classification under multipath fading channels," in *2020 International Conference on Information and Communication Technology Convergence (ICTC)*, 2020, pp. 300–304.

[12] S. Peng, H. Jiang, H. Wang, H. Alwageed, and Y.-D. Yao, "Modulation classification using convolutional neural network based deep learning model," in *2017 26th Wireless and Optical Communication Conference (WOCC)*, 2017, pp. 1–5.

[13] S. Huang, L. Chai, Z. Li, D. Zhang, Y. Yao, Y. Zhang, and Z. Feng, "Automatic modulation classification using compressive convolutional neural network," *IEEE Access*, vol. 7, pp. 79 636–79 643, 2019.

[14] W. Zhang, D. Zhu, Z. He, N. Zhang, X. Zhang, H. Zhang, and Y. Li, "Identifying modulation formats through 2d stokes planes with deep neural networks," *Opt. Express*, vol. 26, no. 18, pp. 23 507–23 517, Sep 2018.

[15] Y. Zeng, M. Zhang, F. Han, Y. Gong, and J. Zhang, "Spectrum analysis and convolutional neural network for automatic modulation recognition," *IEEE Wireless Communications Letters*, vol. 8, no. 3, pp. 929–932, 2019.

[16] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, "Cbam: Convolutional block attention module," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 3–19.