Optics Communications 560 (2024) 130465

Contents lists available at ScienceDirect

Optics Communications

journal homepage: www.elsevier.com/locate/optcom

Intelligent non-cooperative optical networks: Leveraging scattering neural networks with small training data

Yinglin Chen [a], Tianhua Xu [b], Tongyang Xu [c,*]

[a] 6G Research Center, China Telecom Research Institute, Guangzhou 510660, China
[b] School of Engineering, University of Warwick, Coventry CV4 7AL, United Kingdom
[c] School of Engineering, Newcastle University, Newcastle upon Tyne, NE1 7RU, United Kingdom

ARTICLE INFO

Keywords:
Optical fiber communication
Non-cooperative optical network
Machine learning
Wavelet scattering
CNN
Wavelet transform

ABSTRACT

Artificial intelligence (AI) is enabling intelligent communications where learning based signal classification simplifies optical network signal allocation and shifts signal processing pressure to each network edge. This work proposes a non-orthogonal signal waveform framework that leverages its unique spectral compression characteristic as a user address for efficiently forwarding messages to target users. The primary focus of this work lies in the physical layer intelligent receiver design, which can automatically identify different received signal formats without preamble notification in a non-cooperative communication approach. Traditional signal classification methods, such as convolutional neural network (CNN), rely on extensive training, resulting in a heavy dependency on large training datasets. To overcome this limitation, this work designs a specific two-layer scattering neural network that can accurately separate signals even when the training data is limited, leading to reduced training complexity. Its performance remains robust in diverse transmission conditions. Furthermore, the scattering neural network is interpretable because features are extracted based on deterministic wavelet filters rather than training based filters.

1. Introduction

In optical communications, establishing cooperation between a transmitter and a receiver is the conventional setup to achieve reliable signal transmission. As a result, the knowledge of the signal format becomes crucial side information that must be mutually shared between both ends of the communication link. This side information is transmitted from the optical transmitter to inform the receiver about the signal format. However, this process introduces additional overhead, which consumes valuable time, frequency, or space resources.

In post-5G era, ultra-reliable low latency communications (URLLC) [1,2] plays an important role since low latency communications are needed for emerging mission-critical applications such as autonomous driving, industry 4.0, eHealth, and financial services. To reduce latency, signal controlling overhead should be as short as possible. In extreme scenarios, it is desirable to transmit signals without any overheads. Therefore, a solution that eliminates the need for transmitting side information, allowing the receiver to timely extract signal format information from received signals, becomes necessary. Signal classification is a vital technique used in various applications, including recognizing different signal standards, modulation patterns, radio frequency authentication, and radar signal identification. Traditional signal classification relies on the optimal maximum likelihood processing [3] but at the cost of high complexity. With the fast evolution of communication services, more devices are connected resulting in complex signal communication environment. In this case, mathematical based solutions are no longer efficient for signal classification.

Thanks to the advancement of artificial intelligence (AI) [4–6], intelligent solutions are being used in communications. In general, AI is divided into machine learning (ML) and deep learning (DL). Initially, due to limited computation power, ML was commonly used to solve mathematically unachievable tasks. The training process in ML is simple and explainable with well-known algorithms such as support vector machine (SVM) and k-nearest neighbors (KNN). However, these ML algorithms require efficient input features, which should be manually extracted from the training data. These features can range from simple statistical measures like mean and variance of a signal sequence, to more advanced features such as two-dimensional time–frequency grids obtained from wavelet transform [7,8]. Notice that extracting these optimal features often relies on expertise and domain knowledge. Furthermore, low-complexity ML methods struggle to handle complex problems effectively. This limitation has led to the

* Corresponding author.
E-mail address: tongyang.xu@newcastle.ac.uk (T. Xu).

https://doi.org/10.1016/j.optcom.2024.130465
Received 21 November 2023; Received in revised form 10 January 2024; Accepted 11 March 2024
Available online 13 March 2024

growing popularity of DL techniques in various advanced applications such as image processing and pattern recognition. DL offers the advantage of automatic feature extraction through sophisticated neural network architectures, such as convolutional neural networks (CNN) [3]. However, this advantage comes at the expense of large amounts of training data, making it challenging for time-variant signal communications [9–11]. Additionally, the features extracted by DL algorithms lack mathematical explainability, limiting their interpretability in certain applications [12–14].

The stability of channel conditions in fiber transmissions, as opposed to wireless transmissions, presents a unique opportunity for leveraging AI to address challenges in optical communications. AI has shown great promise in tackling physical layer and network layer-related challenges faced in optical communications such as fiber nonlinear noise mitigation and network traffic prediction [15–18]. Despite the promising potential of AI in optical communications, its practical implementation faces certain limitations. One major challenge is the need for a large amount of training data to achieve optimal performance. Collecting and processing such extensive datasets can be resource-intensive and time-consuming. Additionally, the long training time associated with AI models further impedes its widespread application. To fully exploit the capabilities of AI while facilitating dynamic and flexible optical network designs, it is essential to develop advanced algorithms that can effectively work with small training datasets.

The requirement of large training datasets results in increased training complexity and is a crucial challenge [11] for future intelligent communications. To relax the over-dependence on data for model training while ensuring high signal identification accuracy, we aim to develop low training cost AI models that can achieve high accuracy with only a small amount of training data. This research focuses on utilizing a deterministic learning framework known as wavelet scattering [19–22], which enables faster model training using a reduced number of training symbols. Unlike conventional DL approaches that involve complex architectures with multiple neural layers, we focus on a shallow neural network architecture with fewer layers. Furthermore, different to the black-box nature of DL, the wavelet scattering process offers interpretability as each layer is computed based on conventional wavelet transform principles.

The main contributions of this work are listed below:

- We propose a non-orthogonal signal wavelet scattering neural network, specifically designed to enhance the intelligence and efficiency of optical communication networks. By employing this innovative network architecture, the communication system becomes non-cooperative, allowing for improved adaptability.
- The key advantage of the proposed wavelet scattering neural network is its ability to be trained effectively with a small training dataset, leading to a wide range of applications, especially in scenarios where access to abundant training data is severely limited.
- The proposed wavelet scattering neural network facilitates the extraction of multi-layer signal scattering features. These extracted features provide valuable insights into the working mechanism of the neural network, resulting in a more interpretable and transparent design.
- An additional advantage of the proposed wavelet scattering neural network is its ability to preserve distinguishable and informative features even under the impact of strong noise and nonlinearity distortions. This robustness ensures the network's reliable and consistent performance in extreme and challenging conditions, making it a highly dependable and trustworthy solution for critical communication scenarios where noise and signal distortions are prevalent.

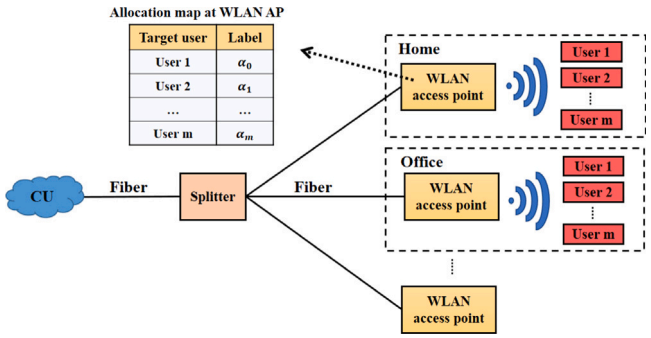

Fig. 1. Illustrative scenario for bandwidth compression labeling signal distribution after fiber transmission. The parameter α quantifies the level of bandwidth compression in received signals, enabling accurate target user classification. CU: central unit; WLAN: wireless local area network.

2. Intelligent network architecture and waveform basics

This work considers a flexible networking system empowered by intelligent algorithms. An example application scenario is illustrated in Fig. 1, where fiber cables reach the user's living or working space to provide high speed broadband service. Aligned with the growing support for fixed-mobile convergence [23], the central unit (CU) could represent an edge node of the 5G core network. Users are connected to the 5G core network via wireless local area network (WLAN) access points (AP), such that they can obtain seamless service continuity regardless of their locations or access technologies. For instance, users can seamlessly switch between 5G accesses via new radio (NR) when outdoors and WLAN indoors.

Multiple users are connected to one WLAN AP, and the number of active users are dynamic with users turning on/off their devices. In traditional systems, the received signal at the user side has to pass cyclic redundancy check (CRC) because the downlink control information (DCI) contains important information such as the UE identity and resource assignment information. To guarantee correct DCI message decoding, a number of CRC unmasking attempts and CRC checks are required. For each attempt, a user must compensate signal timing, frequency, and/or phase impairments by performing complex digital signal processing (DSP) including channel estimation, equalization, demodulation, demapping and channel decoding. These steps introduce significant processing complexity. The best-case scenario occurs when the first attempt of DCI decoding successfully passes the CRC check. Conversely, the worst-case scenario happens when none of the decoding candidates can pass the CRC check. As a result, the DCI decoding latency is a random-like process resulting in latency fluctuations. To reduce such processing latency, it is desired to first classify the incoming signals by an edge node, e.g. the WLAN AP in Fig. 1. The information used for classification is associated with user labels that are uniquely assigned to each user and known by the edge node. Furthermore, to minimize spectral efficiency loss due to overhead, this information is embedded into the signal format. Subsequently, the classified signals are forwarded to their respective users based on their assigned labels, ensuring that each user exclusively receives his intended signal. This intelligent classification and forward strategy simplifies the user side signal processing and reduces jitters. It can also be applied in private networks where network edge nodes have high freedom to design their own transmission protocols.

To realize the strategy illustrated in Fig. 1, unique signal waveform patterns have to be designed with distinguishable features. The non-orthogonal spectrally efficient frequency division multiplexing (SEFDM) signal has the bandwidth compression benefit, which is also a recognizable feature from other signals [24]. The signal waveform is flexible in spectral bandwidth and therefore the bandwidth variations can be

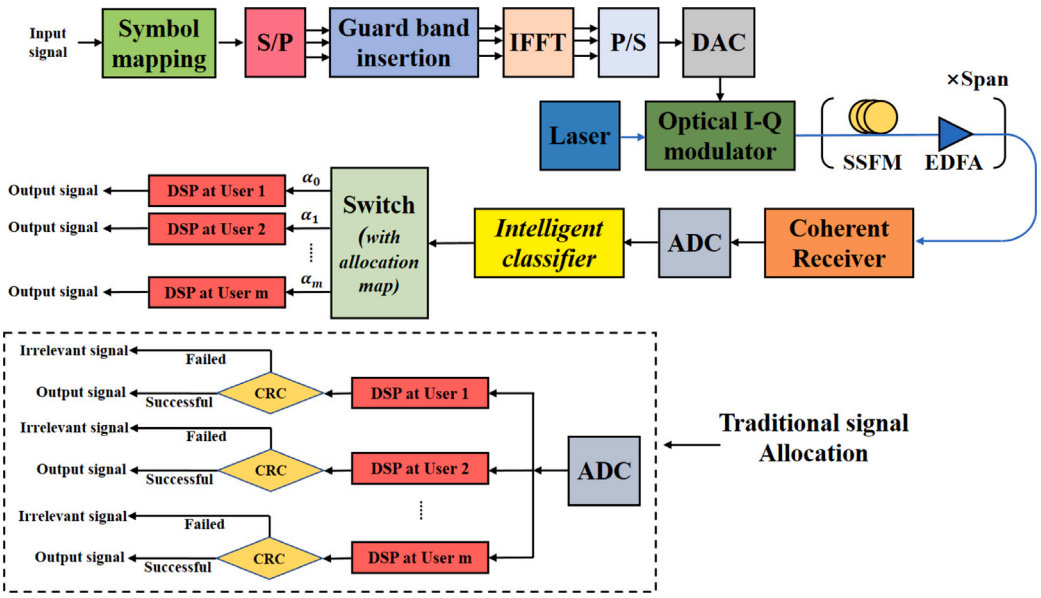

Fig. 2. Block diagram of the proposed optical system model where signals are classified by the intelligent classifier and forwarded to their target users, thereby easing user-side processing compared to the traditional signal allocation method. S/P: serial to parallel; P/S: parallel to serial; DAC: digital-to-analog converter; ADC: analog-to-digital converter; SSFM: split-step Fourier method; EDFA: erbium-doped fiber amplifier; Optical I-Q modulator: optical in-phase and quadrature modulator; DSP: digital signal processing; CRC: cyclic redundancy check.

used as the unique feature for user labeling. A discrete SEFDM signal is defined as

$$x_n = \frac{1}{\sqrt{O}} \sum_{k=0}^{O-1} s_k e^{j2\pi\alpha \frac{kn}{O}}, \tag{1}$$

for $n = 0, 1, \ldots, O-1$, where $O = \rho N$ is the product of the oversampling factor ρ and the number of subcarriers N. s_k is drawn from the data symbol vector S, defined as $S = [s_0, s_1, \ldots, s_{O-1}]$. The bandwidth compression factor (BCF) α is smaller than 1 for SEFDM, indicating that SEFDM subcarriers are non-orthogonally packed. When $\alpha = 1$, (1) describes an orthogonal frequency division multiplexing (OFDM) signal. The fractional Fourier transform due to the presence of α will cause high computational complexity, and therefore an inverse fast Fourier transform (IFFT) architecture is proposed by padding $(\lfloor O/\alpha \rfloor - O)$ zeros at the end of S, where $\lfloor \cdot \rceil$ is the nearest integer function. By doing so, the fractional Fourier transform is simplified to the conventional IFFT operation. The output of the IFFT will be truncated back to an O-length vector with the last few samples discarded.

3. System modeling setup

The application scenario depicted in Fig. 1 can be simplified to an optical system model. In this model, the signals received after long-haul fiber transmission are processed by an intelligent signal classifier and then forwarded to the target users. The setup of this system that operates in a non-cooperative manner is illustrated in Fig. 2.

The system maps the input information bits into complex symbols, which are then converted into parallel symbol vectors to enable multi-carrier transmission. Each symbol vector is oversampled before going through the IFFT operation, which is similar to adding guard band packing on both sides of each vector. After the parallel-to-serial (P/S) conversion, multicarrier OFDM and SEFDM symbols are obtained. The digital signals are then converted into analog signals using a digital-to-analog converter (DAC). To up-convert the baseband electrical signal to an optical signal of 320 GHz bandwidth at a laser central wavelength of 1550 nm, an I-Q modulator comprising two Mach–Zehnder modulators is used. To simulate a long-haul optical fiber transmission system, multiple fiber spans are connected in series, each span covering 80 km. The optical fiber channel is simulated using the nonlinear Schrödinger equation [25] with the split-step Fourier method (SSFM) and a step size of 0.05 km. Meanwhile, an erbium-doped fiber amplifier (EDFA) is applied between each fiber span to amplify distorted optical signals. The optical system model takes into account Kerr fiber non-linearities, such as self-phase modulation (SPM), cross-phase modulation (XPM), and four-wave mixing (FWM). Moreover, the fiber model is configured with a power attenuation constant of 0.2 dB/km, a nonlinear fiber parameter of 1.2 W km, and a chromatic dispersion parameter of 17 ps/(nm km). The impaired optical signals, which have undergone various distortions, are received and down-converted to electrical signals in the coherent receiver after transmission through the optical fiber. The digital signals are obtained using the analog-to-digital converter (ADC) module.

The key module in Fig. 2 is the intelligent classifier which automatically identifies the signal format based on the BCF parameter α. An allocation map is created and saved initially, which is then used for signal distribution. It is noted that the proposed intelligent classifier is applied directly after the ADC module, namely that no DSP is required at the edge node. Furthermore, end users do not need to perform multiple DSP and CRC attempts to identify target received signals. As a result, the proposed intelligent solution can simplify signal processing and enhance the effective throughput of the network. To enable accurate signal distribution, the design of the intelligent classifier is of great importance.

4. Feature extraction approaches

This section introduces three categories of traditional feature extraction approaches, and their performances will be provided in Section 7 for comparisons.

4.1. Statistical feature extraction

Arithmetic mean is a straightforward algorithm where it computes the average value of a dataset as the following

$$\mu = \frac{1}{N_d} \sum_{i=0}^{N_d-1} Y_i, \tag{2}$$

where Y_i is the ith sample of a data stream and N_d is the length of the data stream.

Variance is used to measure the variations of a dataset. A small number of standard deviation indicates that the data values are closer to the mean value while a large number of standard deviation indicates that data are spread out away from the mean value. Its calculation is expressed below

$$Var = \frac{1}{N_d}\sum_{i=0}^{N_d-1}|Y_i-\mu|^2. \tag{3}$$

Skewness [26] is a way to measure the data distribution characteristics. A negative skewness indicates that a dataset distributes more data to the left side relative to its mean value; a positive skewness indicates data is more distributed to the right side of the mean value. The computation of skewness is defined as

$$\kappa = \frac{\frac{1}{N_d}\sum_{i=0}^{N_d-1}(Y_i-\mu)^3}{(\sqrt{\frac{1}{N_d}\sum_{i=0}^{N_d-1}(Y_i-\mu)^2})^3}. \tag{4}$$

The ratio between the maximum value and the minimum value is also studied here. The MaxMin ratio tells the fluctuations of a dataset, and its definition is given by

$$P_{mm} = \frac{max(Y_i)}{min(Y_i)}. \tag{5}$$

Interquartile is a way to measure the data dispersion, which equals the difference between the 25th percentile and 75th percentile. The expression of the interquartile is given as

$$iqr = Q_3 - Q_1, \tag{6}$$

where Q_3 is the 75th percentile and Q_1 is the 25th percentile.

Each aforementioned statistical operation converts an $N_d \times 1$ vector of samples into a scalar, which can be used by SVM to separate different signal classes.

4.2. Time–frequency feature extraction

Wavelet transform is a representative time–frequency analysis method utilizing adaptive windows. Normally, a function ψ, which is the mother wavelet, is used for the windowing function. The mathematical expression of a general wavelet transform [7,8] is defined as

$$WT(a,b) = \int_{-\infty}^{\infty} f(t)\frac{1}{\sqrt{a}}\psi^*(\frac{t-b}{a})\,dt, \tag{7}$$

where a and b are the scale and translation factors, respectively, and ψ^* is the complex conjugate of ψ. A long window, which indicates a stretched wavelet and therefore a large scale a, is used for slowly changing signals at low frequency ranges. A short window, which indicates a compressed wavelet and therefore a small scale a, is used for rapidly changing signals at high frequency ranges. The adaptive wavelet functions are crucially helpful to signals with a wide spectral bandwidth since the signals containing information that cross different frequency bands. The filtering operation is illustrated in Fig. 3, where N_f waveform filters (WF) are applied on the N_d time samples to extract features at different frequencies.

Based on the configurations of the scale factor a and the translation factor b, wavelet transform can be divided into continuous wavelet transform (CWT) and discrete wavelet transform (DWT). To have a high precision time–frequency feature extraction, CWT is commonly used as defined in the following

$$CWT(2^{m/V},b) = \int_{-\infty}^{\infty} f(t)\frac{1}{\sqrt{2^{m/V}}}\psi^*(\frac{t-b}{2^{m/V}})\,dt. \tag{8}$$

Feature extraction resolution is determined by the scale of a wavelet, which is referred to as the value of a. For the fine tuned CWT, the value of a is normally defined as $2^{m/V}$, where $m = 1, 2, 3, \ldots$, and V indicates the number of voices per octave [8]. The larger number of V, the higher resolution can be achieved at feature extraction stage. Moreover, b is defined with integer values.

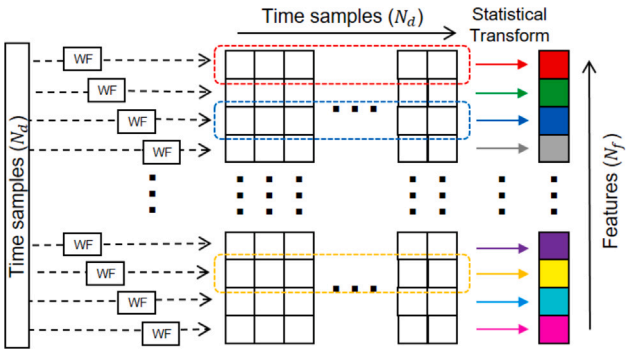

Fig. 3. One-dimensional wavelet feature generation based on wavelet filtering (WF) and statistical feature dimensionality reduction.

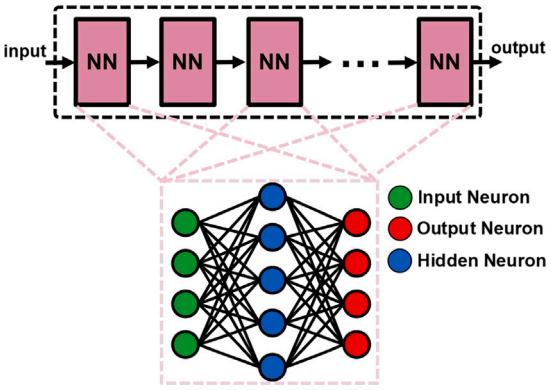

Fig. 4. CNN architecture with neural network (NN) blocks and internal neural connections.

The extracted time–frequency features from wavelet transform of a signal are two dimensional, which are complicated for the use in signal classification. Therefore, dimensionality reduction is of great importance to simplify the classification processing. The commonly used dimensionality reduction method is via statistical approaches from Section 4.1, among which variance and interquartile are selected for their satisfactory performance reported in [27]. By doing so, the original $N_d \times N_f$ feature matrix is simplified into an $1 \times N_f$ feature vector. In this paper, a multiclass error-correcting output codes (ECOC) model [28] is utilized. To separate different signal classes, an one-versus-one [29] coding strategy is implemented, simplifying the multiclass classification task into multiple binary class classification tasks. Multiple binary SVM learners, using a polynomial kernel of order two, are employed for classification.

The ML-based time–frequency wavelet transform solution would achieve faster signal processing and classification. However, one challenge of using manual feature extraction and shallow neural network is its limit in dealing with complicated tasks, in which features are not easy to be extracted. In addition, feature extraction demands specialized expertise, making it difficult for individuals without domain knowledge to determine the most efficient features for a specific task.

4.3. CNN feature extraction

DL is an advancement relative to ML thanks to the computation power advancement in recent years. In this case, a multi-layer neural connection, as depicted in Fig. 4, with huge neuron connections and backward propagation is implementable. The most popular DL

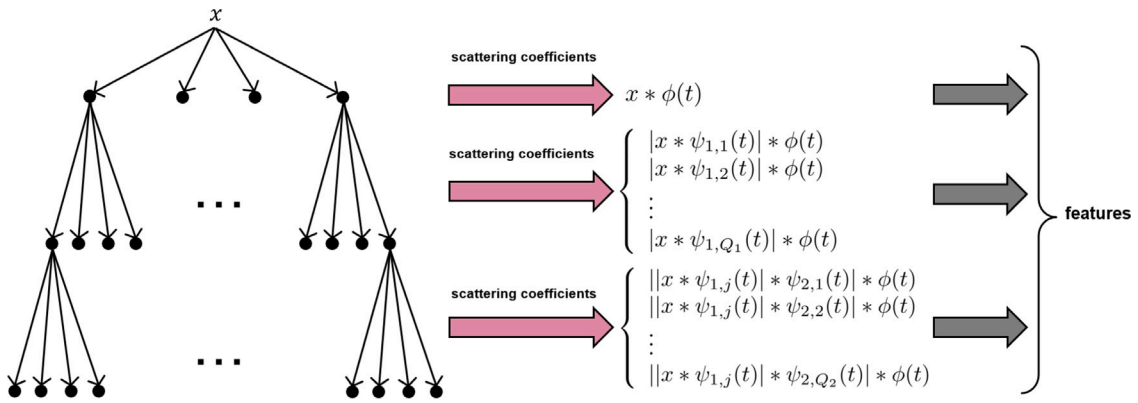

Fig. 5. Wavelet scattering neural network feature extraction.

architecture is CNN, which integrates multiple convolutional layers for automatic feature extraction. It can learn and extract hidden features from a signal without any manual tuning efforts. The mathematical explanations of CNN are as follows:

$$Y_{M-1} = \begin{cases} \sigma(Y_{M-2} * \omega^0_{M-1} + b^0_{M-1}) \\ \sigma(Y_{M-2} * \omega^1_{M-1} + b^1_{M-1}) \\ \vdots \\ \sigma(Y_{M-2} * \omega^{K-1}_{M-1} + b^{K-1}_{M-1}) \end{cases}, \tag{9}$$

where Y_{M-1} represents the feature maps after the $(M-1)$-th convolutional layer operations. For the first layer, i.e. $M = 2$, the input Y_0 is the received signal immediately after the ADC module. Considering K feature filters indicated by $(\omega^0_{M-1}, \omega^1_{M-1}, \ldots, \omega^{K-1}_{M-1})$, there will be K convolution operations performed in this layer. After the addition of the bias given by $(b^0_{M-1}, b^1_{M-1}, \ldots, b^{K-1}_{M-1})$ and the application of the activation function denoted by $\sigma(\cdot)$, Y_{M-1} includes K feature maps. This feature extraction operations are repeated until all convolutional layers are went through. The output of the final convolutional layer, containing useful hidden feature information, will be used for signal class probability computation. During the training process, a backward propagation is iteratively operated to update CNN hyperparameters until performance converges. With the automatic feature extraction capability, CNN can perform better than ML-based approaches.

5. Interpretable wavelet scattering feature extraction

An evolution of the single-layer wavelet transform is its multi-layer wavelet scattering framework [19]. The scattering framework aims to extract stable features that are insensitive to translations and deformations of input signals. This multi-layer scattering framework is particularly useful to SEFDM signal classification since the key feature we are interested in is spectral bandwidth, which is reflected by signal spectral edge. The internal signal variations are not helpful features to improve signal classification accuracy and thus can be ignored.

As mentioned in [19], traditional Fourier transform is not stable to signal deformation since high frequency components are greatly distorted by small deformations. Therefore, work in [19] proposed to use frequency averaging to remove the deformation instability. Since frequency-domain averaging is equivalent to time-domain averaging on a filter. The first step of a wavelet scattering operation is equivalent to a time-average operation, defined as

$$Ws_0 = x * \phi(t), \tag{10}$$

where $\phi(t)$ is a lowpass filter defined by the translation invariance time period T. By doing so, all high frequency components are removed and the output Ws_0 is locally invariant within T.

The use of the lowpass filter $\phi(t)$ causes information loss since high frequency components are filtered out. Therefore, a set of wavelets will be used on the original input signal x to recover high frequency components with a set of wavelet modulus transforms as

$$Wm_0 = \begin{cases} |x * \psi_{1,1}(t)| \\ |x * \psi_{1,2}(t)| \\ \vdots \\ |x * \psi_{1,Q_1}(t)| \end{cases}, \tag{11}$$

where $\psi_{i,j}(t)$ indicates a wavelet at the ith scattering layer with the jth wavelet resolution. The equation array in (11) shows operations at the first scattering layer, namely $i = 1$. The number of wavelets at this scattering layer is decided by the parameter Q_1, which determines wavelet transform resolution and is named the number of voices per octave. As investigated by [19], the commonly used value is $Q_1 = 8$. The first-order wavelet scattering coefficients, after averaging operations, is expressed as

$$Ws_1 = \begin{cases} |x * \psi_{1,1}(t)| * \phi(t) \\ |x * \psi_{1,2}(t)| * \phi(t) \\ \vdots \\ |x * \psi_{1,Q_1}(t)| * \phi(t) \end{cases}. \tag{12}$$

Repeating the similar wavelet convolution and modulus averaging operations, the second-order wavelet scattering coefficients are computed as

$$Ws_2 = \begin{cases} \| x * \psi_{1,j}(t)| * \psi_{2,1}(t)| * \phi(t) \\ \| x * \psi_{1,j}(t)| * \psi_{2,2}(t)| * \phi(t) \\ \vdots \\ \| x * \psi_{1,j}(t)| * \psi_{2,Q_2}(t)| * \phi(t) \end{cases}. \tag{13}$$

Since the second layer wavelet will convolve with all the wavelet modulus coefficients from the first layer, the number of operations will be expanded at the second layer. In (13), the variable $j \in [1, Q_1]$. For each specific value of j, a number of Q_2 wavelet convolutions are needed. The value of Q_2 is typically smaller than Q_1 in order to get a sparse representation. The wavelet convolution and modulus averaging will continue to the next wavelet scattering layer until the performance converges. The number of wavelet scattering layer depends on applications and [19] indicates a two-layer scattering architecture will be sufficient to many applications.

The scattering architecture of wavelet convolution, modulus computation, and averaging operation is regarded as a CNN-like neural network. The benefit of the scattering network is that all the filters are pre-defined by wavelets and there is no need to learn from training. This is beneficial to applications with limited training data. In addition, the feature extraction efficiency is partially decided by the value of wavelet resolution Q. A high value of Q will lead to features with high frequency resolution but at the cost of increased computational complexity. The discovery from [19] shows that an optimal value of Q is between 1 and 8.

6. System setup and model training

This work considers four QPSK-modulated signal classes: OFDM signals and SEFDM signals with $\alpha = 0.9, 0.8, 0.7$. The number of subcarriers is set to $N = 256$ for all signals and the oversampling factor is $\rho = 2$ to ensure sufficient signal resolution. Considering robustness against imperfect timing conditions, the training dataset randomly truncates 256 samples from the original 512 samples per multicarrier symbol, leading to a reduced input size of $N_d = 256$. To make the model robust especially in different noise power conditions, the launch power is set from −20 dBm to 24 dBm, with a 4 dB incremental step. Symbols under different levels of launch power are mixed for training. It is noted that our previous work [10] has comprehensively studied the impact of different fiber lengths. Therefore, this work focuses on one scenario with 30 fiber spans, resulting in a total effective fiber length of 2400 km. The impact of different training data sizes is examined by varying the number of training symbols per signal class (SPSC), indicated by SPSC = 1000, 100, 10, 5, 3, 2. For the evaluation of classification performance, 1000 testing symbols per signal class are generated independently from the training symbols for testing.

Statistical features are always the simplest solutions and should be considered at the beginning. With statistical computations, each multicarrier symbol results in a scalar value as its feature following the operations outlined in Section 4.1.

For time–frequency analysis, the wavelet transform is employed to generate a two-dimensional time–frequency feature grid. Using the Morse wavelet with a scale range of 7 octaves and $V = 10$ voices per octave, a total of $N_f = 140$ frequency scales are obtained, considering both the real and imaginary parts of the signal. Consequently, the CWT produces a two-dimensional 256×140 time–frequency analysis matrix. The feature matrix is then simplified into a 1×140 feature vector by computing its variance and interquartile.

Regarding the CNN classification model, the architecture in Fig. 4 is used. It comprises seven NN blocks, each consisting of convolutional, normalization, ReLU activation, and MaxPool/AveragePool layers. With each NN block defining $K = 64$ feature filters, 64 independent feature maps are generated. The dimension for the first block is $2 \times 256 \times 64$, taking into account the real and imaginary parts of the input QPSK-modulated symbols. After the convolutional layer, the output is normalized, passed through the ReLU activation function, and then downsampled using the MaxPool layer, simplifying the dimension to $2 \times 4 \times 64$ in the last NN block. It is noted that the last NN block uses an AveragePool layer instead of the MaxPool layer to obtain smooth features rather than extreme features. A fully connected layer and a SoftMax layer are packed at the end to classify the signal. The stochastic gradient descent with momentum (SGDM) optimizer is utilized to minimize the cross-entropy loss between the predicted and true signal classes. The optimal CNN classifier is obtained after a predefined number of training iterations through backpropagation operations. The training process consists of 30 epochs, and a mini-batch size of 128 is employed. Moreover, a learning rate of 0.02 is chosen to achieve slow but highly accurate learning.

In terms of wavelet scattering, this work explores a two-layer scattering network. The first scattering layer comprises $Q_1 = 8$ Morlet wavelet filters per octave, providing high-frequency resolution. The second layer, with $Q_2 = 1$, is used for reduced computing complexity. This design aligns with the findings in [19]. To integrate the wavelet scattering network in our system, we configure an invariance scale of 2.5 ms and a sampling frequency of 200 kHz to make the neural network robust under signal variations. For each training symbol, the input consisting of $N_d = 256$ complex-valued samples is initially decomposed into its real and imaginary parts. The resulting 512 real-valued samples are fed into the aforementioned two-layer scattering network to derive scattering coefficients. Subsequently, an ECOC classifier is implemented using binary SVM models with a polynomial kernel of order two and the one-versus-one coding design. This classifier is trained on the scattering coefficients and used for the classification of four non-orthogonal signal classes.

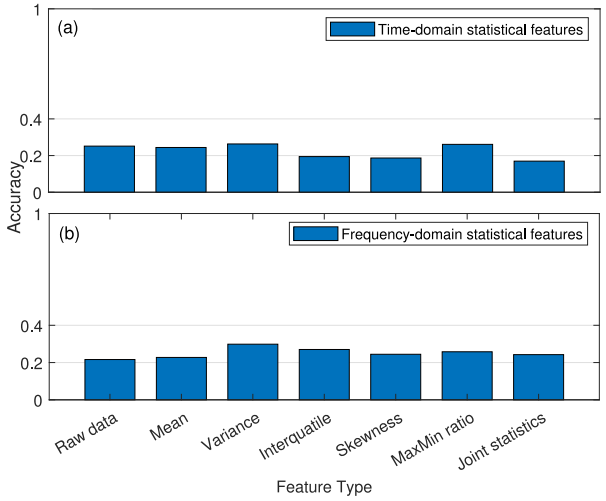

Fig. 6. Classification accuracy of statistical feature-based SVM with SPSC = 1000 and 4 dBm launch power using (a) time-domain and (b) frequency-domain statistical features.

7. Result comparisons and mechanism illustration

This section begins by examining the effects of one dimensional statistical features and their combinations, both in the time-domain and frequency-domain. To explore the capability of using a small amount of training dataset, the second part investigates wavelet scattering models using two-dimensional time–frequency features. The performances of traditional wavelet transform-based and CNN signal classifiers are presented as benchmarks.

In the beginning, to obtain convincing results, SPSC = 1000 symbols are generated, leading to a total of 4000 training symbols with four signal classes. In Fig. 6(a), five time-domain statistical features are extracted from the training dataset. Joint statistics are investigated by combining each statistical feature. In addition, the raw data without any feature extractions is evaluated. The results in Fig. 6(a) demonstrate that all the statistical features, including the joint feature, fail to facilitate proper signal classification using SVM. The accuracy remains limited due to the strong similarity in signal features among different classes. Similar outcomes are obtained with the frequency-domain dataset illustrated by Fig. 6(b), where Fourier transform was applied to the raw data. Therefore, it is inferred that simple statistical features are insufficient in assisting SVM to accurately classify signals, necessitating the use of more sophisticated feature extraction algorithms to achieve better performance.

Beyond the simple statistical feature extractions discussed above, the performance comparison among wavelet transform, CNN, and wavelet scattering classifiers is presented in Fig. 7. With a sufficient number of training symbols given by SPSC = 1000 in Fig. 7(a), three signal classifiers show similar classification accuracy rates. Achieving a highly accurate classifier normally requires a large dataset, which complicates the training process. To address this, a time/power-efficient solution is sought in reducing the number of training symbols while maintaining classification quality. In Fig. 7(b), the number of training symbols per signal class is reduced to SPSC = 100. Both wavelet scattering and wavelet transform enabled classifiers demonstrate high accuracy in identifying signals when the launch power reaches a sufficient level. However, the CNN classifier experiences a drop in classification accuracy even at high launch power due to the reduced training dataset. As the dataset is further reduced, significant changes in classification accuracy occur, as seen in Fig. 7(c) to Fig. 7(f). The traditional CNN classifier cannot work at all, and the previously functional wavelet transform classifier fails to provide competing accuracy results, especially at low launch power. Conversely, the proposed wavelet

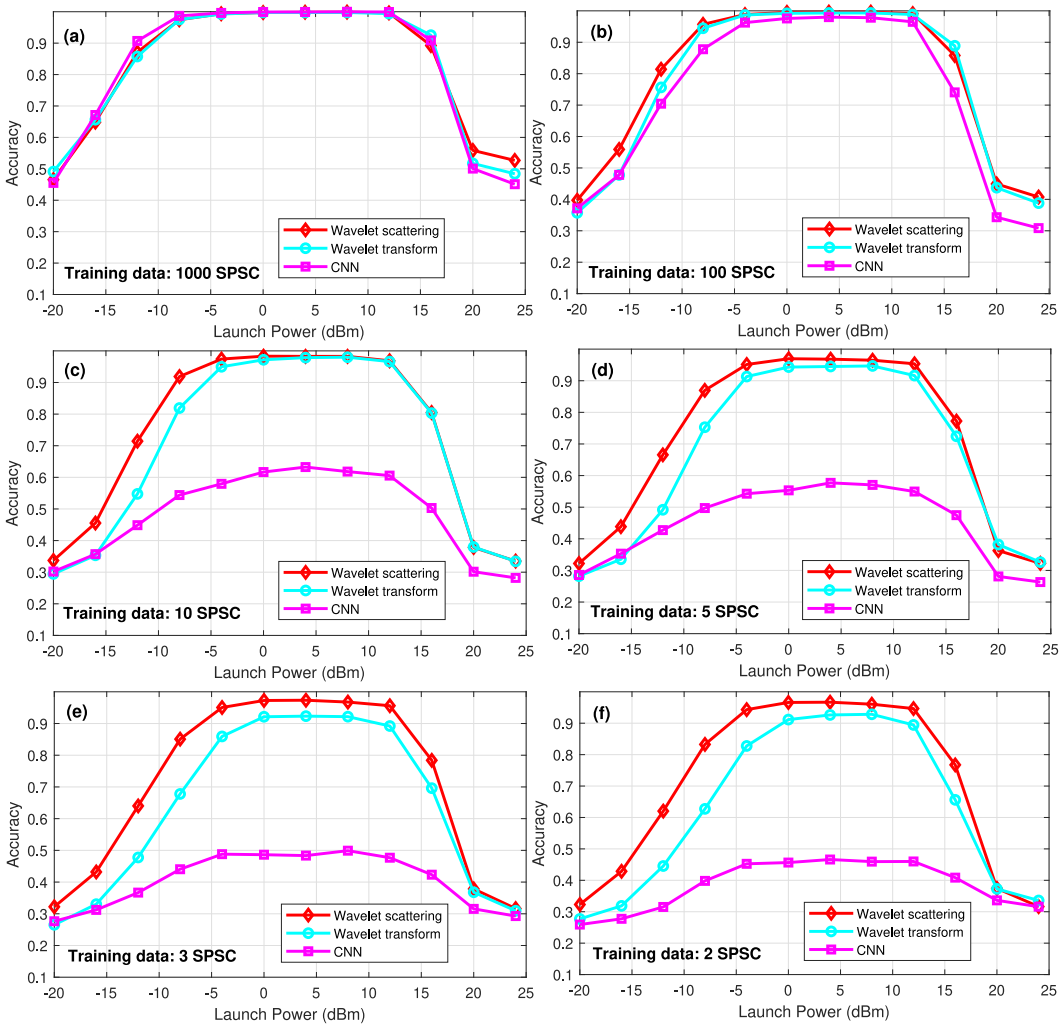

Fig. 7. Classification accuracy of three signal classifiers when different training symbols per signal class (SPSC) values are considered.

scattering classifier maintains a consistently high level of classification accuracy even when provided with only two training symbols per signal class, as demonstrated in Fig. 7(f).

The results presented in Fig. 7 reveal that a small number of training symbols are insufficient to train a reliable CNN classifier. CNN model building relies on the internal neural connection topology, the number of layers, and the weights, all of which must be learned and fine-tuned through iterative training. Therefore, having an adequate number of training symbols is crucial to ensure accurate CNN classifier training. On the other hand, the wavelet transform network has a fixed internal architecture once the wavelet properties and wavelet filters are determined. Therefore, the internal structure of a wavelet transform network is not dependent on training. This characteristic explains why a robust wavelet transform classifier can be trained effectively with a small number of training symbols. Similarly, the wavelet scattering network shares this advantage with the wavelet transform network, as its internal architecture becomes deterministic once the wavelet properties are established. Additionally, the wavelet scattering network offers the further advantage of multiple scattering layers, which enhance the quality of feature extraction. This is why a robust wavelet scattering classifier can be successfully trained even with as few as SPSC = 2 training symbols, whereas the wavelet transform network exhibits reduced classification accuracy under the same training conditions.

For a comprehensive analysis of the results presented in Fig. 7, we further investigate the classification accuracy for each signal class, as depicted in Fig. 8. When provided with sufficient training data (SPSC = 1000), all signal classes are successfully distinguished by the three signal classifiers within the optimal launch power range. Notably, the two edge signal classes, namely OFDM and SEFDM with $\alpha = 0.7$, demonstrate higher accuracy rates, as they are limited to one-side misclassification of signal classes. In the training data limited scenario, a small training dataset of SPSC = 2 fails to train accurate CNN signal classifiers because the accuracy rates for the four signal classes are all at low values in Fig. 8(f). Particularly, the classification accuracy for the signal with $\alpha = 0.9$ falls below 30%, explaining the low average accuracy rates for CNN. CNN's reliance on learning-based feature extraction leads to ineffective feature learning from a small training dataset, resulting in inaccurate classification for all signal classes. The wavelet transform-based classifier's performance also experiences a decline, although its impact is relatively limited. Fig. 8(d) shows that the OFDM signal class is more sensitive to a small training dataset, with its accuracy rate dropping below 85%, leading to a minor decrease in average accuracy. It is noted that all curves in Fig. 8 (d) and (f) are distributed in a random format and the inclusion of them only serves the purpose of visually comparing and highlighting the compromised performance when the training data is exceptionally small. Both wavelet scattering and wavelet transform can efficiently extract signal features for signal classification. However, the multi-layer feature extraction mechanism in wavelet scattering leads to the availability of more useful features, resulting in higher accuracy rates compared to the wavelet transform approach, as illustrated in Fig. 8(b).

Fig. 9 presents an evaluation of the classifiers' performance under three conditions: high noise, strong non-linearity, and optimal launch

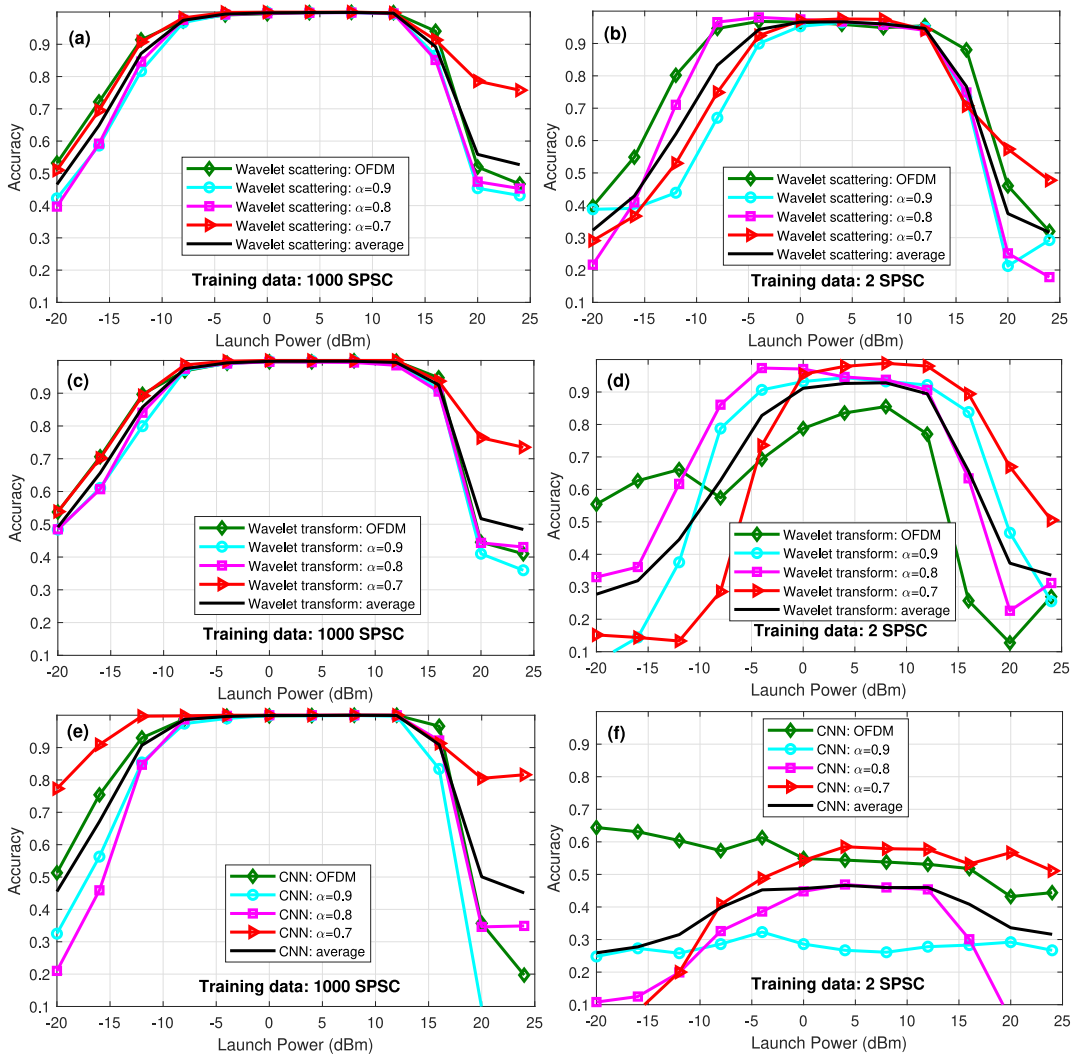

Fig. 8. Classification accuracy of three signal classifiers for individual signal classes and the average accuracy considering training symbols per signal class (SPSC)=1000 and SPSC = 2 scenarios.

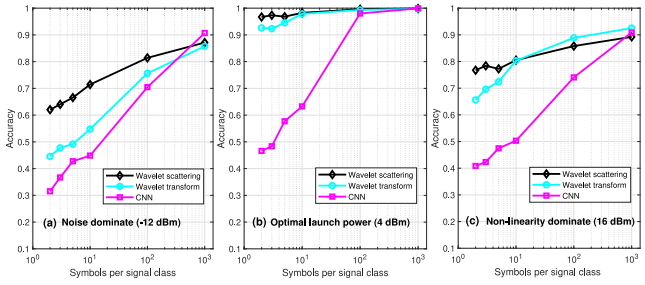

Fig. 9. Evaluation of three signal classifiers under three extreme conditions: (a) Noise dominate, (b) Optimal launch power, and (c) Non-linearity dominate. The launch power level in each condition is given in parentheses.

power. We specifically select the performance at the launch power of −12 dBm to represent the classifiers' performance in noise-dominated conditions. Results show that the proposed wavelet scattering approach exhibits the best performance under all investigated SPSC values except for SPSC = 1000. When it comes to the optimal launch power region represented by 4 dBm, where signals are in their most favorable condition with minimal noise and non-linearity effects, wavelet scattering outperforms the other two methods with consistent and high accuracy. As the launch power increases to 16 dBm, the signals experience higher non-linearity effects. In this scenario, wavelet scattering exhibits the highest accuracy when SPSC is small, and the accuracy results of all three classifiers converge with the increase in SPSC. It is noted that Fig. 9(a) and (c) include extreme channel conditions where noise and non-linearity dominate separately. With a small training dataset, the wavelet scattering network still outperforms others. As the training dataset increases to SPSC = 1000, CNN starts to show minor performance gain. This is because CNN has automatic feature extraction capability, allowing it to identify the optimal features under high noise and non-linearity effects while the other two networks are not able to achieve this. For the wavelet transform method, when non-linearity dominates, its performance surpasses both wavelet scattering and CNN at high SPSC. This suggests that a one-layer network may be more suitable under the impact of non-linearity. The wavelet scattering-based signal classifier is robust against high noise power and non-linearity effects. Its performance gain is particularly evident when SPSC is small, as shown in Fig. 10(b). On the other hand, the powerful CNN classifier fails with the worst performance, as its capability cannot be fully utilized with a small training dataset.

In Fig. 11, the mechanism of wavelet scattering classifier in classifying non-orthogonal signal classes with SPSC = 2 is provided. The plot illustrates the feature coefficients obtained after the wavelet scattering operations shown in Fig. 5. At the optimal launch power, as shown

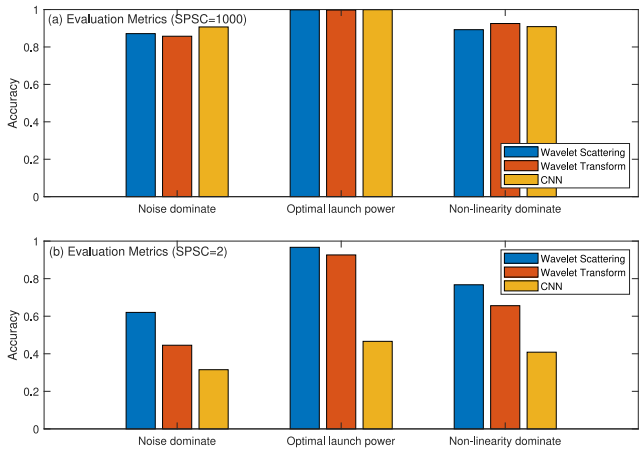

Fig. 10. Evaluation of three signal classifiers under two training conditions: (a) Sufficient training dataset with SPSC = 1000, and (b) Insufficient training dataset with SPSC = 2.

in Fig. 11(a), distinct and separable features are observed for each signal class, especially within the index range of 60 to 75. Notably, the OFDM signal class shows the widest feature variations, while the signal with $\alpha = 0.7$ shows the narrowest variations. These feature maps reflect real-world signal spectral compression variations, where signals with $\alpha = 0.7$ occupy a narrower bandwidth. In Fig. 11(b) and (c), the feature coefficients are affected by noise and non-linearity effects, respectively. It is observed that the amplitude gap between different signal classes becomes narrower under these influences. Despite these distortions, the signal features remain distinguishable, demonstrating the robustness of the wavelet scattering classifier in non-orthogonal signal classification, even under extreme conditions. This robustness sets it apart from black-box CNN classifiers, which lack transparency regarding their underlying mechanisms. Fig. 11 not only explains the functioning of the wavelet scattering classification, but also provides insights into why such classifiers exhibit resilience in the face of high noise power and strong non-linearity effects. Unlike the tunable CNN architecture with variable filters, the wavelet scattering network has a fixed internal architecture once the wavelet properties and wavelet filters are determined. In addition, the process of extracting features from the training data is mathematically transparent and easily comprehensible, as defined by (10)–(13). The transparency and interpretability of the wavelet scattering neural networks are critical for understanding and trusting the models' decisions, making them a valuable tool for intelligent signal identification.

8. Conclusion

This work proposed a wavelet scattering neural network that is capable of effectively classify non-orthogonal signal waveforms. By utilizing the level of spectral bandwidth compression for user labeling, an intelligent signal distribution framework was designed where signals can be classified and forwarded to their target receivers. This innovation eliminates the need for signaling control overheads and simplifies user-side signal processing. To comprehensively evaluate the neural network's performance, a wide range of training data sizes were tested, revealing its robustness and reliability even with extremely small training datasets. In contrast, the traditional CNN classifier heavily relying on iterative training fails to achieve satisfactory accuracy when confronted with limited training data. The proposed wavelet scattering neural network's inner workings were further elucidated by investigating the extracted features. This transparency and interpretability provide valuable insights into the classifier's decision-making process. Moreover, the classifier's performance under different transmission

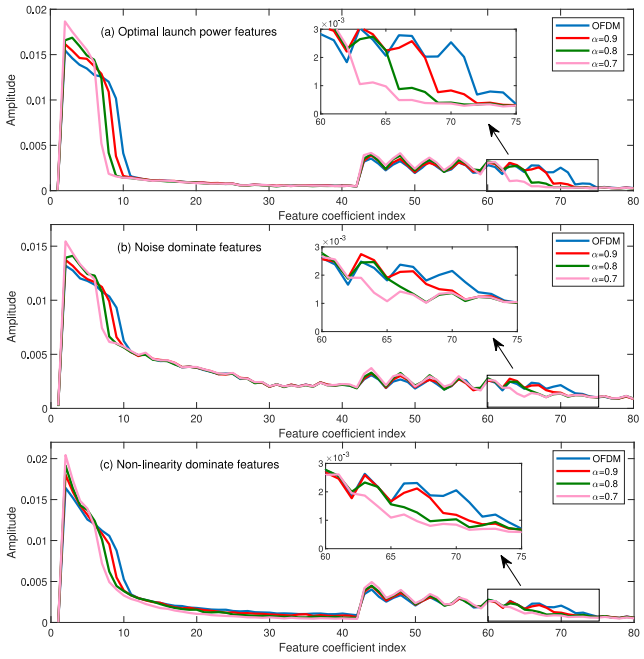

Fig. 11. Signal feature illustration to explain the mechanism of wavelet scattering neural networks in non-orthogonal signal classification. (a) Features at optimal launch power. (b) Features when noise dominates. (c) Features when non-linearity dominates.

conditions was studied, demonstrating its robustness against strong noise and non-linearity effects. This resilience makes it a suitable choice for real-world applications where signal conditions are challenging.

Funding

This work was supported by UK EPSRC (EP/Y000315/1), EU Horizon 2020 MSCA Grant 101008280 (DIOR), and UK Royal Society Grant (IES/R3/223068).

CRediT authorship contribution statement

Yinglin Chen: Writing – review & editing, Writing – original draft, Visualization, Validation, Methodology, Formal analysis, Conceptualization. **Tianhua Xu:** Writing – review & editing. **Tongyang Xu:** Writing – review & editing, Writing – original draft, Visualization, Validation, Supervision, Software, Resources, Project administration, Methodology, Investigation, Funding acquisition, Formal analysis, Data curation, Conceptualization.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

No data was used for the research described in the article.

References

[1] A. Aijaz, Private 5G: The future of industrial wireless, 2020, arXiv:2006.01820.
[2] C. She, C. Sun, Z. Gu, Y. Li, C. Yang, H.V. Poor, B. Vucetic, A tutorial on ultrareliable and low-latency communications in 6G: Integrating domain knowledge into deep learning, Proc. IEEE 109 (3) (2021) 204–246.

[3] F. Hameed, O.A. Dobre, D.C. Popescu, On the likelihood-based approach to modulation classification, IEEE Trans. Wireless Commun. 8 (12) (2009) 5884–5892.
[4] C. Jiang, H. Zhang, Y. Ren, Z. Han, K.C. Chen, L. Hanzo, Machine learning paradigms for next-generation wireless networks, IEEE Wirel. Commun. 24 (2) (2017) 98–105.
[5] J. Kaur, M.A. Khan, M. Iftikhar, M. Imran, Q. Emad Ul Haq, Machine learning techniques for 5G and beyond, IEEE Access 9 (2021) 23472–23488, http://dx.doi.org/10.1109/ACCESS.2021.3051557.
[6] T. O'Shea, J. Hoydis, An introduction to deep learning for the physical layer, IEEE Trans. Cognit. Commun. Netw. 3 (4) (2017) 563–575.
[7] I. Daubechies, The wavelet transform, time-frequency localization and signal analysis, IEEE Trans. Inform. Theory 36 (5) (1990) 961–1005.
[8] S. Mallat, A Wavelet Tour of Signal Processing, Third Edition: The Sparse Way, third ed., Academic Press Inc., Orlando, FL, USA, 2008.
[9] H. Wu, X. Li, Y. Deng, Deep learning-driven wireless communication for edge-cloud computing: opportunities and challenges, J. Cloud Comput. 9 (1) (2020) 21, http://dx.doi.org/10.1186/s13677-020-00168-9.
[10] T. Xu, T. Xu, I. Darwazeh, Deep intelligent spectral labelling and receiver signal distribution for optical links, Opt. Express 29 (24) (2021) 39611–39632.
[11] W. Tong, G.Y. Li, Nine challenges in artificial intelligence and wireless communications for 6G, IEEE Wirel. Commun. 29 (4) (2022) 140–145, http://dx.doi.org/10.1109/MWC.006.2100543.
[12] L. Dai, R. Jiao, F. Adachi, H.V. Poor, L. Hanzo, Deep learning for wireless communications: An emerging interdisciplinary paradigm, IEEE Wirel. Commun. 27 (4) (2020) 133–139, http://dx.doi.org/10.1109/MWC.001.1900491.
[13] G. Villa, C. Tipantuña, D.S. Guamán, G.V. Arévalo, B. Arguero, Machine learning techniques in optical networks: A systematic mapping study, IEEE Access 11 (2023) 98714–98750, http://dx.doi.org/10.1109/ACCESS.2023.3312387.
[14] S. Chakraborty, R. Tomsett, R. Raghavendra, D. Harborne, M. Alzantot, F. Cerutti, M. Srivastava, A. Preece, S. Julier, R.M. Rao, T.D. Kelley, D. Braines, M. Sensoy, C.J. Willis, P. Gurram, Interpretability of deep learning models: A survey of results, in: 2017 IEEE SmartWorld, Ubiquitous Intelligence & Computing, Advanced & Trusted Computed, Scalable Computing & Communications, Cloud & Big Data Computing, Internet of People and Smart City Innovation (SmartWorld/SCALCOM/UIC/ATC/CBDCom/IOP/SCI), 2017, pp. 1–6, http://dx.doi.org/10.1109/UIC-ATC.2017.8397411.
[15] J. Li, D. Wang, S. Li, M. Zhang, C. Song, X. Chen, Deep learning based adaptive sequential data augmentation technique for the optical network traffic synthesis, Opt. Express 27 (13) (2019) 18831–18847, http://dx.doi.org/10.1364/OE.27.018831.
[16] F. Musumeci, C. Rottondi, A. Nag, I. Macaluso, D. Zibar, M. Ruffini, M. Tornatore, An overview on application of machine learning techniques in optical networks, IEEE Commun. Surv. Tutor. 21 (2) (2019) 1383–1408, http://dx.doi.org/10.1109/COMST.2018.2880039.
[17] M. Schädler, G. Böcherer, S. Pachnicke, Soft-demapping for short reach optical communication: A comparison of deep neural networks and Volterra series, J. Lightwave Technol. 39 (10) (2021) 3095–3105, http://dx.doi.org/10.1109/JLT.2021.3056869.
[18] O. Sidelnikov, A. Redyuk, S. Sygletos, M. Fedoruk, S. Turitsyn, Advanced convolutional neural networks for nonlinearity mitigation in long-haul WDM transmission systems, J. Lightwave Technol. 39 (8) (2021) 2397–2406, http://dx.doi.org/10.1109/JLT.2021.3051609.
[19] J. Andén, S. Mallat, Deep scattering spectrum, IEEE Trans. Signal Process. 62 (16) (2014) 4114–4128.
[20] J. Bruna, S. Mallat, Invariant scattering convolution networks, IEEE Trans. Pattern Anal. Mach. Intell. 35 (8) (2013) 1872–1886, http://dx.doi.org/10.1109/TPAMI.2012.230.
[21] S. Mallat, Group invariant scattering, Comm. Pure Appl. Math. 65 (10) (2012) 1331–1398.
[22] E. Oyallon, E. Belilovsky, S. Zagoruyko, Scaling the scattering transform: Deep hybrid networks, in: 2017 IEEE International Conference on Computer Vision, ICCV, 2017, pp. 5619–5628, http://dx.doi.org/10.1109/ICCV.2017.599.
[23] C. Behrens, S. Krauß, E. Weis, D. Breuer, Technologies for convergence of fixed and mobile access: An operator's perspective [invited], J. Opt. Commun. Netw. 10 (1) (2018) A37–A42, http://dx.doi.org/10.1364/JOCN.10.000A37.
[24] T. Xu, I. Darwazeh, Transmission experiment of bandwidth compressed carrier aggregation in a realistic fading channel, IEEE Trans. Veh. Technol. 66 (5) (2017) 4087–4097.
[25] G. Agrawal, Nonlinear Fiber Optics, fifth ed., Academic Press, 2013.
[26] D.P. Doane, L.E. Seward, Measuring skewness: A forgotten statistic? J. Stat. Educ. 19 (2) (2011) 1–18, http://dx.doi.org/10.1080/10691898.2011.11889611.
[27] T. Xu, I. Darwazeh, Wavelet classification for non-cooperative non-orthogonal signal communications, in: 2020 IEEE Globecom Workshops (GC Wkshps), 2020, pp. 1–6, http://dx.doi.org/10.1109/GCWkshps50303.2020.9367556.
[28] T.G. Dietterich, G. Bakiri, Solving multiclass learning problems via error-correcting output codes, J. Artificial Intelligence Res. 2 (1995) 263–286.
[29] A. Rocha, S.K. Goldenstein, Multiclass from binary: Expanding one-versus-all, one-versus-one and ECOC-based approaches, IEEE Trans. Neural Netw. Learn. Syst. 25 (2) (2014) 289–302.