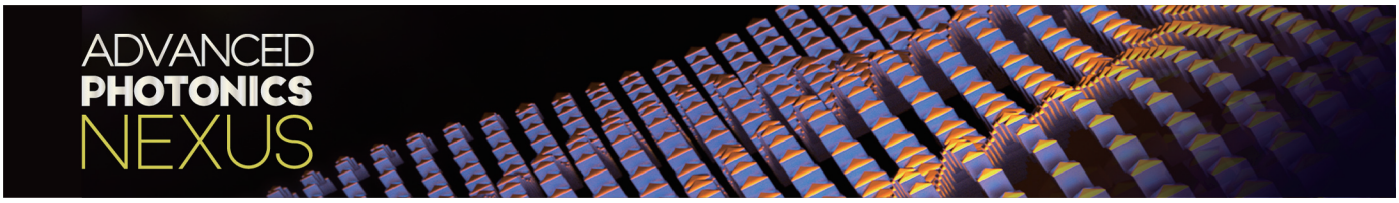
Research Article



Multiparameter performance monitoring of pulse amplitude modulation channels using convolutional neural networks

Si-Ao Li,[a] Yuanpeng Liu,[a] Yiwen Zhang,[a] Wenqian Zhao,[a] Tongying Shi,[a] Xiao Han,[b] Ivan B. Djordjevic,[b] Changjing Bao,[c] Zhongqi Pan,[d] and Yang Yue[e,*]

[a]Nankai University, Institute of Modern Optics, Tianjin, China
[b]University of Arizona, Department of Electrical and Computer Engineering, Tucson, Arizona, United States
[c]University of Southern California, Department of Electrical Engineering, Los Angeles, California, United States
[d]University of Louisiana at Lafayette, Department of Electrical and Computer Engineering, Lafayette, Louisiana, United States
[e]Xi'an Jiaotong University, School of Information and Communications Engineering, Xi'an, China

Abstract. A designed visual geometry group (VGG)-based convolutional neural network (CNN) model with small computational cost and high accuracy is utilized to monitor pulse amplitude modulation-based intensity modulation and direct detection channel performance using eye diagram measurements. Experimental results show that the proposed technique can achieve a high accuracy in jointly monitoring modulation format, probabilistic shaping, roll-off factor, baud rate, optical signal-to-noise ratio, and chromatic dispersion. The designed VGG-based CNN model outperforms the other four traditional machine-learning methods in different scenarios. Furthermore, the multitask learning model combined with MobileNet CNN is designed to improve the flexibility of the network. Compared with the designed VGG-based CNN, the MobileNet-based MTL does not need to train all the classes, and it can simultaneously monitor single parameter or multiple parameters without sacrificing accuracy, indicating great potential in various monitoring scenarios.

Keywords: pulse amplitude modulation; optical performance monitoring; intensity modulation; optical fiber communication; neural network applications.

Received Nov. 21, 2023; revised manuscript received Jan. 20, 2024; accepted for publication Feb. 19, 2024; published online Mar. 14, 2024.

[DOI: 10.1117/1.APN.3.2.026009]

1 Introduction

To meet ever-growing demands for high-capacity optical communication, data centers (DCs) have gradually become the key technology for a myriad of network applications.[1–3] Compared to coherent optical systems, intensity modulation and direct detection (IMDD) of multilevel pulse amplitude modulation (PAM) formats utilize an architecture that is more power-efficient and easy to implement, providing a suitable choice for 100G, 400G intra- and inter-DC networks.[4–8] Currently, various pluggable transceivers based on IMDD and multilevel PAM have been deployed in intra-DC short reach (SR)/long reach (LR) applications.[9,10] In 2020, the industry's first silicon photonics 100G PAM4 dense wavelength division multiplexing solution was commercially available. Using two-wavelength 25 GBd PAM4 signals, it can support 80-km transmission distance and 4 Tb/s communication speed for inter-DC networks.[11] The industry will then be likely marching toward higher level PAM signals, such as PAM-6 or even PAM-8.[12–15] However, as it is intensity-only modulation, an IMDD system has relatively low spectral efficiency (SE), making it difficult to further improve the capacity within limited bandwidth. Alternatively, coherent optical communication re-emerged for practical applications in the early 20th century for high SE transmission, which can provide higher sensitivity and bit rates, and may be used for intra-and inter-DC applications in the near future.[16–18]

*Address all correspondence to Yang Yue, yueyang@xjtu.edu.cn

However, compared with the IMDD, the sensitivity improvement of the coherent technique is at the cost of an additional local oscillator, more complex modulation structures, and higher bit rates at a single wavelength are attributed to more modulation dimensions.

In order to utilize bandwidth resources more efficiently, the wavelength division multiplexing (WDM) system combined with IMDD is widely used in today's DC networks to enlarge the data-carrying capacity.[19,20] Figure 1 shows the structure of both the intra- and inter-DC systems based on WDM-IMDD with eye diagram monitoring using machine learning (ML). The monitor nodes are usually implemented on the transmitter side or in front of the receiver. On the transmitting side, the nodes focus on monitoring the launch power, transmitting optical signal-to-noise ratio (OSNR), signal modulation format (MF), baud rate (BR), and other characteristics to ensure the quality of the generated signal. As for the receiver side, in addition to the relevant parameters of the received signal, the impact of various impairments in the channel has attracted more attention, including dispersion, nonlinear effects, the received OSNR affected by the noise figure of the amplifiers, etc. Eye diagrams of one wavelength channel can be collected through a tunable bandpass filter (TBPF) and digital communication analyzer (DCA). The collected images could then be uploaded to the network cloud, and then the ML methods are utilized to perform multiparameter joint monitoring based on the input images. Afterward, the software-defined networking controller can provide the feedback information to the channel under monitoring in time, so that it can adapt to the link and environmental changes for better performance.

In such a high-capacity and complex system, many parameters need to be monitored in real time. On the one hand, some channel impairments and characteristics need to be monitored to provide appropriate compensation. On the other side, the current real optical networks are increasingly dynamic and reconfigurable, so flexible signal parameter monitoring is also necessary to improve network reliability.[21,22] First of all, in order to use an adaptive MF according to the transmission conditions, a suitable MF identification is needed.[23] The Nyquist shaping with different roll-off factors (ROFs) is usually applied to alleviate the bandwidth requirement of electrical and optical components,[24,25] while probabilistic shaping (PS) distribution schemes[26–28] and forward error correction (FEC) coding with different FEC overheads (OHs) are used for enhancing the tolerance to the OSNR.[29–31] Low OSNR always limits the performance of optical links, and channel impairments, such as chromatic dispersion (CD), could severely distort the signal. Therefore, an accurate and powerful monitoring scheme that can help identify and optimize the optical network would be essential.

Deep learning (DL), as a branch of the ML methods based on deep neural networks, has attracted widespread interest over the past few years in optical performance monitoring (OPM).[32,33] With the ability of feature extraction and self-learning, convolutional neural network (CNN)-based DL can directly process images, such as eye diagrams,[34,35] constellation diagrams,[36–38] amplitude histograms,[39,40] and asynchronous delay tap plots.[41] A constellation diagram is typically obtained by coherent detection, which requires complex hardware. The hardware for acquiring the asynchronous amplitude histogram is relatively simple, and it mainly contains statistical information on the amplitude of the signal within a period of time. Therefore, it is often used for signal OSNR estimation. At the same time, because it contains fewer features of the signal, it is also used for relatively simple classification tasks, such as MF classification. In asynchronous delay tap sampling plots, there is a time delay between two electrical lines before ADC. The time delay is usually equal to 0.5 or 0.25 time periods of the symbol interval, meaning that the BR of the signal needs to be obtained in

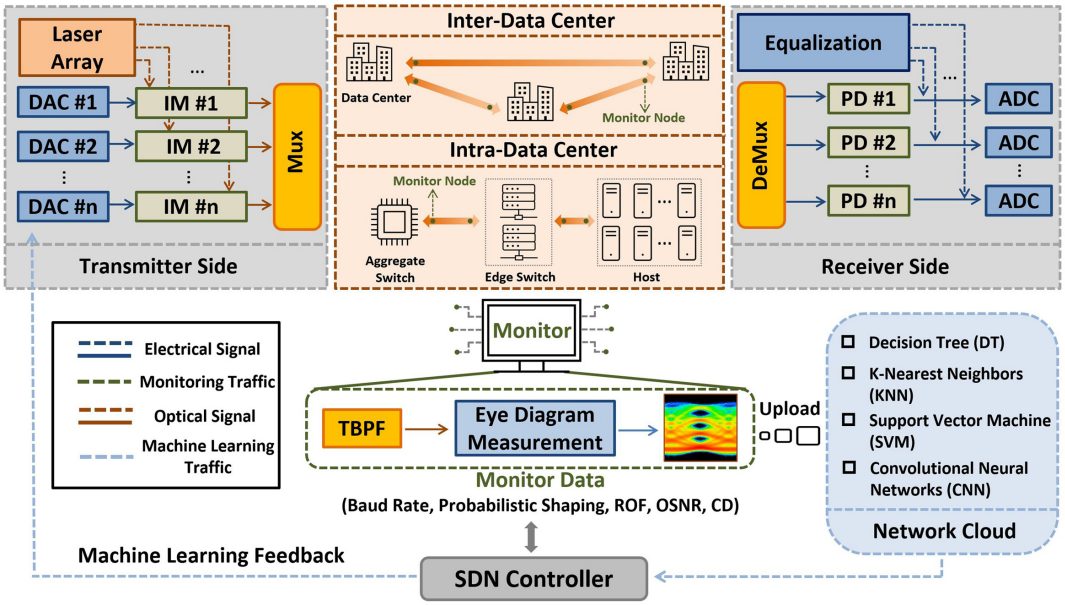


Fig. 1 Conceptual diagram of multiparameter performance monitoring of PAM signals in intra- and inter-data center systems. DAC, digital-to-analog converter; IM, intensity modulator; PD, photodiode; ADC, analog-to-digital converter; TBPF, tunable bandpass filter; SDN, software-defined networking; ROF, roll-off factor; OSNR, optical signal-to-noise ratio; CD, chromatic dispersion.

advance. By contrast, an eye diagram seems to be a suitable way, using low-speed direct detection together with clock recovery. Eye diagrams contain plentiful features of the original digital data and do not require a more complex receiving structure, such as the constellation diagram. For instance, the clock jitter, the rise and fall time, and the noise level of the optical signal could be analyzed from the eye diagram. At the same time, compared to an asynchronous amplitude histogram, an eye diagram uses one more degree of freedom (time). Therefore, it can better identify parameters in the time domain, such as BR. Generally speaking, the original eye diagram contains enough features for the following classification networks, and there is no need to perform digital signal processing (DSP) on the signal in advance to reduce the impairments. This also shortens the time to acquire an image of an eye diagram, giving it the potential for real-time monitoring.

Previous works have demonstrated that CNN-DL can perform OPM from featured images, such as transmitter and dispersion eye closure for PAM4 (TDECQ) estimation,[42] OSNR estimation,[43,44] MF recognition,[45] and bit-rate identification.[46,47] However, the above demonstrations focus more on channel impairments and MFs, and little joint monitoring of digital signal parameters. Meanwhile, most of them are aiming for coherent optical channels and can only jointly monitor a few parameters. To pursue more efficient dynamic optical IMDD-based DC networks, various digital shaping technologies, especially PS and Nyquist shaping, will gradually play important roles in industrial applications. Therefore, simultaneously monitoring both digital signal parameters and optical link parameters will be necessary. Moreover, due to the urgent demand of applications, such as emerging large-scale cloud computing and high-definition videos, passive optical networks gradually become key parts in optical transmission links for providing broadband connectivity. The protection and monitoring of optical line terminal and optical network unit are also much-needed issues.[48,49] Optical monitoring capabilities can be used to enable new ways of managing traffic. For example, routing decisions based on performance monitoring is a possibility. By monitoring the channel quality and link security and constantly updating the routing lookup table, the traffic with large capacity and priority can be dynamically adjusted to the high-performance optical channel so as to ensure that the data channel reaches an acceptable BER, and the whole network achieves sufficient transmission and protection capacity.[50]

In this paper, we demonstrate joint monitoring of a PAM-based IMDD channel for six parameters of digital signal and optical link (MF, BR, PS, ROF, OSNR, and CD) using CNN-based DL and eye diagrams on the receiver side. Here, a visual geometry group (VGG)-based CNN model with less computational cost is designed and optimized for much more efficient classification. The experimental results indicate that the designed VGG-based CNN outperforms the other three traditional ML methods, including support vector machines (SVMs), k-nearest neighbors (KNNs), decision trees (DTs), and gradient-boosted decision trees (GBDTs).[51–54] A high prediction accuracy of 97.16% is achieved for jointly monitoring up to six parameters including 3 to 8-ary PAM, 25 to 27 GBd BR, 0.8 to 1 PS coding rate, 0.1 to 0.5 ROF, 25 to 40 dB OSNR, and −120 to 0 ps/nm CD. Moreover, the VGG-based CNN shows the stability of monitoring, with the highest accuracy greater than 96% in different scenarios, providing the possibility of dynamically monitoring and optimizing channel performance.

Furthermore, the other three modern CNN networks are also compared with the designed VGG-based model, including ResNet-18, MobileNetV3, and EfficientNetV2. All of them can achieve >96% accuracy. By contrast, the proposed VGG-based model with fewer layers has smaller memory usage, and the lightweight MobileNetV3 has fewer parameters and floating-point operations per second (FLOPs) using mobile inverted bottleneck convolutional (MBConv). Both are more cost-efficient and resource-friendly for channel monitoring of IMDD-based DC networks. Meanwhile, we also noticed that EfficientNetV2 has the potential to achieve higher accuracy through an optimized combination of MBConv and Fused-MBConv, which also provides ideas for using fewer resources to achieve higher monitoring accuracy of PAM-based communication.

Finally, a multitask learning (MTL) model combined with MobileNetV3 is further designed, and the output neurons of six-parameter joint monitoring in MTL could be reduced from 1728 to 21. The MTL model can not only carry out joint monitoring of multiple parameters but also monitor each single parameter separately at the same time. Compared with the other CNN methods, the accuracy of MTL could also be achieved above 95% in various monitoring tasks without training all 1728 classes.

The rest of the paper is organized as follows: In the next section, the optical experimental setup of PAM-based transmission and the used ML algorithms are introduced. Next, the results of ML and the VGG-based CNN are compared. Furthermore, the four modern CNN models mentioned above are discussed, including their accuracy and computational resources. Furthermore, the performance of MTL combined with CNN is investigated and discussed in detail. Finally, a conclusion is made for multiparameter performance monitoring of PAM channels using CNN.

2 Experimental Setup and Methods

2.1 Optical Experimental Transmission Link

The experimental setup shown in Fig. 2(a) is used to capture the eye diagrams of the PAM-based IMDD channel with different parameters related to both the digital signal (BR, ROF, MF, and PS) and optical link (OSNR and CD). In the offline DSP, PS coding and Nyquist shaping are performed in sequence to change the number of bits per symbol and the ROF. The sampling rate of the digital-to-analog converter (DAC) used here is up to 96 GSamples/s. For the 25 to 27 GBd signals that need to be monitored in this work, it could perform upsampling by up to 3 times, corresponding to a sampling rate of 75 to 81 GSamples/s. Therefore, to better perform Nyquist shaping and consider the limitation of the DAC, a threetime upsampling is adopted for different BRs. Meanwhile, the pre-equalization based on feed-forward equalization with seven taps is utilized for various BRs, and then the predistortion is also used to compensate for the nonlinearity introduced by E/O devices, such as an intensity modulator. The processed data are then sent to a 20 GHz DAC.

As for the optical link, a flat noise is achieved from the amplified spontaneous emission (ASE) noise source filtered by TBPF1, then amplified by an erbium-doped fiber amplifier (EDFA1). The center of the noise is at the wavelength of coherent transmitter and its spectral width is 1.5 nm. The noise can be

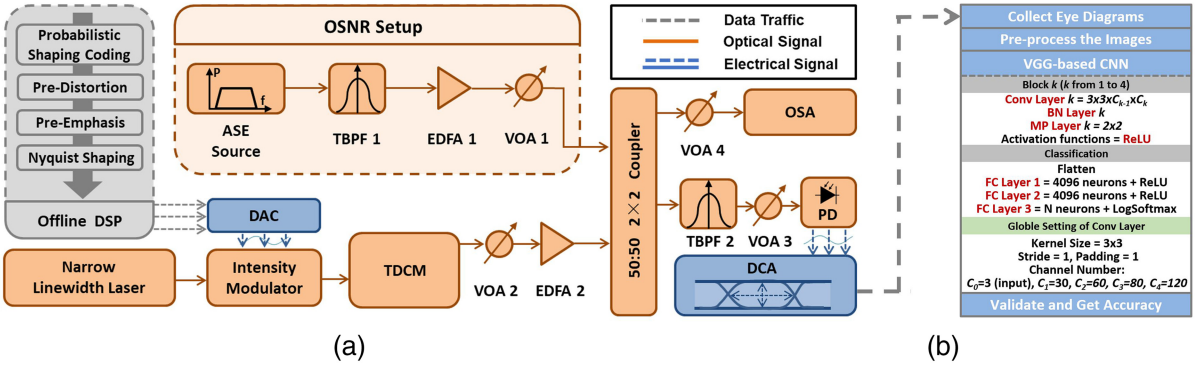


Fig. 2 (a) Experimental setup used to collect eye diagrams. ASE, amplified spontaneous emission; TBPF, tunable bandpass filter; EDFA, erbium-doped fiber amplifier; VOA, variable optical attenuator; DSP, digital signal processing; DAC, digital-to-analog converter; TDCM, tunable dispersion compensation module; PD, photodiode; OSA, optical spectrum analyzer; DCA, digital communication analyzer. (b) The structure of the VGG-based CNN model for classification. Conv, convolutional; BN, batch normalization; MP, max pooling; FC, fully connected.

adjusted through a variable optical attenuator (VOA1) to control the OSNR values in the channel. In addition, a tunable dispersion compensation module (TDCM) is used in the transmission line to emulate the accumulated CD in practice.

The outputs of EDFA2 (signal) and VOA1 (noise) are further combined through a 3-dB optical coupler. One of the outputs is used for the optical spectrum analyzer (OSA) to measure the OSNR, while the other one is first injected into TBPF2 to remove the out-of-band ASE noise. Then, the eye diagrams are captured by a 22.5-GHz DCA after controlling the received optical power at 3 dBm through VOA3. Some of the captured eye diagrams with chosen parameters are listed in Fig. 3. Here, six parameters are selected for joint monitoring (PAM order, 3 to 8; BR, 25 to 27 GBd; PS coding rate, 0.8 to 1; ROF, 0.1 to 0.5; OSNR, 25 to 40 dB; CD, −120 to 0 ps/nm). It is worth noting that there are two main reasons why the range of the negative CD value of TDCM is used. (1) The adjustable range of negative CD value is larger, which can emulate a larger dispersion accumulation. (2) Since dispersion only affects the phase of the signal and causes pulse broadening, the impact of positive or negative CD values on the eye diagram should be similar. It is equally suitable for demonstrating the potential of ML for dispersion classification. Particularly, the small monitoring range of BR takes the OH into account for various FEC coding. In the inter- and intra-DC applications, different FEC encoders are usually selected with different OHs. Therefore, to achieve a given bit rate under different OHs, slightly adjusting the BR is needed. The OHs of different FEC encoders could be further obtained from the classification of the BR. Besides, the achievable information rate (AIR) of PS follows the equation:

$$\mathrm{AIR} = N \times \left[-\sum_{i=1}^{M} \left(\frac{1}{M} \times \log_2 \frac{1}{M} \right) \right] = N \times \log_2 M, \tag{1}$$

where M is the order of PAM signals, $\log_2 M$ represents the number of bits/symbols of the uniform PAM signals, and N (0.8 to 1) is the coding rate needing to be monitored. N refers to the multiple of the AIR of the probabilistic-shaped signal relative to that of the uniform signal. In PS, the probability of amplitudes is commonly generated according to the Maxwell–Boltzmann (MB) distribution, as shown in Eq. (2). Therefore,

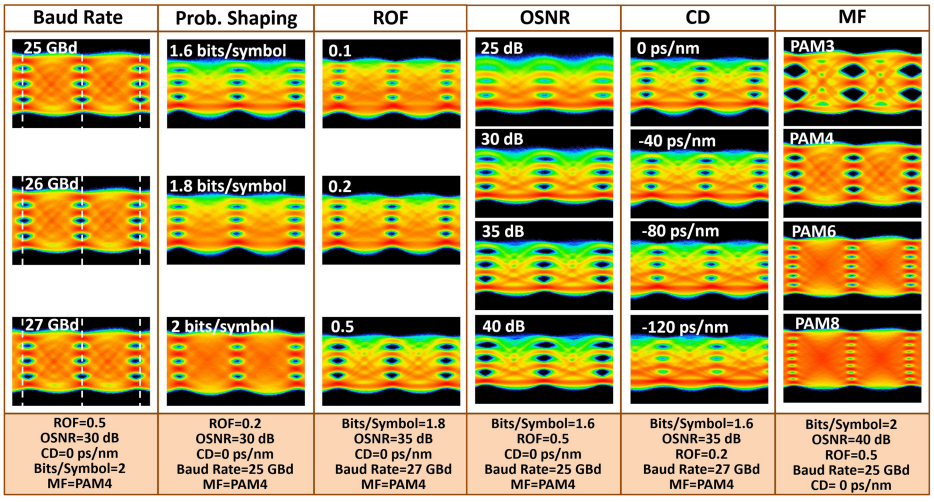


Fig. 3 Eye diagrams of PAM signals with different MFs, BRs, PS, ROFs, OSNR, and CD.

the entropy rate of the PS PAM signal could be expressed in Eq. (3). Since the entropy rate should be numerically equal to AIR, the expression for N could be further obtained as shown in Eq. (4),

$$P_X(x) = \frac{e^{-\lambda x^2}}{\sum_{x' \in X} e^{-\lambda x'^2}} \quad \{X = [\pm 1, \pm 3, \ldots, \pm(M-1)]\}, \tag{2}$$

$$H(X) = -\sum_{x \in X} P_X(x) \log_2 P_X(x) \quad \{X = [\pm 1, \pm 3, \ldots, \pm(M-1)]\}, \tag{3}$$

$$N = \frac{-\sum_{x \in X} P_X(x) \log_2 P_X(x)}{\log_2 M} \quad \{X = [\pm 1, \pm 3, \ldots, \pm(M-1)]\}. \tag{4}$$

The chosen interval of different parameters also is shown in Fig. 3: three parameters contain four values, and the other three contain three values. Meanwhile, 10 images are collected for each class so that there is a total of 1728 $(4^3 \times 3^3)$ classes and 17,280 images in the data set.

2.2 Algorithms of ML

All collected images are then divided into training (70%) and testing (30%) sets for classification based on ML including three traditional ML methods (SVM, KNN, and DT) and three modern CNN models (VGG-based, ResNet, MobileNetV3, and EfficientNetV2) for comparison.

2.2.1 Traditional ML methods

SVM can be defined as a linear classifier with the largest margin in a high-dimensional feature space, aiming at finding the hypothesis space that can correctly divide the training set.[51] It is a popular method for solving small and medium data samples and nonlinear, high-dimensional classification problems. An important property of SVM is that most of the training sets do not need to be retained, and the final model is only related to the support vector after the training is completed. As a supervised learning method, DT uses information entropy as a measure to construct a tree structure with the fastest decrease in entropy. It uses training data to establish a model based on the principle of minimizing the loss function and uses the decision model to classify new data sets.[53] The advantage of DT over the other pattern recognition techniques is the interpretability of the constructed model consisting of feature selection, DT generation, and pruning. The main idea of KNN is using principal component analysis to extract the features. For an n-dimensional input vector, it corresponds to a point in the feature space, and the output is the label corresponding to the feature vector. After inputting the unlabeled data, KNN algorithms compare it with the feature corresponding to the data in the data set, and then extract the label with the closest feature, which is called the "nearest neighbor." Finally, the class with the most occurrences among the k most similar data is selected as the classification result.[52] The GBDT model is an additive model, which trains N regression trees serially and finally adds up the results of the N regression trees, thus obtaining a strong learner. It is a tree ensemble method that builds a DT learner at a time by fitting the gradients of the residuals of the previously constructed tree learners.[54] In these traditional ML methods, the color histograms of RGB combined with a histogram of oriented gradients (HOGs) are used as the features. The color histograms mainly contain color information of the input images, while HOG contains contour information of the images. In these parameters, the choice of "pixels per cell" in HOG could greatly affect accuracy. Figure 4 lists the HOG and color histograms corresponding to different pixels per cell and MFs. The colored lines in the color histograms represent the gray-scale distribution under the corresponding red, green, and blue channels. The range of gray-scale range is 0 to 255.

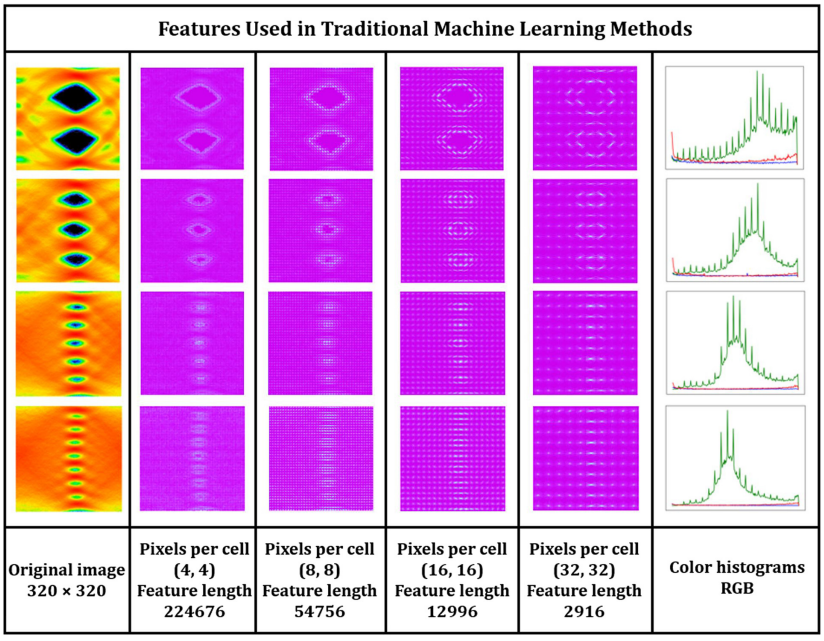


Fig. 4 Features and parameters used in traditional ML methods (KNN, SVM, DT, and GBDT).

It can be clearly seen that the contour is clearer when there are fewer pixels in each cell, but the number of features would be more, increasing the computational cost. Here, the 16 pixels × 16 pixels per cell are chosen for KNN and SVM, corresponding with 12,996 features. Therefore, by adding the number of features in color histograms (256 × 3), the total length of the features used is 13,764. As for DT and GBDT, fewer features show better accuracy in classification tasks. So, the 64 pixels × 64 pixels per cell are chosen for DT and GBDT, corresponding with 576 features, and the total length of the features used is 1344. The corresponding selected parameters are shown in Table 1.

2.2.2 Modern CNN models

Due to different network architectures and hyperparameter choices, the performance of CNN will vary greatly. After much trial-and-error and research in previous works, numerous refined modern CNN models debuted. They have different layer structures and combinations to achieve higher accuracy while also further reducing the computing resources. Here, a VGG-based model is mainly used. The commonly used ResNet-18, the lightweight MobileNetV3, and the recent EfficientNetV2 are also under consideration. In addition, before training the CNN models, the images need to be preprocessed. Data augmentation is first implemented to enlarge the data set for all the models, including the center cropping and color jitter. The image is cropped to a size of 320 × 320 × 3 first, and then a color jitter is performed. Finally, the image processed by color jitter is used together with the original cropped image as the data set. The input eye diagrams are then normalized using the mean and standard deviation to enhance the data's responsiveness to the activation function.

For a typically VGG-based block, it includes a multiple convolutional (Conv) layer, a batch normalization (BN) layer, a max pooling (MP) layer, and an activation function,[55] as listed in Fig. 5. In the proposed technique, we used a VGG-based model with four blocks, as shown in Fig. 2(b). Compared to the commonly used VGG-11 model, here we reduce the number of Conv and MP layers by half, achieving a much better trade-off between accuracy and computational cost. Moreover, the kernel size of each Conv layer is also shrunk by 2 to 4 times. Through further optimization, the number of neurons in the output layer of the traditional VGG-11 model is finally reduced from 512 to 120 in our VGG-based model, enabling more efficient training. The 3 × 3 Conv layers with different channels are chosen in each block; the details of each layer are listed in Table 2. The BN layer is used for speeding up the convergence of the network along with avoiding the vanishing gradient problem. The following 2 × 2 MP layer is utilized to half the size of images in height and width, and the rectified linear unit (ReLU) is chosen as a nonlinear activation function. The classifier is

Table 1 Parameters used in hog and color histograms.

		HOG			Color histogram			
Method	Input image size	Orientation	Pixels per cell	Cells per block	Bin	Range	Channel	Feature length
KNN, SVM	320 × 320 × 3	9	16 × 16	2 × 2	256	0 to 255	3 (RGB)	13,764
DT, GBDT	320 × 320 × 3	9	64 × 64	2 × 2	256	0 to 255	3 (RGB)	1344

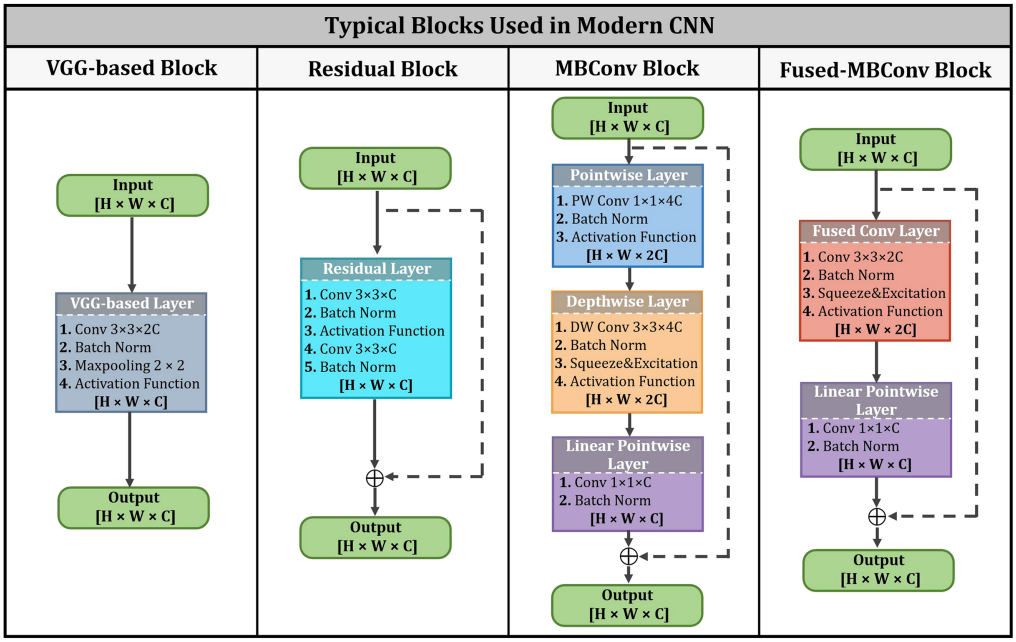


Fig. 5 Typical algorithm architectures applied in the VGG-based model, ResNet-18, MobileNetV3, and EfficientNetV2. PW, point-wise; DW, depth-wise.

Table 2 Structure of the VGG-based model.

Input size	Filter size	Layer	Output size
320 × 320 × 3	3 × 3 × 3 × 30	Conv.1	320 × 320 × 30
320 × 320 × 30	2 × 2 × 30	MP.1	160 × 160 × 30
160 × 160 × 30	3 × 3 × 30 × 60	Conv.2	160 × 160 × 60
160 × 160 × 60	2 × 2 × 30	MP.2	80 × 80 × 60
80 × 80 × 60	3 × 3 × 60 × 80	Conv.3	80 × 80 × 80
80 × 80 × 80	2 × 2 × 80	MP.3	40 × 40 × 80
40 × 40 × 80	3 × 3 × 80 × 120	Conv.4	40 × 40 × 120
40 × 40 × 120	2 × 2 × 120	MP.4	20 × 20 × 120
48,000	48,000 × 4096	FC.1	4096
4096	4096 × 4096	FC.2	4096
4096	4096 × *N*	FC.3	*N*

N, number of classes; Conv, convolutional layer; MP, max pooling layer; BN layer and ReLU in the structure do not change the size of the output; FC, fully connected.

composed of three fully connected (FC) layers, flattening, and memorizing the data. The soft-max algorithms are used in the last FC layer for classification problems. In addition, several hyperparameters are also optimized for a better performance as well as a higher training speed. Backpropagation and stochastic gradient descent are chosen as the optimizer to train the model, with a learning rate of 0.00092 and a momentum of 0.9.

With the network depth increasing, the prediction accuracy of the deeper network might get saturated and then degrade rapidly. This degradation occurs in both the training set and the test set, indicating that it is not due to overfitting. This may be caused by the gradient vanishing. Specifically, the weights of hidden layers closer to the input layer update slowly as they approach the input layer. Moreover, even when deeper networks start to converge, a degradation problem emerges. For instance, the current number of layers may be optimal for the model. Increasing the number of layers requires maintaining the same performance by ensuring that the input and output of the added layers remain unchanged, effectively implementing identity mapping. However, the model struggles to learn this identity mapping, which is a primary factor contributing to performance degradation. To overcome the degradation problem, ResNet was introduced with the basic block shown in Fig. 5. Instead of each directly fitting a desired underlying mapping, we explicitly make these layers fit a residual mapping. In general, the desired underlying mapping is denoted as $H(x)$, and x represents the input features. Mapping used here refers to a kind of function representing a processing of the input features. In ResNet, it is recast into $F(x)+x$ and makes the residual layers fit the mapping of $F(x)=H(x)-x$. It is easier to optimize the residual mapping than the original one. Here, the typical ResNet-18 with 17 Conv layers and 1 FC layer is selected for classification. The details of the network can be found in Ref. 56.

MobileNetV3 is a lightweight model with fewer parameters and computations, which is achieved by the MBConv blocks depicted in Fig. 5. An MBConv is based on separable convolution containing $(1\times 1\times M\times N)$ point-wise (PW) and $(K\times K\times 1\times N)$ depth-wise (DW) convolution. M represents the number of channels, while N and K represent the size and the number of convolution kernels, respectively. Therefore, the computation reduction of the separable convolution used in the MBConv block over the regular one can be described as

$$\frac{H\times W\times C\times N\times 1^2+H\times W\times N\times 1\times K^2}{K\times K\times C\times N\times H\times W}=\frac{1}{K^2}+\frac{1}{C}, \tag{5}$$

where the size of the input is described as (H,W,C). Meanwhile, MobileNetV3 introduces a squeeze-and-excitation (SE) layer to extract the correlation features between channels after performing DW convolution. A linear PW layer is also used at the end of the block, which differs from the regular PW layer in that it removes the activation function. There are two new MobileNet models proposed in Ref. 57, and the MobileNetV3-S is chosen here with fewer layers.

The recent EfficientNetV2 utilizes the newly emerged Fused-MBConv block to achieve more effective training. In the Fused-MBConv block, the PW and DW convolution in the original MBConv block are replaced with a regular Conv layer to fully utilize modern accelerators, as shown in Fig. 5. Meanwhile, the neural architecture search is used to search for the different combinations of MBConv and Fused-MBConv blocks to achieve the best trade-off between training speed and accuracy. Besides, EfficientNetV2 prefers smaller 3×3 kernel sizes, but that means more layers should be added to compensate for the reduced receptive field. The searched model EfficientNetV2-S is considered in the comparison here, according to Ref. 58.

3 Results

In the first instance, we compared the classification accuracies of VGG-based CNN with the ones of the other three traditional ML methods (SVM, DT, and KNN) under different parameter combinations. Furthermore, we adopted four modern CNN models (VGG-based, ResNet-18, MobileNetV3-S, and EfficientNetV2-S) to achieve a joint classification of all six parameters and explore the most cost-efficient method.

3.1 VGG-Based CNN and Traditional ML Methods

First, the single-parameter classification tasks of each model are tested, including OSNR, CD, ROF, PS, BR, and MF. For instance, when monitoring OSNR, the remaining five parameters are randomly combined. In this case, the designed VGG-based CNN is used as the comparison model; the results are presented in Table 3 for a clearer comparison. It is evident that KNN, SVM, and VGG-based CNN perform well in all single-parameter classification tasks, achieving accuracy above 95%. In the classification tasks of PS and MF, all ML methods exhibit high accuracy, indicating that these two parameters have distinct and easily distinguishable characteristics. These results can also be observed from the eye diagrams depicted in Fig. 3. MF corresponds to the number of eyes in eye diagrams, while PS corresponds to the probability of occurrence of different amplitudes, both of which can be readily identified from the diagrams. However, the classification accuracy of DT and GBDT for some parameters is not satisfactory. Therefore, for these parameters, the confusion matrix is provided for DT and GBDT in Fig. 6 to facilitate a more detailed analysis. Regarding DT, in the OSNR task, adjacent OSNR values are more likely to be confused as OSNR increases, particularly for 35 and 40 dB. In the CD task, classification failures mainly occur between −40 and −80 ps/nm. For the ROF task, the slight difference between shaped waveforms at ROF of 0.1 and 0.2 results in a higher likelihood of confusion, as is evident from the accuracy scores. In the BR task, a 1-GBd variation corresponds to 1.5 ps in the time domain. Due to the slight difference in bit period, the probability of misclassification is higher, with the probabilities of being classified into other classes hovering around 20%. The effects caused by BR change on eye diagrams are relatively weak especially when values of ROF and OSNR are small, which also increases the difficulty of classification. Compared to DT, GBDT demonstrates better accuracy in each single-parameter classification task due to gradient-based calculations and fitting. However, the accuracy for the BR task remains relatively low. The confusion matrix reveals that the main challenge lies in distinguishing between 26 and 27 GBd.

Then, in order to further test the performance of each method on multiparameter classification, all six parameters except for the MF are divided into two sets: digital signal parameters (PS, ROF, and BR) and optical link parameters (OSNR and CD), corresponding to 27 (3^3) and 16 (4^2) classes, respectively.

Table 3 Accuracy of single-parameter classifications of different ML methods.

Method	OSNR (%)	CD (%)	ROF (%)	BR (%)	PS (%)	MF (%)
KNN	99.44	99.92	95.58	95.41	99.96	100
DT	92.03	92.48	78.88	61.71	99.54	95.81
SVM	99.32	99.78	97.02	97.31	99.92	100
GBDT	99.61	98.82	95.41	88.37	99.98	99.81
VGG-based CNN	99.01	100	98.7	99.21	100	100

CD, chromatic dispersion; ROF, roll-off factor; BR, baud rate; PS, probabilistic shaping; MF, modulation format.

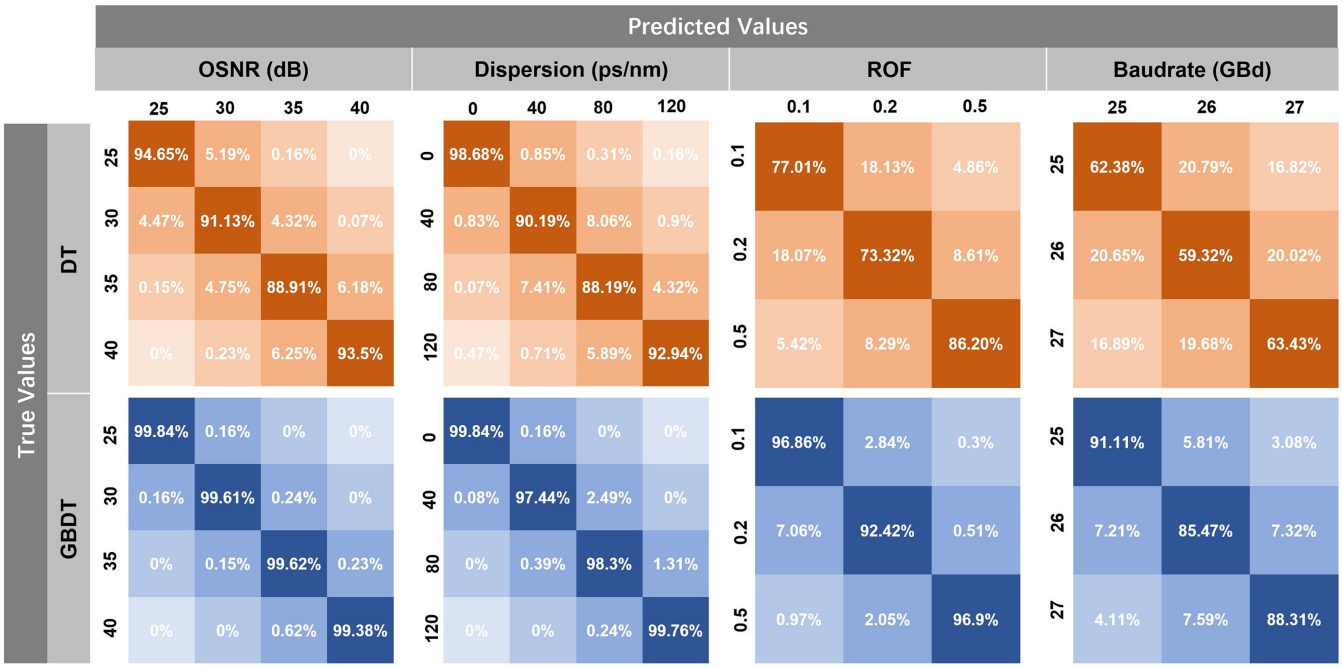


Fig. 6 Confusion matrices of DT and GBDT for OSNR, CD, ROF, and BR classification tasks.

The best learning parameters have been chosen for different methods. Figures 7(a) and 7(b) show the accuracy of jointly monitoring two optical link parameters and three digital signal parameters under different MFs and learning methods. It can be clearly seen that in jointly monitoring two optical link parameters (16 classes), the other four learning methods work well except for the DT, achieving higher accuracy of more than 97%. With the increase of PAM order, the accuracy of each method is reduced. That is because the higher PAM order is more sensitive to the changes in the parameters, resulting in a more distorted eye diagram, which is difficult to identify. However, the accuracy of GBDT significantly drops when jointly monitoring three digital signal parameters (36 classes). This outcome aligns with the accuracy of single-parameter classification tasks and the corresponding confusion matrix presented in Table 3 and Fig. 6. Both GBDT and DT exhibit relatively low accuracy in the classification of BR. Furthermore, the decrease in accuracy for DT and GBDT in the joint monitoring of three digital signal parameters can also be attributed to the increase in the number of classes. In contrast, VGG-based CNN, SVM, and KNN outperform the other two methods in both joint monitoring tasks. Notably, the classification accuracy of VGG-based CNN and SVM is less affected as the PAM order increases, whereas KNN is more significantly influenced by the MF in the joint monitoring of digital signal parameters.

Next, the complexity of the classification is increased to 432 classes for each MF under different combinations of all the digital signal and optical link parameters. For different ML methods, different parameters are optimized. In the KNN model, multiple trainings are performed by sweeping models with different k values, and the model with the highest accuracy is ultimately selected. Similarly, in the DT model, the best combination of parameters is obtained by sweeping the max depth and max features. Max depth is related to the number of layers of DT, and max features refer to the maximum number of features considered during the division of DT. As for SVM, a linear SVM classifier is chosen, and the error penalty C is optimized in different conditions. For GBDT, the primary parameters being swept are the maximum depth, learning rate, and number of iterations, all aimed at achieving better accuracy. The main parameter selections for the aforementioned methods are presented in Table 4. As shown in Fig. 7(c), under more complex

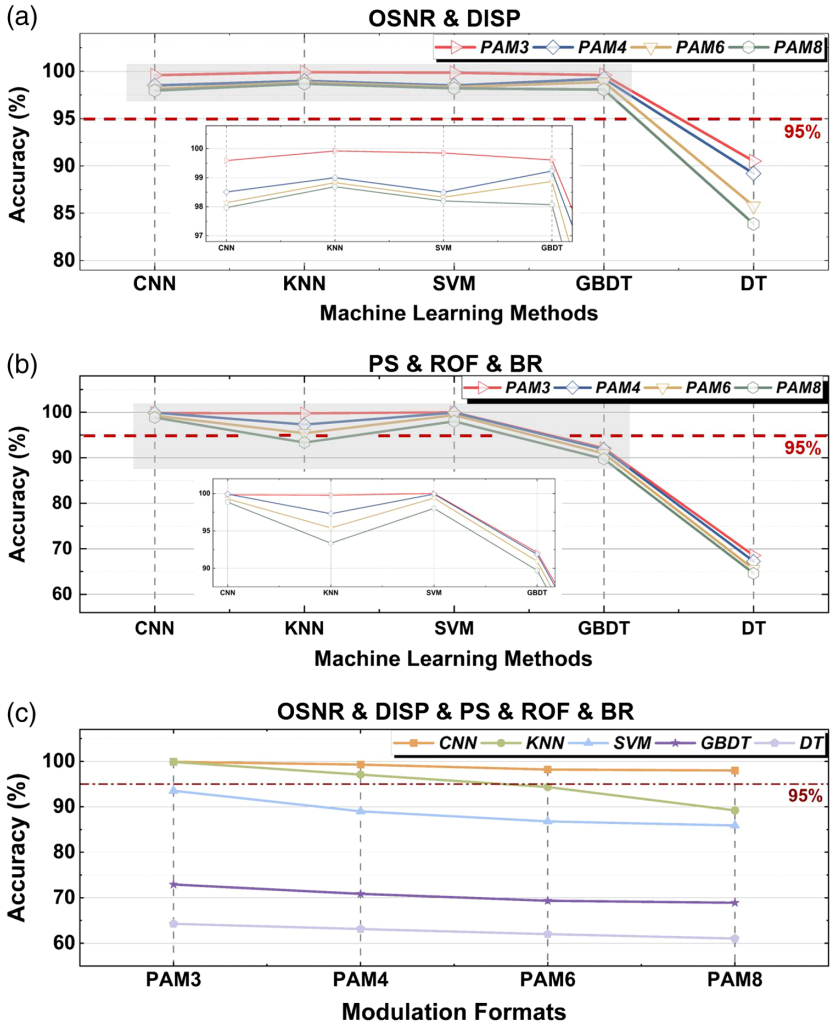


Fig. 7 Accuracy of joint monitoring parameters with different ML methods for (a) digital signal parameters and (b) optical link parameters. (c) Accuracy for all the 432 classes for each MF with different five-parameter combinations.

Table 4 Parameters selected of traditional ML methods.

Method	PAM3	PAM4	PAM6	PAM8	All classes
KNN	$k = 3$	$k = 3$	$k = 4$	$k = 3$	$k = 5$
DT	MD = 400	MD = 400	MD = 400	MD = 400	MD = 400
	MF = 300	MF = 400	MF = 300	MF = 900	MF = 600
SVM	$C = 10$	$C = 8$	$C = 8$	$C = 10$	$C = 10$
GBDT	LR = 0.1	LR = 0.1	LR = 0.1	LR = 0.1	LR = 0.1
	MD = 6	MD = 6	MD = 6	MD = 7	MD = 7
	iter = 350	iter = 400	iter = 370	iter = 420	iter = 500

MD, max depth; MF, max features; LR, learning rate; iter, iterations; C, error penalty.

conditions, the accuracy of SVM and KNN decreased from 93.52% to 85.90% and 99.85% to 89.2%, respectively, with the increase of PAM orders. On the other hand, the VGG-based CNN consistently maintained an accuracy of over 97% for all conditions, surpassing the other four methods. Similarly, Fig. 7(c) shows that compared with SVM and VGG-based CNN, the performance of KNN is significantly affected by the MF. With an increase in PAM order from 3 to 8, the accuracy of KNN correspondingly dropped by ∼10%.

Finally, when the complexity of the classification is increased to 1728 classes under different combinations of all the six parameters including MFs, the advantages of CNN over traditional ML become more obvious. The main parameter selections of the traditional ML methods are also shown in the last column of Table 4. From Fig. 8, it can be clearly seen that under the most complex conditions, 97.61% accuracy is achieved by VGG-based CNN, which is ∼7% higher than the highest accuracy among the other four methods. In summary, with the increase of classification classes, the accuracy of each method decreases, and it is more obvious for traditional ML methods. In contrast, VGG-based CNN exhibits better robustness in both single-parameter monitoring and multiparameter joint monitoring, consistently achieving an accuracy above 97%. Nevertheless, in the joint monitoring of a small number of parameters, some traditional ML methods, such as SVM, KNN, and GBDT, are also competitive options.

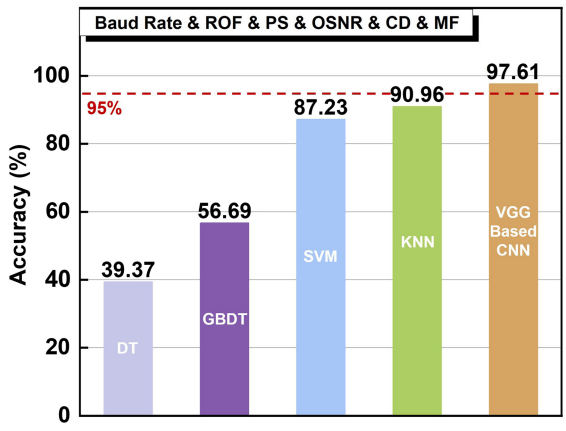


Fig. 8 Accuracy for all 1728 classes with different six-parameter combinations using DT, GBDT, KNN, SVM, and VGG-based CNN.

Among the four traditional ML methods, KNN outperforms the other three methods, reaching an accuracy of 90.96%, even in the most complex six-parameter joint monitoring scenario. However, the performance of KNN is more susceptible to the MF. As for DT and GBDT, they perform less well on tasks that jointly monitor more parameters. This may be due to the unsuitability of the feature combination of HOG and color histograms for these two methods. Therefore, for DT and GBDT, in order to achieve higher accuracy, more appropriate features should be found and selected. In fact, taking more attempts at feature selection and parameter optimization could indeed improve the accuracy of traditional ML methods, but the workload and time consumption of manual feature selection are also issues worthy of attention. In real-time optical network monitoring, accurately identifying the most suitable features is challenging, highlighting the importance of automatic feature extraction offered by DL methods. Therefore, due to the higher accuracy and the ability to adapt to various simple and complex scenarios, CNN represents a better choice for accurate performance monitoring of a PAM-based channel.

In addition, to address the feature selection challenge in the GBDT model, a CNN followed by a GBDT model is employed for comparison. Initially, the designed VGG-based CNN is used for feature extraction, and the output features are then processed using GBDT instead of FC layers. The size of output features is 4096×1. Table 5 shows the comparison of the accuracy of CNN, CNN+GBDT, and GBDT under different classification tasks. "All class" in Table 5 refers to the joint monitoring of all six parameters. After feature extraction by CNN, the accuracy for monitoring BR and ROF is significantly improved using GBDT. Therefore, in the joint monitoring of three digital signal parameters, CNN + GBDT achieves an accuracy improvement of nearly 7% compared to using GBDT alone. When performing six-parameter joint monitoring, the accuracy of CNN + GBDT also sees substantial improvement, although it still falls short of using CNN alone. This could be attributed to the relatively large number of classes (1728) for GBDT. In addition, utilizing CNN + GBDT involves training both the CNN and GBDT models, which results in a longer training time compared to using FC layers for classification. It needs to be admitted that by changing the output feature length of CNN, the performance of CNN + GBDT could be further optimized, but it needs a long time to do a grid search to find the best feature length. Therefore, using CNN seems to be a more effective choice for multiparameter performance monitoring in comparison.

Table 5 Accuracy of classifications of GBDT, VGG-based CNN, and VGG-based CNN + GBDT.

Method	OSNR (%)	CD (%)	ROF (%)	BR (%)	PS (%)	MF (%)	OSNR and CD (%)	ROF and PS and BR (%)	All classes (%)
CNN	99.01	100	98.7	99.21	100	100	99.32	99.12	97.61
CNN + GBDT	99.16	99.61	98.81	98.1	100	100	99.53	98.11	83.13
GBDT	99.61	98.82	95.41	88.37	99.98	99.81	99.03	91.47	56.69

CD, chromatic dispersion; ROF, roll-off factor; BR, baud rate; PS, probabilistic shaping; MF, modulation format.

3.2 Comparisons among Four Modern CNN Models

Accuracy curves of the first 30 epochs of each model are plotted in Fig. 9(a). Analysis shows that the accuracy of all the models can converge to a stable value within 30 epochs, where the convergence rate is related to the learning rate.

Figure 9(b) uses the box plot to show the accuracy distribution of the last 150 epochs of each CNN model. For the results of each model, the upper and lower ends of the black solid line represent the maximum and minimum of the data points, respectively, while the points that are not within the range are regarded as outliers. The two sides of the box are the upper and the lower quartile, and the dashed line within the box indicates the median. The dotted line on the right represents the normal distribution fitted according to the data points. Basically, all these four models can reach a high accuracy of >96% for jointly monitoring all six parameters of the PAM-based channel. Among them, EfficientNetV2-S takes the leading role, with an accuracy distribution centered on around 97.5%, while ResNet-18 seems to be stabler, with smaller fluctuations of data points.

Some key parameters reflecting the complexity of the models are also compared in Table 6. FLOPs correspond to the computation time, and the parameters correspond to the consumption of resources. The FLOPs of the convolutional layers can be calculated by Eq. (3), and the FLOPs of FC layers are expressed in Eq. (4),

$$\mathrm{FLOPs}_{\mathrm{CNN}} = 2 \times (C_i \times k_w \times k_h) \times C_o \times W \times H, \tag{6}$$

$$\mathrm{FLOPs}_{\mathrm{FC}} = (2 \times I) \times O, \tag{7}$$

where C_i and C_o refer to the channels of input and output, respectively. H and W represent the length and width of the output feature map, k_w and k_h are the length and width of the convolution kernel, and I and O represent the number of the input and output neurons, respectively.[59] Although EfficientNetV2-S has the highest accuracy, its FLOPs and parameters are the largest among the four models, meaning greater resource consumption and longer computation time. In contrast, MobileNetV3-S realizes fewer parameters and FLOPs using the MBConv block. Meanwhile, the VGG-based model designed here utilizes fewer Conv layers to achieve the least memory usage. Neither of these two methods comes at the expense of accuracy. Therefore, MobileNetV3-S and the designed few-layer VGG-based model

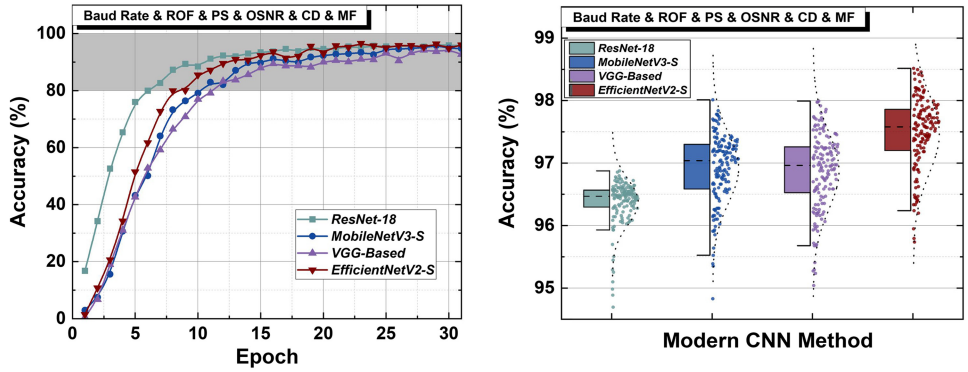


Fig. 9 (a) Accuracy curves and (b) distributions of VGG-based model, ResNet-18, MobileNetV3-S, and EfficientNetV2-S.

Table 6 Computational cost per image of the modern CNN models.

Model name	Input size	FLOP	Parameter	Memory
MobileNetV3-S	224×224×3	64.36 M	3.28 M	18.44 MB
VGG-based	224×224×3	573.13 M	120 M	15.00 MB
Resnet-18	224×224×3	1.82 G	12.06 M	28.53 MB
EfficientNetV2-S	224×224×3	2.9 G	23.41 M	139.00 MB

FLOPs, floating point operations; memory, the usage of video memory.

seems to be more cost-efficient for jointly monitoring all six parameters. It further indicates that for joint monitoring of PAM channels, fewer Conv layers are sufficient for feature extraction. In other words, it seems that a network with fewer layers including MBConv and Fused-MBConv blocks can be designed to achieve the best trade-off between accuracy and computational cost.

4 Discussion

In previous results, the potential of single-parameter as well as up to six-parameter joint monitoring of PAM signals using different CNN models has been demonstrated. However, it should be noted that the number of the final output neurons N of each CNN model needs to be changed according to the classes of the task. In addition, the trained model can only perform a specific classification task and requires retraining for other monitoring tasks. This lack of flexibility makes it challenging to apply the model to scenarios with broader requirements. In communication networks, there is a greater need for a model that can perform joint monitoring of multiple parameters while simultaneously monitoring each single parameter. To address these limitations, an MTL model combined with CNN is designed here to improve the applicability of the network. As a DL model, MTL has been applied in many applications.[60] Compared to single-task learning, MTL enables simultaneous monitoring of multiple single-parameter tasks. Common parameters are shared among different tasks in MTL, and the losses of different tasks jointly update the shared layer. The MTL model used in this work has six tasks, which are MF, BR, CD, OSNR, ROF, and PS, respectively. Therefore, the output neurons of six-parameter joint monitoring in MTL could be reduced from 1728 to 21. The designed structure is shown in Fig. 10.

Before training the model, the input image undergoes preprocessing. Data augmentation is implemented to enlarge the data set for the model, the color jitter, random cropping, and center cropping with a size $320 \times 320 \times 3$ are utilized to increase data set diversity. Following that, the input eye diagrams are normalized using mean and standard deviation to make the data better respond to the activation function. Next, the data set (1728 classes) is divided into 80% for training (1382 random classes) and 20% for testing (346 classes). When all the six subtasks are classified correctly, it means that the corresponding six-parameter joint monitoring is correct. Consequently, the MTL model can perform joint monitoring of multiple parameters while simultaneously monitoring each single parameter. In this work, the lightweight MobileNetV3-Small is selected for the convolutional processing of MTL to achieve more efficient training. The specific parameters utilized in the MobileNetV3-Small are also shown in Table 7.

For classification tasks, cross-entropy is usually used as the loss function, as shown in Eq. (8),

$$L = \frac{1}{N}\sum_{i=1}^{N} L_i = -\frac{1}{N}\sum_{i=1}^{N}\sum_{c=1}^{M} y_{ic}\,\log(p_{ic}), \tag{8}$$

where N is the number of samples and c represents the specific class in the task. y_{ic} is the label of task and p_{ic} denotes the predicted possibility. So, the final loss function of the MTL model can be expressed as

$$L = \sum_{i=1}^{k} \lambda_i L_i, \tag{9}$$

where k represents the total number of tasks in the MTL model, and λ_i is the weight of the ith task in the loss function. λ_i could be adjusted to change the importance of each task in the MTL model by further searching the grid. The loss weights corresponding to six tasks are listed in Table 8.

Figure 11 shows the accuracy comparison between VGG-based CNN and MTL. Both models achieve high accuracy (>97%) in all single-parameter classification tasks. Moreover, in the six-parameter joint monitoring, although MTL exhibits slightly lower accuracy than VGG-based CNN, it still maintains a high accuracy, exceeding 95%. The accuracy of MTL in various monitoring tasks is comparable to that of CNN, while offering the advantage of not requiring training on all 1728 classes. In addition, MTL allows for simultaneous monitoring of both single parameters and multiple parameters, providing greater flexibility compared to a CNN designed for a single task. Furthermore, the MTL model presents opportunities for further optimization. First, the structure of the FC layer corresponding to multiple tasks can be further optimized to achieve better accuracy. This may involve increasing the number of FC layers or modifying the output neurons of hidden layers. Second, the weights of loss functions can be updated using GradNorm, a method that dynamically adjusts the weight of each loss function, saving time compared to grid searching.[61]

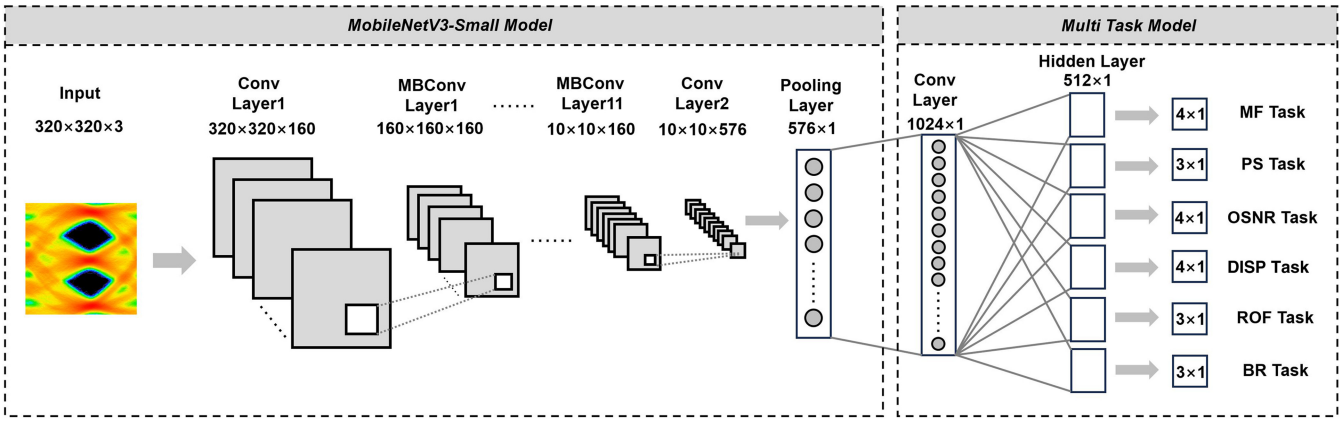


Fig. 10 Structure of MTL model combined with MobileNetV3-Small.

Table 7 Structure of MobileNetV3-Small.

Input	Operator	Out channel	SE	NL	Stride
320×320×3	Conv 3×3	16	—	HS	2
160×160×16	MBConv 3×3	16	Yes	RE	2
80×80×16	MBConv 3×3	24	—	RE	2
40×40×24	MBConv 3×3	24	—	RE	1
40×40×24	MBConv 5×5	40	Yes	HS	2
20×20×40	MBConv 5×5	40	Yes	HS	1
20×20×40	MBConv 5×5	40	Yes	HS	1
20×20×40	MBConv 5×5	48	Yes	HS	1
20×20×48	MBConv 5×5	48	Yes	HS	1
20×20×48	MBConv 5×5	96	Yes	HS	2
10×10×96	MBConv 5×5	96	Yes	HS	1
10×10×96	MBConv 5×5	96	Yes	HS	1
10×10×96	Conv 1×1	576	Yes	HS	1
10×10×576	Pooling 7×7	576	—	—	1
1×1×576	Conv 1×1, NBN	1280	—	HS	1

SE, whether there is a squeeze-and-excite in that block; NL, nonlinearity used; HS, H-swish; RE, ReLU; NBN, no batch normalization.

Table 8 Weight of the tasks in the loss function.

Task	BR	MF	ROF	PS	OSNR	CD
Weight	1.01	0.81	0.99	0.78	0.98	0.79

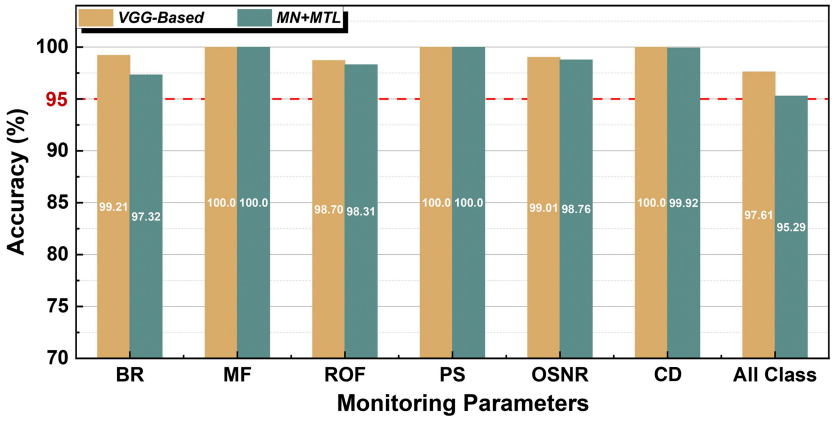


Fig. 11 Accuracy of different monitoring tasks using MTL and VGG-based CNN.

5 Conclusion

We have proposed utilizing an eye diagram to monitor multiple performance parameters of an IMDD channel in optical communication systems using CNN-based methods. Our designed VGG-based CNN model with fewer Conv layers can achieve joint monitoring of both digital signal parameters and optical link parameters of the PAM-based IMDD communication system with higher accuracy and less memory usage. Compared to the traditional ML methods, the proposed VGG-based CNN model can automatically extract features and successfully deal with a variety of complex monitoring issues. With a high accuracy of 97.16% for up to six parameters, joint monitoring, including BR, PS, ROF, OSNR, CD, and MFs, the VGG-based CNN model shows greater potential than the other traditional ML methods in various applications of both static and dynamic optical networks.

Furthermore, we have also compared four CNN models, including the proposed VGG-based model, ResNet-18, MobileNetV3, and EfficientNetV2. With less computational

cost, the VGG-based model and MobileNetV3 reduce the hardware requirements. For jointly monitoring all six parameters of the PAM-based channel, the EfficientNetV2 with the highest accuracy of 97.55% and the lightweight MobileNetV3 with the fewest FLOPs also provide bright ideas for further optimization of the network. For more complex IMDD-based DC networks using PAM signals, a cost-efficient CNN with fewer layers with an optimal combination of MBConv and Fused-MBConv blocks can be further designed to achieve higher accuracy without the increment of computational resources.

Finally, an MTL model combined with MobileNetV3 is designed to address the limitations of the VGG-based CNN and provide more flexibility for applications. The model achieves high accuracy (>95%) in all classification tasks. Furthermore, MTL offers the advantage of not requiring training on all 1728 classes, enabling simultaneous monitoring of both single parameters and multiple parameters by calculating the gradient of multiple losses. This enhanced flexibility surpasses that of a CNN designed for a single task, highlighting its potential in various monitoring scenarios.

Disclosures

The authors have no relevant financial interests relating to the study and no other potential conflicts of interest to disclose.

Code and Data Availability

The data that support the findings of this study are available from the corresponding author upon reasonable request.

Acknowledgments

This work was supported by the National Key Research and Development Program of China (Grant No. 2019YFB1803700) and the Key Technologies Research and Development Program of Tianjin (Grant No. 20YFZCGX00440).

References

1. X. Zhou, R. Urata, and H. Liu, "Beyond 1 Tb/s intra-data center interconnect technology: IM-DD OR coherent?" *J. Lightwave Technol.* **38**(2), 475–484 (2020).
2. E. Maniloff, S. Gareau, and M. Moyer, "400G and beyond: coherent evolution to high-capacity inter data center links," in *Proc. Opt. Fiber Commun. Conf. and Exhibit. (OFC)*, p. M3H.4 (2019).
3. Q. Cheng et al., "Recent advances in optical technologies for data centers: a review," *Optica* **5**(11), 1354–1370 (2018).
4. S. T. Le et al., "Beyond 400 Gb/s direct detection over 80 km for data center interconnect applications," *J. Lightwave Technol.* **38**(2), 538–545 (2020).
5. Z. Qu et al., "Single-lambda 100G-PAM4 QSFP28 transceiver for 80-km C-band transmission," *Proc. SPIE* **11038**, 1130806 (2020).
6. S.-R. Moon et al., "Realization of real-time DSP for C-band PAM-4 transmission in inter-datacenter network," *Opt. Express* **28**(2), 1269–1278 (2020).
7. N. Eiselt et al., "Evaluation of real-time 8×56.25 Gb/s (400G) PAM-4 for inter-data center application over 80 km of SSMF at 1550 nm," *J. Lightwave Technol.* **35**(4), 955–962 (2016).
8. H. Mardoyan et al., "84-, 100-, and 107-GBd PAM-4 intensity-modulation direct-detection transceiver for datacenter interconnects," *J. Lightwave Technol.* **35**(6), 1253–1259 (2017).
9. M. Chagnon, "Direct-detection technologies for intra-and inter-data center optical links," in *Proc. Opt. Fiber Commun. Conf. and Exhibit. (OFC)*, p. W1F.4 (2019).
10. A. Mecozzi and M. Shtaif, "Information capacity of direct detection optical transmission systems," *J. Lightwave Technol.* **36**(3), 689–694 (2017).
11. "COLORZ®," Inphi Corporation, https://www.marvell.com/products/optical-modules (accessed 15 Oct. 2021).
12. D. Zou et al., "100G PAM-6 and PAM-8 signal transmission enabled by pre-chirping for 10-km intra-DCI utilizing MZM in C-band," *J. Lightwave Technol.* **38**(13), 3445–3453 (2020).
13. H. Xin et al., "120 GBaud PAM-4/PAM-6 generation and detection by photonic aided digital-to-analog converter and linear equalization," *J. Lightwave Technol.* **38**(8), 2226–2230 (2020).
14. J. Zhang et al., "280 Gb/s IM/DD PS-PAM-8 transmission over 10 km SSMF at O-band for optical interconnects," in *Proc. Opt. Fiber Commun. Conf. (OFC)*, p. M4F.1 (2020).
15. T. A. Eriksson et al., "56 Gbaud probabilistically shaped PAM8 for data center interconnects," in *Proc. Eur. Conf. Opt. Commun. (ECOC)* (2017).
16. R. A. Linke and A. H. Gnauck, "High-capacity coherent lightwave systems," *J. Lightwave Technol.* **6**(11), 1750–1769 (1988).
17. D.-S. Ly-Gagnon et al., "Coherent detection of optical quadrature phase-shift keying signals with carrier phase estimation," *J. Lightwave Technol.* **24**(1), 12–21 (2006).
18. M. Morsy-Osman and D. V. Plant, "A comparative study of technology options for next generation intra- and inter-datacenter interconnects," in *Proc. Opt. Fiber Commun. Conf. and Exhibit. (OFC)*, p. W4E.1 (2019).
19. J. Wei et al., "System aspects of the next-generation data-center networks based on 200G per lambda IMDD links," *Proc. SPIE* **11308**, 1130805 (2020).
20. E. El-Fiky et al., "400 Gb/s O-band silicon photonic transmitter for intra-datacenter optical interconnects," *Opt. Express* **27**(7), 10258–10268 (2019).
21. A. Nag, M. Tornatore, and B. Mukherjee, "Optical network design with mixed line rates and multiple modulation formats," *J. Lightwave Technol.* **28**(4), 466–475 (2009).
22. W. Wei, C. Wang, and J. Yu, "Cognitive optical networks: key drivers, enabling techniques, and adaptive bandwidth services," *IEEE Commun. Mag.* **50**(1), 106–113 (2012).
23. Q. Cai et al., "Modulation format identification in fiber communications using single dynamical node-based photonic reservoir computing," *Photonics Res.* **9**(1), B1–B8 (2021).
24. X. Han et al., "Joint probabilistic-Nyquist pulse shaping for an LDPC-coded 8-PAM signal in DWDM data center communications," *Appl. Sci.* **9**(23), 4996 (2019).
25. M. Zhu et al., "Optical single side-band Nyquist PAM-4 transmission using dual-drive MZM modulation and direct detection," *Opt. Express* **26**(6), 6629–6638 (2018).
26. K. Wang et al., "High-speed PS-PAM8 transmission in a four-lane IM/DD system using SOA at O-band for 800G DCI," *IEEE Photonics Technol. Lett.* **32**(6), 293–296 (2020).
27. Z. Qu and I. B. Djordjevic, "On the probabilistic shaping and geometric shaping in optical communication systems," *IEEE Access* **7**, 21454–21464 (2019).
28. L. Sun et al., "Dyadic probabilistic shaping of PAM-4 and PAM-8 for cost-effective VCSEL-MMF optical interconnection," *IEEE Photonics J.* **11**(2), 7202611 (2019).
29. M. N. Sakib and O. Liboiron-Ladouceur, "A study of error correction codes for PAM signals in data center applications," *IEEE Photonics Technol. Lett.* **25**(23), 2274–2277 (2013).
30. T. Yoshida, M. Karlsson, and E. Agrell, "Performance metrics for systems with soft-decision FEC and probabilistic shaping," *IEEE Photonics Technol. Lett.* **29**(23), 2111–2114 (2017).
31. S.-R. Moon et al., "C-band PAM-4 signal transmission using soft-output MLSE and LDPC code," *Opt. Express* **27**(1), 110–120 (2019).
32. R. Gu, Z. Yang, and Y. Ji, "Machine learning for intelligent optical networks: a comprehensive survey," *J. Network Comput. Appl.* **157**, 102576 (2020).

33. W. S. Saif et al., "Machine learning techniques for optical performance monitoring and modulation format identification: a survey," *IEEE Commun. Surv. Tutor.* **22**(4), 2839–2882 (2020).
34. V. S. Ghayal and R. Jeyachitra, *Advances in Electrical and Computer Technologies*, Springer, Singapore (2020).
35. S. Savian et al., "Joint estimation of IQ phase and gain imbalances using convolutional neural networks on eye diagrams," in *Proc. Conf. Lasers and Electro-Opt. (CLEO)*, p. STh1C.3 (2018).
36. D. Wang et al., "Intelligent constellation diagram analyzer using convolutional neural network-based deep learning," *Opt. Express* **25**(15), 17150–17166 (2017).
37. S. Peng et al., "Modulation classification based on signal constellation diagrams and deep learning," *IEEE Trans. Neural Networks Learn. Syst.* **30**(3), 718–727 (2018).
38. K. Jiang et al., "A novel digital modulation recognition algorithm based on deep convolutional neural network," *Appl. Sci.* **10**(3), 1166 (2020).
39. H. Lv et al., "Joint OSNR monitoring and modulation format identification on signal amplitude histograms using convolutional neural network," *Opt. Fiber Technol.* **61**, 102455 (2021).
40. J. Du et al., "A CNN-based cost-effective modulation format identification scheme by low-bandwidth direct detecting and low rate sampling for elastic optical networks," *Opt. Commun.* **471**, 126007 (2020).
41. S. D. Dods and T. B. Anderson, "Optical performance monitoring technique using delay tap asynchronous waveform sampling," in *Proc. Opt. Fiber Commun. Conf. and Exhibit. (OFC)*, p. OThP5 (2006).
42. D. Lippiatt et al., "Impairment identification for PAM-4 transceivers and links using machine learning," in *Proc. Opt. Fiber Commun. Conf. and Exhibit. (OFC)*, p. W7A.5 (2021).
43. T. Tanimura et al., "Convolutional neural network-based optical performance monitoring for optical transport networks," *J. Opt. Commun. Networking* **11**(1), A52–A59 (2019).
44. D. Wang et al., "Modulation format recognition and OSNR estimation using CNN-based deep learning," *IEEE Photonic. Technol. Lett.* **29**(19), 1667–1670 (2017).
45. Z. Zhao et al., "A modulation format identification method based signal amplitude sorting and ratio calculation," *Opt. Commun.* **470**, 125819 (2020).
46. M. C. Tan et al., "Simultaneous optical performance monitoring and modulation format/bit-rate identification using principal component analysis," *J. Opt. Commun. Networking* **6**(5), 441–448 (2014).
47. X. Fan et al., "Joint optical performance monitoring and modulation format/bit-rate identification by CNN-based multi-task learning," *IEEE Photonics J.* **10**(5), 1–12 (2018).
48. M. Yang et al., "FPGA-based real-time soft-decision LDPC performance verification for 50G-PON," in *Proc. Opt. Fiber Commun. Conf. and Exhibit. (OFC)*, p. W3H.2 (2019).
49. M. M. Rad et al., "Passive optical network monitoring: challenges and requirements," *IEEE Commun. Mag.* **49**(2), S45–S52 (2011).
50. Z. Pan, C. Yu, and A. E. Willner, "Optical performance monitoring for the next generation optical communication networks," *Opt. Fiber Technol.* **16**(1), 20–45 (2010).
51. M. Awad and R. Khanna, Efficient Learning Machines, Springer, Berkeley, California (2015).
52. S. Zhang et al., "Learning k for KNN classification," *ACM Trans. Intell. Syst. Technol.* **8**(3), 1–19 (2017).
53. C. H. Gladwin, *Ethnographic Decision Tree Modeling*, Sage (1989).
54. J. H. Friedman, "Greedy function approximation: a gradient boosting machine," *Ann. Stat.* **29**(5), 1189–1232 (2001).
55. M. Mateen et al., "Fundus image classification using VGG-19 architecture with PCA and SVD," *Symmetry* **11**(1), 1 (2019).
56. K. He et al., "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vision and Pattern Recognit.*, pp. 770–778 (2016).
57. A. Howard et al., "Searching for MobileNetV3," in *Proc. IEEE/CVF Int. Conf. Comput. Vision*, pp. 1314–1324 (2019).
58. M. Tan and Q. V. Le, "EfficientNetV2: smaller models and faster training," in *Proc. 38th Int. Conf. Mach. Learn. (PMLR)*, pp. 10096–10106 (2021).
59. P. J. Freire et al., "Performance versus complexity study of neural network equalizers in coherent optical systems," *J. Lightwave Technol.* **39**(19), 6085–6096 (2021).
60. H. Luo et al., "Cost-effective multi-parameter optical performance monitoring using multi-task deep learning with adaptive ADTP and AAH," *J. Lightwave Technol.* **39**(6), 1733–1741 (2021).
61. Z. Chen et al., "GradNorm: gradient normalization for adaptive loss balancing in deep multitask networks," in *Proc. Int. Conf. Mach. Learn.*, pp. 794–803 (2018).

Si-Ao Li received his BS degree in optical information science and technology from the Dalian University of Technology, China, in 2018. He is currently pursuing his PhD in optical engineering at the Institute of Modern Optics, Nankai University, China.

Yang Yue is a professor in the School of Information and Communications Engineering, Xi'an Jiaotong University, China. He is the founder and current PI of Intelligent Photonics Applied Technology Lab (iPatLab). His current research interest is intelligent photonics, including optical communications, optical perception, and optical chips. He is a fellow of SPIE, and a senior member of IEEE and Optica. He has published over 270 journal papers (including Science) and conference proceedings with >12,000 citations.

Biographies of the other authors are not available.