IEEE ComSoc IEEE Communications Society | IEEE COMPUTER SOCIETY | IEEE Signal Processing Society | VTS | IEEE Transactions on Machine Learning in Communications and Networking

Received 17 December 2023; revised 31 March 2024, 29 May 2024, and 3 June 2024; accepted 5 June 2024.
Date of publication 11 June 2024; date of current version 18 June 2024.

The associate editor coordinating the review of this article and approving it for publication was X. Chu.

Digital Object Identifier 10.1109/TMLCN.2024.3412054

Joint SNR and Rician K-Factor Estimation Using Multimodal Network Over Mobile Fading Channels

KOSUKE TAMURA[1] (Graduate Student Member, IEEE), SHUN KOJIMA[2] (Member, IEEE), PHUC V. TRINH[2] (Member, IEEE), SHINYA SUGIURA[2] (Senior Member, IEEE), AND CHANG-JUN AHN[1] (Senior Member, IEEE)

[1]Graduate School of Engineering, Chiba University, Chiba 263-8522, Japan
[2]Institute of Industrial Science, The University of Tokyo, Tokyo 153-8505, Japan
CORRESPONDING AUTHOR: S. KOJIMA (s.kojima@ieee.org)

This work was supported in part by JST ACT-X under Grant JPMJAX22A5, in part by JSPS KAKENHI under Grant 22K14253 and Grant 22K04085, and in part by the Support Center for Advanced Telecommunications Technology Research (SCAT).

ABSTRACT This paper proposes a novel joint signal-to-noise ratio (SNR) and Rician K-factor estimation scheme based on supervised multimodal learning. In the case of using machine learning to estimate the communication environment, achieving high accuracy requires a sufficient amount of training data. To solve this problem, we introduce a multimodal convolutional neural network (CNN) structure using different waveform formats. The proposed scheme obtains ‘‘feature diversity’’ by increasing the modalities from the same received signal, such as sequence data and spectrogram image. Especially with a limited dataset, training convergence is accelerated since different features can be extracted from each modality. Simulations demonstrate that the presented scheme achieves superior performance compared to conventional estimation methods.

INDEX TERMS Convolutional neural network (CNN), multimodal learning, Rayleigh fading, Rician fading, Rician K-factor estimation, SNR estimation.

I. INTRODUCTION

With the proliferation of the Internet of Things (IoT), the demand for wireless communications is increasing. In IoT, efficient communication is required under limited power and computing resources. Therefore, it is important to promptly recognize the surrounding environment and implement it to reflect in the overall system for efficient communication [1], [2], [3], [4]. Many methods for estimating the communication environment have been studied. In particular, signal-to-noise ratio (SNR) and Rician K-factor, which indicate the effects of noise and direct and scattered waves, are important factors that determine the communication environment [5], [6], [7], [8]. These estimation methods are classified into two main categories: non-machine learning (ML)-based and ML-based approaches.

As non-ML-based approaches, numerous schemes have been extensively researched. In SNR estimation, these can be classified as data-aided (DA) requiring additional reference signal and non-data-aided (NDA) without reference signal [5], [6]. The DA approach provides simple and accurate SNR estimation; it sacrifices spectral efficiency due to the use of many reference signals [9]. In contrast, the NDA approach generally requires relatively long observation time to achieve satisfactory performance. It is unsuitable for time-varying conditions, such as in fast-moving environments [10], [11]. K-factor estimation of the non-ML-based approach includes using the moments of the received signal for estimation [7], [8] and estimating based on the envelope of the received signal [12]. For K-factor estimation, most of these are DA-based approaches to cancel self-interference in the modulated phase, leading to loss of spectral efficiency [13], [14].

On the other hand, ML-based approaches overcome these challenges, such as spectral efficiency issues due to DA, model mismatches, and limited observation time. ML has a strong generalization capability by absorbing knowledge of the communication environment and making per-packet estimates without relying on reference signals. Additionally, ML-based estimation exhibits flexibility in preventing mismatches; thus, it can adapt dynamically to changing scenarios and characteristics [5], [6]. With these characteristics, ML applications in wireless communication technology have been actively investigated in recent years, not only in SNR and K-factor estimation [11], [14], [15], [16], [17] but also in automatic modulation classification and localization [18], [19].

ML-based estimations are superior to non-ML-based estimations in accuracy and robustness, although the challenge lies in obtaining sufficient training data for constructing ML models. Many studies have addressed this challenge using transfer learning (TL) [15], [19] and data augmentation with a generative adversarial network (GAN) [20], [21]. TL is a training technique aimed at improving the performance of the target domain by transferring knowledge from different but related source domains. A typical example involves transferring a model previously trained on a large dataset when insufficient data is available for a specific task and then fine-tuning it to suit the requirements of the new task [22]. Data argumentation with GAN is a technique that generates an unlimited amount of labeled data from generators that have completed training and adds them to the training dataset. GAN consists of two neural networks, a generator, and a discriminator. The generator takes a random variable as input and generates images by repeated up-sampling with transpose convolution. The discriminator evaluates whether the input data is generated or obtained from a training dataset and updates each parameter by feeding it back as a loss [23].

While the ML performance can be improved through data expansion, these schemes have several challenges. In TL, the performance heavily depends on the similarity between the source network's training data and the target data intended for transfer. As the complexity of the target training model increases, it becomes more challenging to match the source model. In addition, collecting an appropriate dataset for a source TL model in wireless communications is remarkably expensive, posing challenges for effective implementation [22]. Data augmentation with GAN requires sufficient data and can collapse when trained with insufficient training data, the same as TL. Furthermore, GAN needs substantial computation to train two networks [24]. As a result, data augmentation for ML in wireless communications remains a challenge in terms of both data collection and computational resources.

Furthermore, accurate SNR and K-factor information is needed without delay to optimize communication parameters in increasingly complex wireless communications. In AMC, the SNR table for control undergoes significant changes, and the throughput performance is degraded due to fluctuating K-factors; thus, it is crucial to have a fast and accurate estimation of K-factors along with SNR estimation [25]. On the other hand, the joint estimation of SNR and K-factor proposed in [13] and [26] are an autocorrelation function and moment-based method that estimates the SNR and K-factor stepwise, respectively. These methods suffer from low estimation accuracy and increased processing time.

Therefore, it is necessary to develop a technique that enables high-performance ML even from limited training data without increasing the computational complexity. In this paper, we propose a multimodal learning-based network to jointly estimate (classify) the communication environment, such as SNR and K-factor. In the proposed network, the output of CNN uses classification. While regression is often utilized for estimating SNR and K-factor using ML, it is sufficient to know what range the estimates fall in for the actual parameter control, and the discrete value of classification is practical enough [27]. Furthermore, classification can result in a lighter network than regression, making it more suitable for situations like ours, where the training dataset is limited [28]. The proposed network simultaneously inputs the sequence data of the received signal and the spectrogram image obtained by short-time Fourier transform (STFT) into the network. This conversion from the same signal to numerical and image formats can increase the number of modalities; thus, it maximizes efficiency and extracts the features necessary for estimation from a small amount of wireless signal data. Since each modality has a different feature extraction robustness, it enables the obtaining of “feature diversity”. The multiple different feature extraction networks are implemented in parallel and combined to achieve both high efficiency in feature extraction and low computational complexity of the networks.

The contributions of this paper are as follows:

1) We explore a scheme that simultaneously captures SNR and K-factor with high accuracy and low complexity. Simultaneous estimation of SNR and K-factor from received signal information without reference signals enables appropriate communication parameter control while minimizing overhead.
2) We propose a multimodal network for wireless environment estimation and investigate its feasibility. In the proposed scheme, the sequence data and spectrogram of the received signal are used as inputs, and the features are extracted separately to improve the performance through feature diversity. Despite limited training data, the proposed scheme demonstrates superior estimation accuracy and processing speed.
3) We reveal the optimal network structure for estimating SNR and K-factor. When conducting multimodal learning, the timing of fusing feature extractions significantly influences accuracy [29]. This paper assumes a network structure with three depths, categorizing it into early fusion, mid-term fusion, and late fusion, and by altering the quantity of training data, it clarifies the optimal fusion stage regardless of the data volume.

In Section II, we overview the related work about estimating the communication environment parameters using ML and multimodal learning. Section III describes the transmission system and the fading model. Section IV shows the proposed multimodal network and describes multimodal training in detail. Section V presents the effectiveness of the proposed network through numerical evaluations. Finally, Section VI provides the conclusion.

II. RELATED WORK

SNR estimation using ML has been widely studied, not only in wireless [11], [14], [15], [16], [17] but also in optical communications [30], [31]. In SNR estimation, ML-based schemes have been shown to have higher accuracy by using image data converted from the sequence data, such as constellations and spectrograms [11], [15], or the sequence data itself [16]. In [11], it trained the network on spectrogram images with frequency, time, and received power data, allowing for joint SNR and Doppler shift estimation. In [15], SNR is estimated by applying TL with constellation image, and [16] shows a higher estimation accuracy by inputting time signals. On the other hand, K-factor estimation is performed using constellation without any information from the transmitter [14]. Further, [17] introduced CNN-based K-factor estimation using spectrum to improve the estimation accuracy. These schemes are good examples of how ML-based estimation of communication environment can perform better than conventional non-ML-based estimation. However, these schemes assume sufficient training data and have not been evaluated for training on limited datasets.

Multimodal learning, in which different data are trained simultaneously, has been studied in several fields such as acoustic signal processing [32], [33] and wireless communications [34], [35], [36], [37]. Specifically, acoustic signal is transformed into numerical values and spectrograms to classify scenes [32], [33]. In wireless communications, multimodal learning has been used for channel estimation in MIMO to achieve more advanced estimation than other ML-based estimations [34]. In [35], despite differences in UAV sensors, multimodal learning accurately estimates non-intuitive features by identifying correlations. In [36], AMC performance has significantly improved due to multimodal characteristics over training from a single feature by simultaneously training time-domain and frequency-domain relationships specific to wireless signals. The paper [37] shows that converting the received signal into the radio image and handcrafted features and then training them simultaneously can achieve advanced estimation even when the SNR is low. Thus, the multimodal learning network shows the potential to outperform networks that are trained from a single modality by effectively exploiting the ability to train from multiple features in wireless communications.

Notably, the works discussed above mainly focus on large datasets and do not address the challenges posed by limited datasets, which are the focus of our paper. Furthermore, to the best of our knowledge, no work has been done using multimodal networks to estimate the communication environment. The network presented in this paper does not require any other sensors to be attached to it since it increases the number of modalities by using the same received signal. Furthermore, its potential is exploited by inputting numeric and image data. Because multiple waveform formats are used as input and feature extraction by individual CNNs is performed in parallel, it efficiently extracts features even from a small amount of data and achieves highly accurate estimation.

III. SYSTEM MODEL

This section describes the characteristics of the Rayleigh and Rician fading model and the transmitter and receiver architecture assumed in this paper.

A. FADING MODEL

1) RAYLEIGH FADING

Let $r(t)$ be the received signal in multipath fading. We assume that the in-phase and quadrature components of this signal follow the Gaussian distribution $\mathcal{N}(0, \sigma^2)$, and define $z(t)$ as $z(t) = |r(t)|$. The probability distribution function (PDF) of z $(z \geq 0)$ follows the Rayleigh distribution, which can be expressed as

$$\begin{aligned} f_z(z) &= \frac{2z}{\bar{P}_r} \exp\left(-\frac{z^2}{\bar{P}_r}\right) \\ &= \frac{z}{\sigma^2} \exp\left(-\frac{z^2}{2\sigma^2}\right), \end{aligned} \tag{1}$$

where $\bar{P}_r = 2\sigma^2$ represents the average received power.

2) RICIAN FADING

We consider the case where a direct wave is occurring; the received signal is a superposition of the line-of-sight (LOS) component and non-LOS (NLOS) components. Therefore, the in-phase and quadrature components satisfy $\mathcal{N}(a, \sigma^2)$. The PDF of the envelope z $(z \geq 0)$ becomes

$$f_z(z) = \frac{z}{\sigma^2} \exp\left\{\frac{-(z^2 + a^2)}{2\sigma^2}\right\} I_0\left(\frac{za}{\sigma^2}\right), \tag{2}$$

where a^2 represents the received power of the LOS component, and $2\sigma^2$ represents the received power of the NLOS components. I_0 denotes the modified Bessel function of the zeroth order. The power ratio between the LOS and NLOS components is defined as the K-factor

$$K = \frac{a^2}{2\sigma^2}. \tag{3}$$

As K increases, the signals in the LOS path become more dominant, and the channel behaves more like a frequency flat fading [38].

B. CHANNEL MODEL

The impulse response of a multipath fading channel is denoted as

$$\begin{aligned} h_d(\tau) &= \sum_{m=0}^{M-1} h_m(t)\delta(\tau - \tau_m) \\ &= \sum_{m=0}^{M-1} h_m(t)\delta\left(\tau - \frac{m}{B}\right), \end{aligned} \tag{4}$$

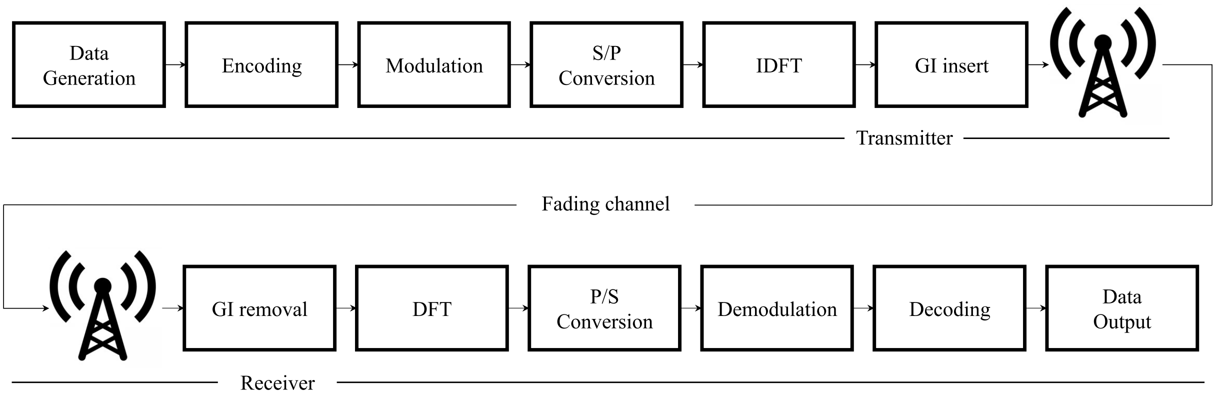


FIGURE 1. **Illustration of the transmitter and receiver structure.**

where

$$M = \lfloor BT_m \rfloor + 1, \tag{5}$$

M represents the discrete path number, and h_m represents the complex gain of each path. τ_m is discrete timing, and B represents the bandwidth, T_m means the multipath spread, and $\delta(\cdot)$ represents the delta function. In this case, the LOS fading channel component, denoted as $h_{LOS}(\tau)$, and the complex channel coefficient h_0 are expressed as follows

$$h_{LOS}(\tau) = h_0 \delta(t - \tau_1), \tag{6}$$

$$h_0 = g_0 \exp(j\theta_0). \tag{7}$$

where g_0 represents its magnitude, and θ_0 represents its phase angle [39]. The path gain of the Rician fading channel, denoted as $h(\tau)$, is given by

$$h(\tau) = \sqrt{\frac{K}{K+1}} h_{LOS}(\tau) + \frac{1}{\sqrt{(K+1)}} h_d(\tau). \tag{8}$$

It is clear from (8) that the LOS component disappears when K is 0, thus (8) becomes $h(\tau) = h_d(\tau)$ in (4), which reduces to Rayleigh fading. Therefore, the channel response depends on the Rician K factor.

C. TRANSMITTER

This paper assumes orthogonal frequency-division multiplexing (OFDM) for transmission. After encoding the data to be transmitted, the signal is modulated. The modulated signal goes through serial-to-parallel (S/P) conversion and is then subjected to inverse discrete Fourier transform (IDFT) to make it into a time-domain signal. A guard interval is inserted to create the transmitted signal to prevent interference. The transmitted signal in the time domain is

$$s(t) = \sum_{w=-\infty}^{\infty} \sqrt{\frac{2P_t}{N_s}} c\,(t - wT) \cdot \left[\sum_{v=1}^{N_s} d(v, w) \exp\left(\frac{j2\pi v\,(t - wT)}{T_s} \right) \right], \tag{9}$$

where

$$c(t) = \begin{cases} 1 & (-T_i \leq t \leq T_s) \\ 0 & \text{otherwise}, \end{cases} \tag{10}$$

where P_t represents the average transmit power, T is the symbol duration, and T_s is the effective symbol length without the guard interval. $d(v, w)$ is the v-th subcarrier of the w-th modulated symbol, which satisfies $E\left[|d(v, w)|\right] = 1$. The guard interval T_i is used, and the relationship $T = T_s + T_i$ is satisfied. The transmitted symbol is represented using any window function defined by the rectangular pulse $c(t)$.

D. RECEIVER

In the time domain, the received signal, denoted as $r(t)$, is affected by fading and additive white Gaussian noise (AWGN) $n(t)$. It can be represented as

$$r(t) = \int_{-\infty}^{\infty} h(\tau, t) s(t - \tau) d\tau + n(t). \tag{11}$$

Here, we assume the channel state is nearly constant over symbol time T. The channel response in the frequency domain can be approximated using the channel response in the time domain as

$$H\left(\frac{v}{T_s}, wT\right) \simeq \int_{-\infty}^{\infty} h(\tau, t + wT) \cdot c(t - \tau) \exp\left(\frac{-j2\pi w\tau}{T_s}\right) d\tau. \tag{12}$$

The received signal of the frequency domain is expressed as

$$\tilde{r}(v, w) \simeq \sqrt{\frac{2P_t}{N_s}} H(v, w) d(v, w) + n(v, w), \tag{13}$$

where $H(v, w)$ is channel matrix. As shown in Fig. 1, signals obtained are processed in the reverse order of the transmitter, and the hard decision is made [11], [40]. As can be seen from (13), the received signal is greatly affected by the channel matrix $H(v, w)$ and $n(v, w)$. In a fading environment, rapid power attenuation occurs, and the values of the channel matrix change significantly. This is reflected significantly in sequence data and spectrograms, making estimation difficult. Also, when the SNR defined by P_t and $n(t)$ becomes small, it greatly affects the spectrogram.

IV. PROPOSED MULTIMODAL NETWORK FOR JOINT SNR AND RICIAN K-FACTOR ESTIMATION

As mentioned in the previous section II, many ML-based communication environment estimation schemes exist. Nevertheless, most focus on a single modality, and utilizing multiple modalities is key to achieving higher accuracy and low latency. This paper proposes a joint SNR and K-factor estimation scheme by the multimodal network with sequence and spectrogram input. The proposed network consists of two main phases: an extraction phase and a fusion phase. A series of training steps allows for effective learning even with limited datasets.

Sequence data and spectrograms are input and trained separately in the extraction phase. Sequence data is good at extracting features related to specific power levels. The spectrogram can effectively capture features related to power fluctuations caused by phenomena such as fading. This intensive, individualized training allows each modality to extract the features it is best at. While in the fusion phase, feature maps from which each data is best at extracting features are fused. Estimation accuracy is highly dependent on the architecture of this fusion network [29]. This paper examines the most suitable timing to perform the fusion process to optimize performance.

A. SPECTROGRAM IMAGE

In ML applications in wireless communications, waveform formats such as the spectrum, constellation, and spectrogram are generally effective when utilizing CNNs, which specialize in feature extraction in images [11], [14], [17]. Our previous works [11], [41] have shown that spectrogram images, which are three-dimensional data (time, frequency, and power), are suitable for SNR estimation. Therefore, in this paper, spectrogram images shown in Fig. 2 are used as the image input for the proposed network. Spectrogram images are obtained by performing STFT on sequence data using a window function. The window function for spectrogram $\xi(t)$ is defined as

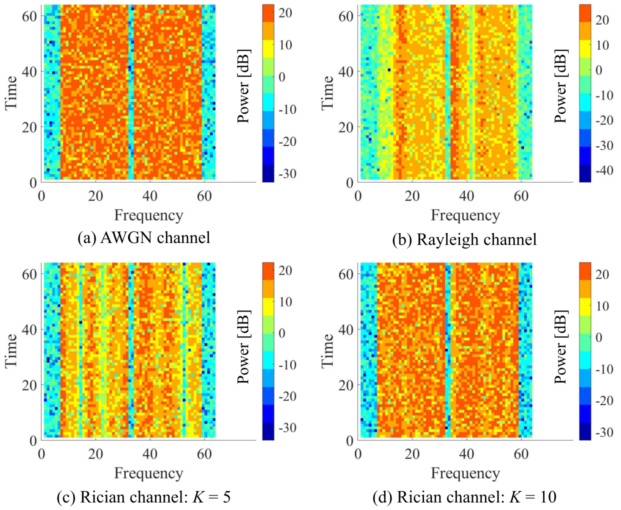
$$\xi(t) = \begin{cases} 1 & (0 \leq t \leq T) \\ 0 & \text{otherwise.} \end{cases} \tag{14}$$

With this function, we can obtain the signal with complex elements R_{spg} as

$$\begin{aligned} R_{spg}(v, w) &= \int_{-\infty}^{\infty} r(\tau)\xi(\tau - w)e^{-j2\pi vw}dt \\ &= \int_{0}^{T} r(\tau)e^{-j2\pi vw}dt. \end{aligned} \tag{15}$$

Therefore, spectrogram P_{spg} is represented as

$$P_{spg}(v, w) = \left|R_{spg}(v, w)\right|^2. \tag{16}$$



(a) AWGN channel

(b) Rayleigh channel

(c) Rician channel: $K = 5$

(d) Rician channel: $K = 10$

FIGURE 2. Spectrogram images of OFDM signal acquired in various channels at SNR 20 dB.

Fig. 2 shows the difference in spectrogram images of OFDM signals affected by a fading channel with an SNR of 20 dB. Fig. 2(a) shows the case of the AWGN channel. The power of the data portion exhibits minimal fluctuation, and there is a distinct difference between the power of the data portion and the power of the portion without data. Fig. 2(b) shows the case of the Rayleigh channel. The received power is not as apparent as Fig. 2(a) due to significant signal distortion from Rayleigh fading and multipath delay. Figs. 2(c) and (d) are derived from Rician channels. Fig. 2(d) with $K = 10$ is more similar to the AWGN image than Fig. 2(c) with $K = 5$. This indicates that with higher values of K, the channel becomes closer to flat fading, as shown in (8). Thus, feature extraction using CNN is possible because fading features clearly appear as power fluctuations in the spectrogram.

B. CNN: FEATURE EXTRACTION

In general, CNNs specialize in image recognition and comprise several convolutional layers, pooling layers, and one or more fully connected layers. In the convolutional layer, features are extracted from the input data using various filters with different sizes. The filters are smaller matrices compared to the input data. They slide over the input data, performing element-wise multiplication with the filter, and the results are summed up to generate a new feature map [42]. The activation functions and pooling layers enable complex expressions and efficient computation by reducing the feature map.

The output of the $(l+1)$-th convolutional layer, denoted as $Y^{l+1}_{o,c',m',n'}$, is calculated as follows

$$Y^{l+1}_{o,c',m',n'} = \sum_{c=1}^{C}\sum_{p=1}^{P}\sum_{q=1}^{Q} W^{l+1}_{c',c,p,q} \cdot X^{l}_{o,c,m,n} + B^{l+1}_{c'}, \tag{17}$$

where o ($o = 1, 2, \cdots, O$) represents the mini-batch index, c ($c = 1, 2, \cdots, C$) and c' ($c' = 1, 2, \cdots, C'$) represent the

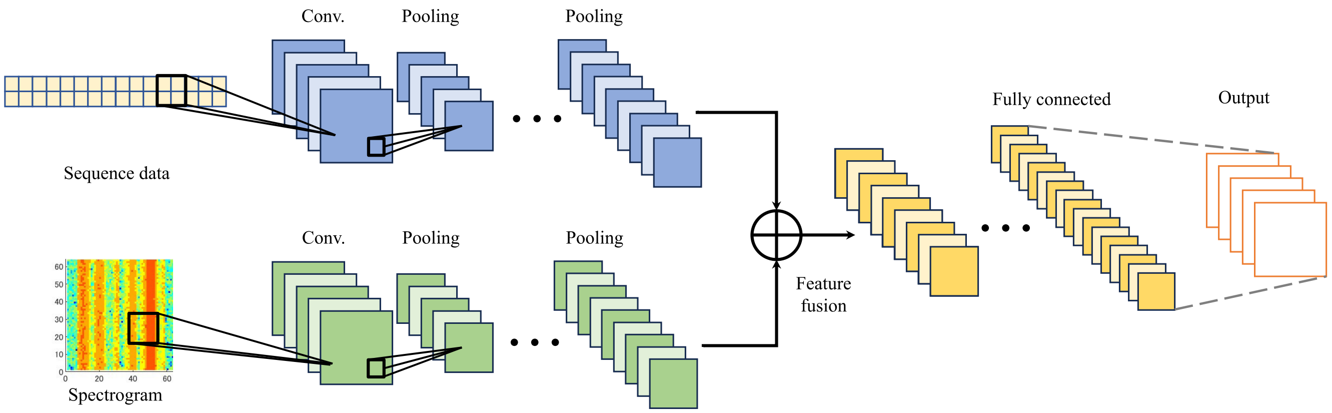


FIGURE 3. **Multimodal network structure of the proposed scheme.**

input and output channels. m' and n' denote the height and width of the output data. p and q represent the height and width of the filter. $X^{l}_{o,c,m,n}$ indicates the output at position (m, n) of the c-th channel in the l-th layer for the o-th mini-batch. $W^{l+1}_{c'c,p,q}$ corresponds to the weight at position (p, q) of the filter between the c-th and c'-th channels in the $(l+1)$-th layer. $B^{l+1}_{c'}$ represents the bias of the c'-th channel in the $(l+1)$-th layer. The activation function and pooling used in this paper is the ReLU function and max-pooling, which is defined as

$$X^{l+1}_{o,c,m,n} = \begin{cases} Y^{l+1}_{o,c',m',n'} & (Y^{l+1}_{o,c',m',n'} > 0) \\ 0 & \text{otherwise}, \end{cases} \tag{18}$$

$$Y^{l+1}_{o,c',m',n'} = \max_{p,q} \left(X^{l}_{o,c,m,n} \right), \tag{19}$$

where $p \in [1, P]$, $q \in [1, Q]$. Through (17) to (19), the feature extraction of the CNN is performed [11]. This series of layers is defined as a training block.

C. NETWORK TRAINING

The overview of the proposed network is shown in Fig. 3. In the proposed scheme, two different CNNs are used to extract features from sequence data and spectrograms, respectively, and finally combined to estimate SNR and K-factor at the same time. Therefore, we need to select the optimal network structure, and three patterns are considered here: early fusion, mid-term fusion, and late fusion. In early fusion, sequence and image data are passed through a training block once each, then fused and estimated after two training blocks. In mid-term fusion, the two modalities go through two training blocks, feature fusion is performed, and estimation is performed after one training block. In late fusion, the two modalities experience three training blocks each, are combined by fully connected layers, and are estimated. Each of the fusion architecture layers and their sizes are shown in detail in Fig. 4.

In general, CNNs require a real vector as input [43]. Therefore, received signal sequences, complex data, cannot be input directly. The sequence data proposed in this paper is divided into the real and imaginary parts of the (13) signal for training. During the training process, the input sequence data is given as

$$r_{train} = \left(\Re\left(\tilde{r}\right), \Im\left(\tilde{r}\right) \right), \tag{20}$$

Here, the output of the previously defined training block in i-th layer as

$$\mathcal{O}_{training} = f^{i}(\cdot). \tag{21}$$

Sequence and spectrogram image data are separately trained in the extraction phase. We denote the input feature maps of the sequence data and spectrogram data as X_{sq} and X_{sp}, respectively. They are individually trained to the j-th layer. The output at the j-th layer is given by

$$Q^{j} = f^{j}(X_{sq}), \tag{22}$$

$$P^{j} = f^{j}(X_{sp}). \tag{23}$$

The output of the fusion operation in the j-th layer is then as follows

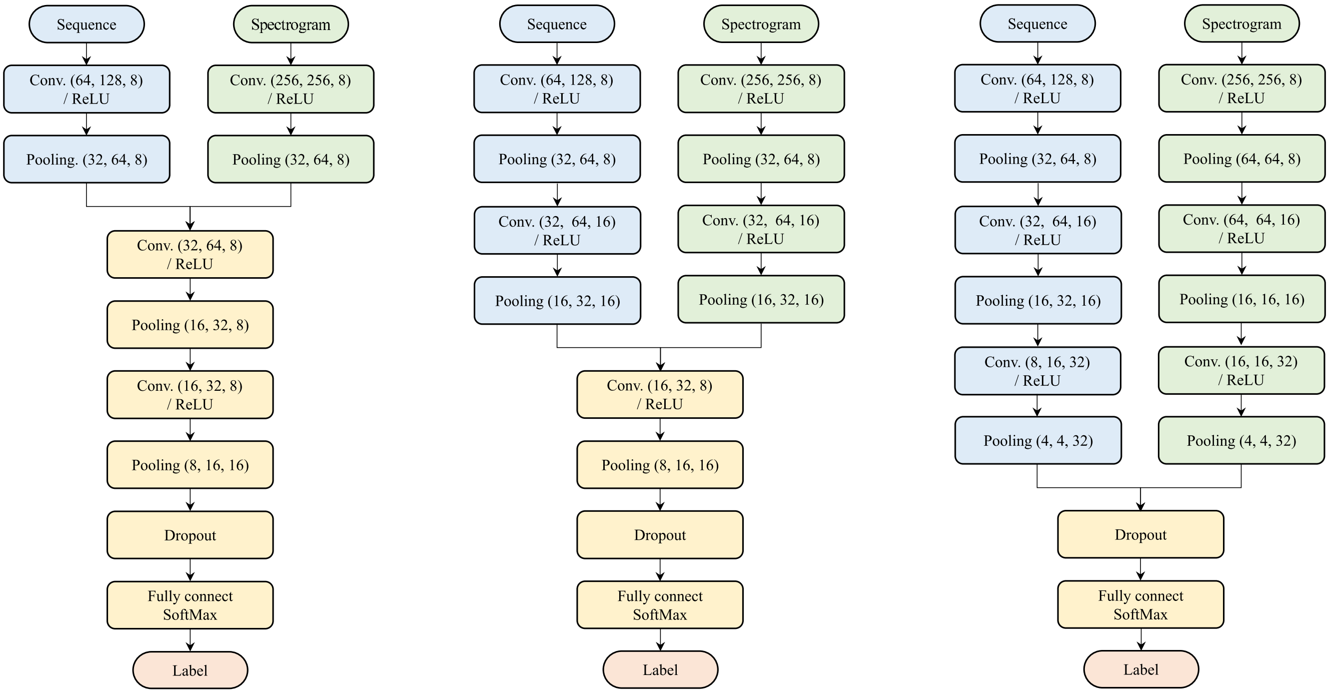
$$R^{j} = \mathcal{A}[Q^{j}, P^{j}], \tag{24}$$

where R indicates the output, and $\mathcal{A}$ means element-wise addition [44]. This fused feature map R^{j} is retrained in both Early and Mid-term Fusion. After that, these feature maps go through the dropout layer. The dropout layer prevents overlearning by overwriting input features with random zeros with arbitrary probability as

$$D^{k} \sim \mathcal{B}(p), \tag{25}$$

$$\tilde{R} = D^{k} * R, \tag{26}$$

where $\mathcal{B}(p)$ is Bernoulli distribution. D^{k} is a vector of independent Bernoulli random variables, each with probability $p = 1$, and $(*)$ means element-wise product [45]. In this paper, 50% of the input features are set to 0. The features trained through the fully connected and softmax layers are probabilistically classified into i classes. The process carried out by the softmax layer involves the inputs $x_i \in$



(a) Early fusion

(b) Mid-term fusion

(c) Late fusion

FIGURE 4. **Architecture of a multimodal network, where (a) is early fusion, (b) is mid-term fusion, and (c) is late fusion.**

$x_1, x_2, \cdots x_N$, which are the outputs of the fully connected layer. The softmax operations represent it

$$Y_{class} = \frac{e^{x_i}}{\sum_{j=1}^{N} e^{x_j}} = \exp\left(x_i - \ln\left(\sum_{j=1}^{N} e^{x_j}\right)\right), \tag{27}$$

where $i = 1, 2, \cdots, N$ [46]. The communication environment is estimated by this classification result. After offline training is completed, the trained network can be used online to estimate and classify parameters such as SNR and K-factor to determine the communication channel environment.

Sequence data can take negative values, yet it does not have color information. On the other hand, as shown in (16), while spectrograms cannot take negative values and suffer from conversion errors, they are suitable for CNNs because the power distortion related to SNR and K-factor clearly appears as a feature. By combining modalities with different features at the same time, the proposed network can train to compensate for each other, even in limited training data. As a result, it is possible to extract features from limited training data with a lightweight CNN, and stable estimation can be realized.

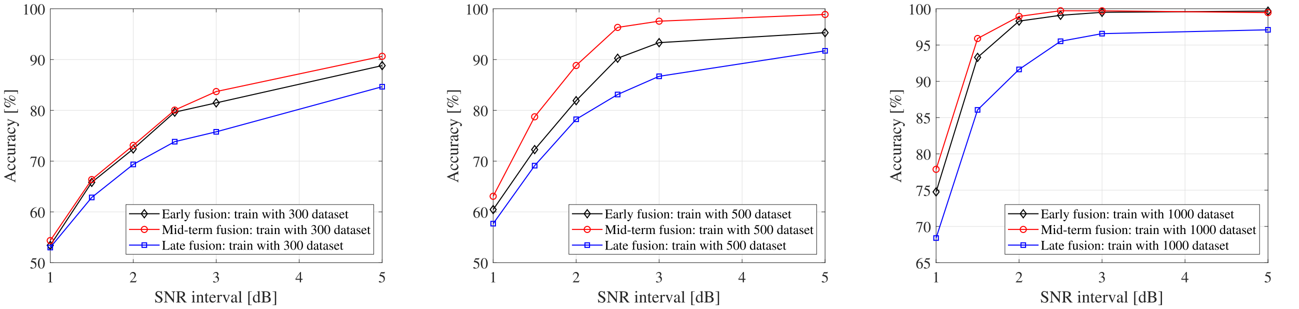
V. NUMERICAL RESULTS

In this section, we evaluate the estimation accuracy of the proposed multimodal network compared to various benchmark schemes. The simulation environment used was MATLAB 2022a, Intel(R) Core(TM) i9-10900K CPU, NVIDIA RTX 3090 GPU. We generate wireless signal data using parameters outlined in Table 1 by Monte Carlo simulations. During the network's offline training phase, we allocate 80% of the training dataset for actual training and 20% for validation. We employ the Adam optimization algorithm [47], conduct training over 50 epochs, and set the initial learning rate at 0.0001.

TABLE 1. **Simulation parameters.**

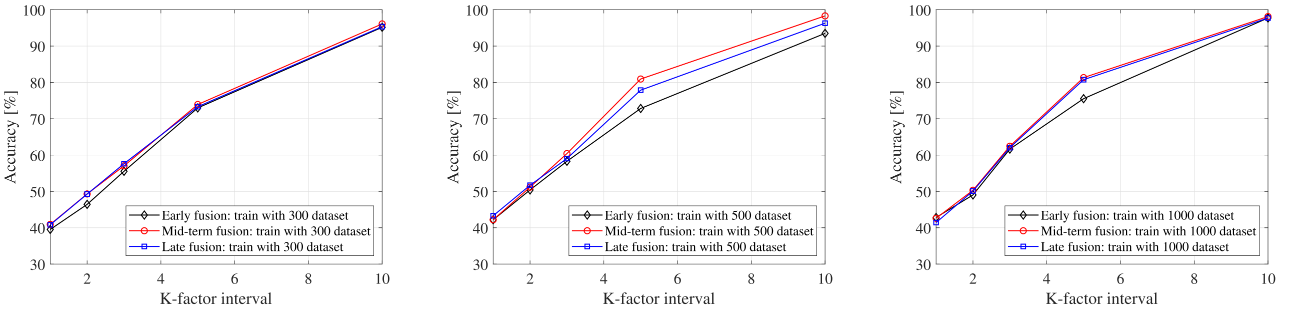
Parameter	Value
IFFT size	64
Number of pilot symbols	2
Number of subcarriers	50
Number of information symbols	32
Length of guard interval	16
Modulation	64QAM
Bandwidth	20 MHz
Doppler shift	1 Hz
Number of paths	15
Fading	Rayleigh / Rician fading

This paper estimates SNR or K-factor by classifying based on either SNR or K-factor intervals. For example, we classify SNR values at 3 dB intervals into five classes, starting from 0 dB as {0, 3, 6, 9, 12} dB. Also, K-factor estimation is conducted similarly, classifying the K-factor



(a) Training on a dataset with 300 samples. (b) Training on a dataset with 500 samples. (c) Training on a dataset with 1000 samples.

FIGURE 5. **Accuracy of SNR estimation in different multimodal network structures: early, mid, late fusion.**



(a) Training on a dataset with 300 samples. (b) Training on a dataset with 500 samples. (c) Training on a dataset with 1000 samples.

FIGURE 6. **Accuracy of K-factor estimation in different multimodal network structures: early, mid, late fusion.**

prepared according to the K-factor interval, starting from 0. For instance, if there are three classes and the K-factor interval is 5, the estimated K-factor is classified into one of the classes $\{0, 5, 10\}$. Unless otherwise specified, we classify SNR into five classes and K-factor into three classes. Here, classification accuracy is defined as the ratio of the number of test data that can be correctly classified to the number of total test data.

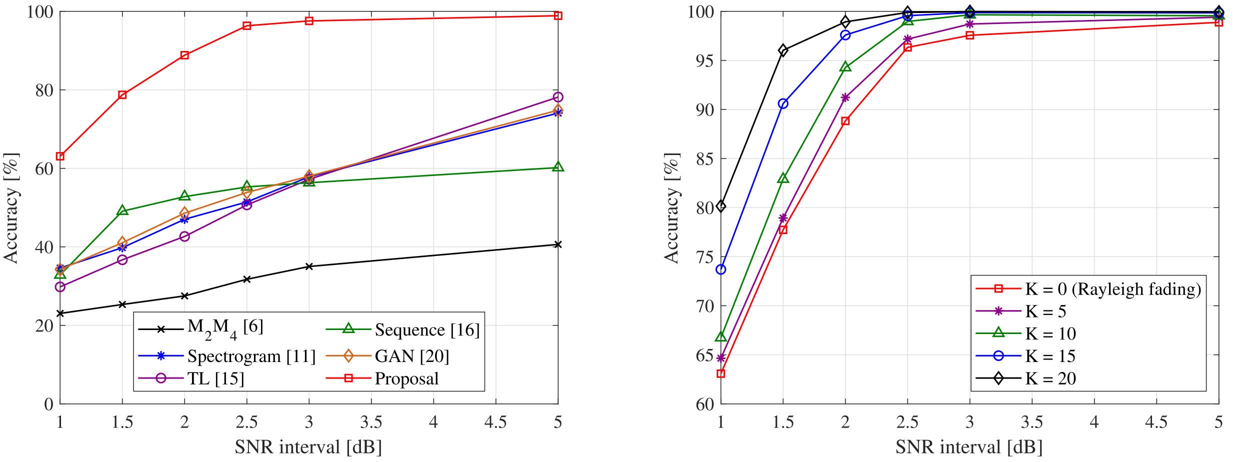
A. OPTIMAL STRUCTURE DETERMINATION

We compare the proposed multimodal network structures shown in Fig. 4 to determine which are optimal for SNR and K-factor estimation. Here, each network is respectively trained on a data set consisting of 300, 500, and 1000 packet samples for each class. The architecture is determined by comparing the estimation results for each parameter on the test dataset, which differs from the training dataset. At this time, the SNR estimation is evaluated in a Rayleigh fading environment, and the K-factor estimation is evaluated in a Rician fading environment with an SNR of 20 dB.

The results of SNR estimation are shown in Figs. 5(a)–(c). It can be observed that the mid-term fusion architecture has the highest estimation accuracy among the three architectures. Mid-term fusion architecture effectively extracts features from sequence and spectrogram data, calibrates by retraining after fusion, and achieves the highest SNR estimation accuracy for each training dataset. On the contrary, in late fusion architecture, the potential of the proposed multimodal learning is not fully exploited, and the sequence and spectrogram data are not well calibrated since they are not retrained through training blocks after fusion. Early fusion architecture requires more training from each modality, resulting in feature vector redundancy.

K-factor estimation results are shown in Figs. 6(a)–(c). It is clear that, unlike SNR estimation, the estimation accuracy does not change significantly regardless of the network's structure when trained on the 300 datasets shown in Fig. 6(a). We also compare the estimation using the network trained on the 500 dataset shown in Fig. 6(b) with that using the network trained on the 1000 dataset shown in Fig. 6(c). The results of the mid-term fusion architecture do not change significantly when the training dataset is doubled, but the late fusion approaches the mid-term fusion by 2.49% when the K-factor interval is 5. On the other hand, early fusion also improves estimation accuracy as the training dataset becomes richer.

These results indicate that the ideal architecture differs based on the estimation target. For SNR estimation, achieving higher accuracy is fused to an earlier stage, while in K-factor estimation, superior accuracy is associated with a later stage of fusion. This indicates the need to adjust the architecture according to the target estimation task. The



(a) Comparison of the proposed scheme with conventional schemes.

(b) Evaluation of the proposed scheme with varying K-factors.

FIGURE 7. Accuracy of SNR estimation.

TABLE 2. Impact of the amount of training dataset on ML-Based SNR estimation [%].

Scheme\Dataset	100	300	500	1000	2000	5000
Spectrogram [11]	51.80	54.94	57.85	58.20	60.31	62.72
TL [15]	54.26	56.70	57.20	57.65	59.07	61.32
Sequence [16]	33.55	44.10	56.34	60.61	75.69	91.32
GAN [20]	49.42	55.03	58.02	58.86	60.82	62.77
Proposal	**55.66**	**83.70**	**97.56**	**99.72**	**99.74**	**99.82**

influence of architecture on the estimation target occurs when one modality cannot train effectively due to a negative impact on the overall estimation. Mid-term fusion, which showed the most stable estimation accuracy for both SNR and K-factor estimation, will be used as multimodal in subsequent evaluations.

B. COMPARISON RESULTS IN SNR ESTIMATION

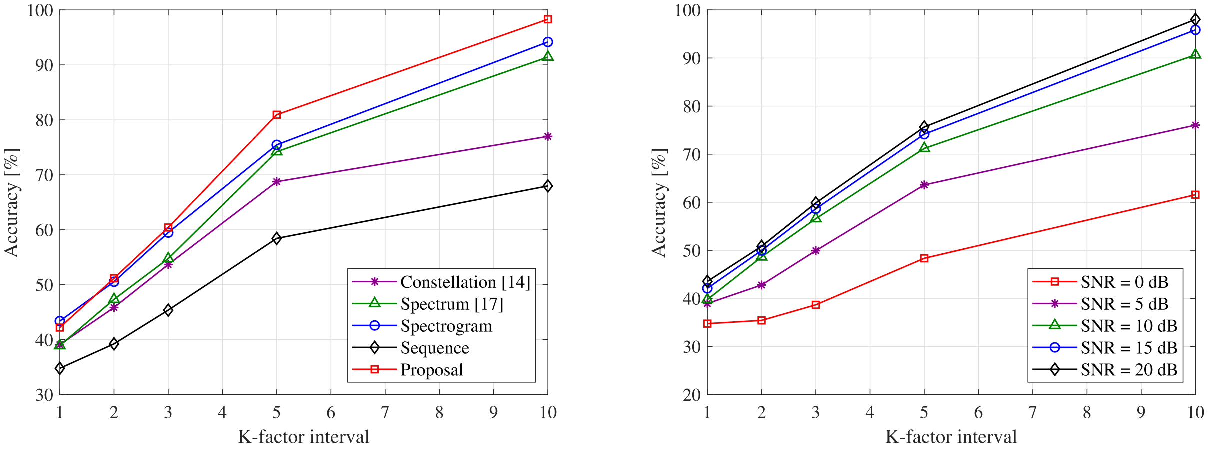
To validate the effectiveness of the proposed scheme, we compare SNR estimation accuracy with that of conventional schemes. Here, we evaluate the estimation accuracy by classifying five SNR classes in a Rayleigh fading environment. As conventional schemes, moment-based SNR estimation [6], CNN only with spectrogram images [11], TL using constellations [15], CNN-LSTM only with sequence data [16], and data augmentation with GAN [20] are compared to the proposed scheme used in the previous study. A limited dataset of 500 data for each SNR is used as the training dataset in all ML-based schemes. For a fair comparison, the proposed approach assumes SNR estimation only.

Fig. 7(a) shows the SNR estimation accuracy compared with various schemes. From this figure, the proposed multimodal estimation achieves the best accuracy of all schemes. M$_2$M$_4$, a non-ML-based scheme, has lower SNR estimation accuracy than any other ML-based schemes due to its vulnerability to noise. From the perspective of feature extraction by ML, comparing schemes that use sequence data and spectrogram images as input, sequence data provides better estimation accuracy in regions with narrow SNR intervals. In contrast, spectrogram images perform better in regions with wide SNR intervals. This indicates that the strength of feature extraction differs depending on the modality. The proposed multimodal learning network is based on sequence data and spectrogram. It increases feature diversity and compensates for the other modality even when features extracted from one modality do not effectively capture the model. This estimation scheme has proven effective in challenging feature extraction due to limited training data and unpredictable power fluctuations caused by Rayleigh fading. The strategy of increasing feature diversity results in stable estimation and maximizes accuracy.

TABLE 3. Impact of the number of classes on the proposed SNR estimation [%].

SNR: Number of classes	4	5	6	7
Accuracy	99.58	98.88	95.2	90.91

Table 2 illustrates the impact of the amount of dataset size with various ML-based SNR estimation schemes. Here, we assume the five classifications with a 3 dB SNR interval under a Rayleigh fading channel and training datasets of various sizes, including 100, 300, 500, 1000, 2000, and 5000, are prepared for offline training. The table reveals that the accuracy of all estimation schemes consistently improves as the dataset size increases. It is observed that image-based schemes yield higher accuracy for smaller datasets. In contrast, with larger datasets, the sequence data-based scheme becomes more accurate. The proposed multimodal



(a) Comparison of the proposed scheme with conventional schemes.

(b) Evaluation of the proposed scheme with varying SNR.

FIGURE 8. Accuracy of K-factor estimation.

TABLE 4. Impact of the amount of training dataset on ML-based K-factor estimation [%].

Scheme\Dataset	100	300	500	1000	2000	5000
Constellation [14]	41.58	50.41	53.67	54.72	56.5	59.47
Spectrum [17]	42.03	51.49	54.75	55.89	56.32	58.77
Spectrogram	46.9	54.96	59.48	60.53	60.88	61.56
Sequence	38.42	42.65	45.37	55.1	58.22	60.42
Proposal	**49.13**	**57.63**	**60.41**	**62.04**	**62.23**	**64.59**

estimation approach, which leverages both sequence and image data, surpasses other schemes in terms of accuracy, highlighting the advantage of combining data modalities for enhanced performance.

Table 3 presents the individual SNR estimation result using the proposed network. For SNR estimation, we selected Rayleigh fading as the channel model, setting the SNR interval to 5 dB. The result in Table 3 indicates that for the individual SNR estimation, accuracy declines as the number of classes increases. However, accuracy remains above 90%, even with increased classes.

We evaluate SNR estimation accuracy and the impact of channel K-factor. Fig. 7(b) shows the estimation accuracy in a Rician fading environment in case of changing the K-factor. As shown in this figure, SNR estimation accuracy improves as K increases. In particular, when K is 20, 80.15% accuracy is obtained even when the SNR interval is 1 dB. In SNR estimation, obtaining high accuracy is difficult due to fluctuation caused by sudden changes in received power due to fading. As the K-factor increases, the channel state becomes closer to flat fading, as shown in Fig. 2, and the fluctuations due to received power fluctuations become reduced. As a result, SNR estimation accuracy improves with increasing K-factor. When the SNR interval exceeds 2 dB, the estimation accuracy exceeds 88.83% in any K environment, confirming that the effect of the K-factor on SNR estimation is relatively small.

TABLE 5. Impact of the number of classes on the proposed K-Factor estimation [%].

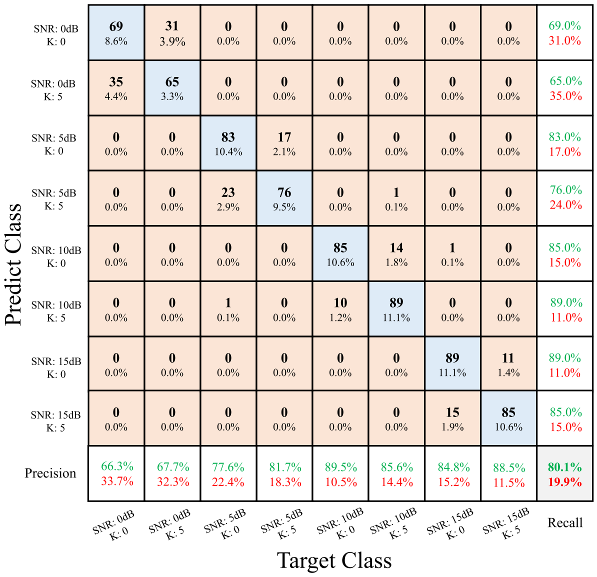
K-factor: Number of classes	2	3	4	5
Accuracy	87.3	80.93	73.32	67.91

C. COMPARISON RESULTS IN K-FACTOR ESTIMATION

We compare the proposed multimodal network with several benchmarks for K-factor estimation: ML using constellation [14], spectrum [17], sequence data, and spectrogram. Each network is trained on a limited dataset of 500 for each class. In this evaluation setup, we consider the classification problem of three K-factor patterns for an unknown input, and the SNR is kept at 20 dB. For a fair comparison, the proposed approach assumes K-factor estimation only.

Fig. 8(a) shows the comparison results of the K-factor estimation with various benchmarks. The proposed multimodal network can train two modalities as well as SNR estimation, and it is more specific than any of the other schemes, which confirms that it can improve estimation accuracy. On the other hand, unlike SNR estimation, K-factor estimation using ML with spectrograms is consistently more accurate than with sequence data. This indicates that the optimal modality depends on the classification target.

Table 4 illustrates the impact of varying training dataset sizes on the accuracy of ML-based K-factor estimations. The SNR is fixed at 20 dB, and the K-factor interval is three. Training dataset sizes of 100, 300, 500, 1000, 2000, and 5000 are prepared for offline training. The results confirm a trend similar to that observed from previous SNR estimation: increasing the size of the dataset improves accuracy across all schemes. However, in contrast to SNR estimation, the network trained with spectrograms consistently achieves higher estimation accuracy than those using sequence estimation, regardless of the dataset size. The difference, also confirmed by the evaluation in Fig. 8, suggests that the optimal modality



	SNR: 0dB K: 0	SNR: 0dB K: 5	SNR: 5dB K: 0	SNR: 5dB K: 5	SNR: 10dB K: 0	SNR: 10dB K: 5	SNR: 15dB K: 0	SNR: 15dB K: 5	Recall
SNR: 0dB K: 0	69 8.6%	31 3.9%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	69.0% 31.0%
SNR: 0dB K: 5	35 4.4%	65 3.3%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	65.0% 35.0%
SNR: 5dB K: 0	0 0.0%	0 0.0%	83 10.4%	17 2.1%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	83.0% 17.0%
SNR: 5dB K: 5	0 0.0%	0 0.0%	23 2.9%	76 9.5%	0 0.0%	1 0.1%	0 0.0%	0 0.0%	76.0% 24.0%
SNR: 10dB K: 0	0 0.0%	0 0.0%	0 0.0%	0 0.0%	85 10.6%	14 1.8%	1 0.1%	0 0.0%	85.0% 15.0%
SNR: 10dB K: 5	0 0.0%	0 0.0%	1 0.1%	0 0.0%	10 1.2%	89 11.1%	0 0.0%	0 0.0%	89.0% 11.0%
SNR: 15dB K: 0	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	89 11.1%	11 1.4%	89.0% 11.0%
SNR: 15dB K: 5	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	15 1.9%	85 10.6%	85.0% 15.0%
Precision	66.3% 33.7%	67.7% 32.3%	77.6% 22.4%	81.7% 18.3%	89.5% 10.5%	85.6% 14.4%	84.8% 15.2%	88.5% 11.5%	80.1% 19.9%

FIGURE 9. Confusion matrix when there are 4 classes of SNRs and 2 classes of K-factors.

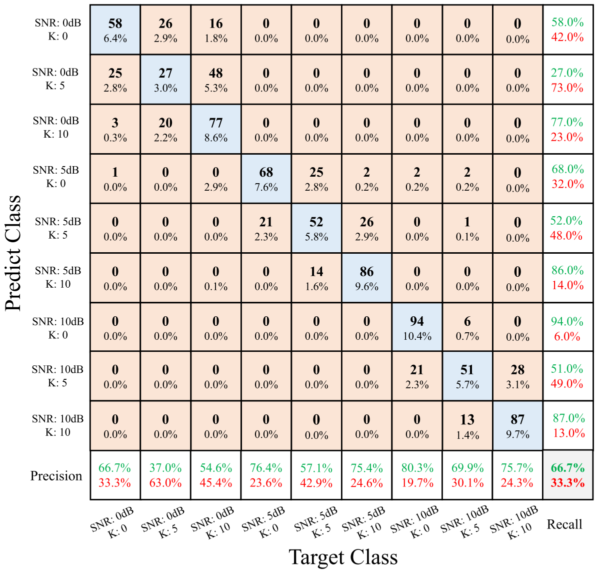
varies depending on the estimation target. Moreover, the proposed approach enhances accuracy by combining features extracted from sequences with those from spectrograms, yielding superior accuracy compared to other schemes.

Table 5 shows the K-factor estimation result using the proposed network. The SNR is fixed at 20 dB for K-factor estimation, and the K-factor interval is set to 5. The result in Table 5 represents that, similar to Table 3, accuracy declines as the number of classes increases, and estimation accuracy drops below 70% when the class count exceeds 7.

We evaluate the effect of SNR variation on K-factor estimation in the proposed network in Fig. 8(b). We assume environments with SNRs of 0 dB, 5 dB, 10 dB, 15 dB, and 20 dB. As in the previous sections, we limit the training of the network to 500 training data. As the SNR increases, the estimation accuracy of the K-factor improves. On the other hand, as the SNR becomes low, the estimation accuracy deteriorates significantly. This indicates that the effect of noise has a significant impact. For K-factor estimation, when the SNR is 0 dB, the K-factor estimation accuracy is 34.76% when the K-factor interval is 1, and 61.57% when the interval is 10. When the SNR is 20 dB, the accuracy is 43.54% when the K-factor interval is 1, and 98.03% when the interval is set to 10. From these results, it can be concluded that the influence of SNR on K-factor estimation is always significant. This is because multimodal training with small datasets is susceptible to unexpected noise due to the specificity of the estimation results.

D. JOINT ESTIMATION OF SNR AND K-FACTOR

Table 6 shows the joint estimation accuracy of SNR and K-factor, classifying SNR into 4 to 7 classes and K-factor into 2 to 5 classes. In the case of four-class SNR estimation, there is a 15.75% difference in accuracy between two and five classes of K-factor estimation. However, as the SNR class increases, this difference diminishes, with the largest classification of seven SNR classes showing a difference of 10.22%. This demonstrates that as the SNR class increases, the influence of K-factor estimation becomes weak, which means that the accuracy of SNR estimation has a significant impact on joint estimation. Overall, estimation accuracy decreases as the number of classes increases, although the decrease in estimation accuracy becomes smaller as the SNR class increases. The joint estimation accuracy deteriorates due to the increase in the number of classes compared to the individual estimation results in Tables 2 and 4. However, the proposed scheme is superior to conventional methods in estimation accuracy, confirming its effectiveness. Moreover, estimating SNR and K-factor jointly reduces training and computation costs compared to estimating each of them separately.



	SNR: 0dB K: 0	SNR: 0dB K: 5	SNR: 0dB K: 10	SNR: 5dB K: 0	SNR: 5dB K: 5	SNR: 5dB K: 10	SNR: 10dB K: 0	SNR: 10dB K: 5	SNR: 10dB K: 10	Recall
SNR: 0dB K: 0	58 6.4%	26 2.9%	16 1.8%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	58.0% 42.0%
SNR: 0dB K: 5	25 2.8%	27 3.0%	48 5.3%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	27.0% 73.0%
SNR: 0dB K: 10	3 0.3%	20 2.2%	77 8.6%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	77.0% 23.0%
SNR: 5dB K: 0	1 0.0%	0 0.0%	0 2.9%	68 7.6%	25 2.8%	2 0.2%	2 0.2%	2 0.2%	0 0.0%	68.0% 32.0%
SNR: 5dB K: 5	0 0.0%	0 0.0%	0 0.0%	21 2.3%	52 5.8%	26 2.9%	0 0.0%	1 0.1%	0 0.0%	52.0% 48.0%
SNR: 5dB K: 10	0 0.0%	0 0.0%	0 0.1%	0 0.0%	14 1.6%	86 9.6%	0 0.0%	0 0.0%	0 0.0%	86.0% 14.0%
SNR: 10dB K: 0	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	94 10.4%	6 0.7%	0 0.0%	94.0% 6.0%
SNR: 10dB K: 5	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	21 2.3%	51 5.7%	28 3.1%	51.0% 49.0%
SNR: 10dB K: 10	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	13 1.4%	87 9.7%	87.0% 13.0%
Precision	66.7% 33.3%	37.0% 63.0%	54.6% 45.4%	76.4% 23.6%	57.1% 42.9%	75.4% 24.6%	80.3% 19.7%	69.9% 30.1%	75.7% 24.3%	66.7% 33.3%

FIGURE 10. Confusion matrix when there are 3 classes of SNRs and 3 classes of K-factors.

TABLE 6. SNR and K-Factor joint estimation accuracy [%].

K-factor\SNR classes	4	5	6	7
2	80.10	76.97	75.05	72.37
3	74.37	73.05	71.35	69.33
4	67.51	66.02	65.63	65.10
5	64.35	64.14	62.97	62.15

We discuss the confusion matrix regarding estimation accuracy to examine further the effectiveness of the proposed joint SNR and K-factor estimation. Fig. 9 shows the confusion matrix when SNR is classified into four classes and K-factor into two classes, and Fig. 10 shows the confusion matrix for SNR classified into three classes and K-factor into three classes. From Fig. 9, it can be confirmed that as the

TABLE 7. **Computational complexity for SNR estimation.**

Scheme	Processing time [ms]	FLOPs	Accuracy [%]
M_2M_4 [6]	3.70×10^{-3}	39.0 K	23.04
Spectrogram [11]	1.78	33.1 M	34.60
TL [15]	5.51	30.9 G	29.80
Sequence [16]	0.82	10.7 M	32.83
GAN [20]	1.80	33.1 M	34.35
Proposal	**2.24**	**40.2 M**	**63.08**

SNR increases, the estimation accuracy increases. This is consistent with Fig 7(b). Fig. 10 shows that when classifying the K-factor as 5, the estimation accuracy is low at all SNRs. This significantly affects the overall decrease in estimation accuracy compared to Fig. 9. This result suggests that it may not be possible to effectively classify when the K-factor is 5 and when it is 10.

E. COMPUTATIONAL COMPLEXITY

We evaluate the computational complexity of the proposed and benchmark schemes. Here, we consider the SNR estimation of five classifications with 1 dB SNR intervals. Table 7 comprehensively compares different schemes, focusing on processing time, floating point operations (FLOPs), and accuracy. Processing time is defined as the time required to compute SNR from a test signal in non-ML-based approaches and the time required to process a test signal through a trained ML model in ML-based approaches. FLOPs, a metric of computational complexity, is determined by the total number of multiply-adds, which is the most used definition [48], [49]. According to this table, non-ML-based approaches minimize processing time. On the other hand, it can also be confirmed that the accuracy is the poorest. Although ML-based approaches have high estimation accuracy, their FLOPs tend to be large. Meanwhile, the proposed scheme achieves the best accuracy while keeping the increase in processing time and FLOPs at the same level as other ML-based schemes.

VI. CONCLUSION

In this paper, we proposed the joint SNR and K-factor estimation using a multimodal network. In the proposed scheme, the received signal is converted into two modalities: sequence data and spectrogram image. Then, the SNR and K-factor are estimated by taking two forms of sequence data and spectrogram as inputs, extracting features from each input respectively, and finally combining them. Feature diversity is obtained by converting a single received signal into two modalities for feature extraction, thus achieving sufficient performance even from a limited dataset. In the simulation, we have compared SNR and K-factor estimation with conventional schemes regarding estimation accuracy and computational complexity. The results have demonstrated that the proposed scheme achieves the highest accuracy and allows simultaneous estimation of SNR and K-factor at reasonable processing speeds.

REFERENCES

[1] S. Falahati, A. Svensson, T. Ekman, and M. Sternad, "Adaptive modulation systems for predicted wireless channels," *IEEE Trans. Commun.*, vol. 52, no. 2, pp. 307–316, Feb. 2004.

[2] Z. Yang, Z. Ding, P. Fan, and N. Al-Dhahir, "A general power allocation scheme to guarantee quality of service in downlink and uplink NOMA systems," *IEEE Trans. Wireless Commun.*, vol. 15, no. 11, pp. 7244–7257, Nov. 2016.

[3] S. Shao, M. Nazzal, A. Khreishah, and M. Ayyash, "Self-optimizing data offloading in mobile heterogeneous radio-optical networks: A deep reinforcement learning approach," *IEEE Netw.*, vol. 36, no. 2, pp. 100–106, Mar. 2022.

[4] Y.-N. Lee, A. Ashikhmin, and J.-T. Chen, "Impact of soft channel construction on iterative channel estimation and data decoding for multicarrier systems," *IEEE Trans. Wireless Commun.*, vol. 7, no. 7, pp. 2762–2770, Jul. 2008.

[5] H. Xu and H. Zheng, "The simple SNR estimation algorithms for MPSK signals," in *Proc. 7th Int. Conf. Signal Process. (ICSP)*, Aug. 2004, pp. 1781–1785.

[6] D. R. Pauluzzi and N. C. Beaulieu, "A comparison of SNR estimation techniques for the AWGN channel," *IEEE Trans. Commun.*, vol. 48, no. 10, pp. 1681–1691, Oct. 2000.

[7] L. J. Greenstein, D. G. Michelson, and V. Erceg, "Moment-method estimation of the Ricean K-factor," *IEEE Commun. Lett.*, vol. 3, no. 6, pp. 175–176, Jun. 1999.

[8] A. Naimi and G. Azemi, "Moment-based Ricean K-factor estimation in the presence of shadowing," in *Proc. 9th Int. Symp. Signal Process. Appl.*, Feb. 2007, pp. 1–4.

[9] A. Das and B. D. Rao, "SNR and noise variance estimation for MIMO systems," *IEEE Trans. Signal Process.*, vol. 60, no. 8, pp. 3929–3941, Aug. 2012.

[10] A. Wiesel, J. Goldberg, and H. Messer-Yaron, "SNR estimation in time-varying fading channels," *IEEE Trans. Commun.*, vol. 54, no. 5, pp. 841–848, May 2006.

[11] S. Kojima, K. Maruta, Y. Feng, C.-J. Ahn, and V. Tarokh, "CNN-based joint SNR and Doppler shift classification using spectrogram images for adaptive modulation and coding," *IEEE Trans. Commun.*, vol. 69, no. 8, pp. 5152–5167, Aug. 2021.

[12] A. Doukas and G. Kalivas, "Rician K factor estimation for wireless communication systems," in *Proc. Int. Conf. Wireless Mobile Commun. (ICWMC)*, Jul. 2006, p. 69.

[13] Y. Chen and N. Beaulieu, "Estimation of Ricean K parameter and local average SNR from noisy correlated channel samples," *IEEE Trans. Wireless Commun.*, vol. 6, no. 2, pp. 640–648, Feb. 2007.

[14] G. Lu, Q. Zhang, X. Zhang, F. Shen, and F. Qin, "CNN based Rician K factor estimation for non-stationary industrial fading channel," in *Proc. IEEE Global Conf. Signal Inf. Process. (GlobalSIP)*, Nov. 2018, pp. 594–598.

[15] X. Xie, S. Peng, and X. Yang, "Deep learning-based signal-to-noise ratio estimation using constellation diagrams," *Mobile Inf. Syst.*, vol. 2020, pp. 1–9, Nov. 2020.

[16] T. Ngo, B. Kelley, and P. Rad, "Deep learning based prediction of signal-to-noise ratio (SNR) for LTE and 5G systems," in *Proc. 8th Int. Conf. Wireless Netw. Mobile Commun. (WINCOM)*, Oct. 2020, pp. 1–6.

[17] M. Alymani, M. H. Alhazmi, A. Almarhabi, H. Alhazmi, A. Samarkandi, and Y. Yao, "Rician K-factor estimation using deep learning," in *Proc. 29th Wireless Opt. Commun. Conf. (WOCC)*, Newark, NJ, USA, May 2020, pp. 1–4.

[18] A. P. Hermawan, R. R. Ginanjar, D. Kim, and J. Lee, "CNN-based automatic modulation classification for beyond 5G communications," *IEEE Commun. Lett.*, vol. 24, no. 5, pp. 1038–1041, May 2020.

[19] K. Bu, Y. He, X. Jing, and J. Han, "Adversarial transfer learning for deep learning based automatic modulation classification," *IEEE Signal Process. Lett.*, vol. 27, pp. 880–884, 2020.

[20] Z. Tang, M. Tao, J. Su, Y. Gong, Y. Fan, and T. Li, "Data augmentation for signal modulation classification using generative adverse network," in *Proc. IEEE 4th Int. Conf. Electron. Inf. Commun. Technol. (ICEICT)*, Aug. 2021, pp. 450–453.

[21] B. Tang, Y. Tu, Z. Zhang, and Y. Lin, "Digital signal modulation classification with data augmentation using generative adversarial nets in cognitive radio networks," *IEEE Access*, vol. 6, pp. 15713–15722, 2018.

[22] M. Wang, Y. Lin, Q. Tian, and G. Si, "Transfer learning promotes 6G wireless communications: Recent advances and future challenges," *IEEE Trans. Rel.*, vol. 70, no. 2, pp. 790–807, Jun. 2021.

[23] I. Goodfellow et al., "Generative adversarial nets," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 27, Dec. 2014, pp. 2672–2680.

[24] V. Kushwaha and G. C. Nandi, "Study of prevention of mode collapse in generative adversarial network (GAN)," in *Proc. IEEE 4th Conf. Inf. Commun. Technol. (CICT)*, Dec. 2020, pp. 1–6.

[25] S. Kojima, K. Shima, K. Maruta, and C. Ahn, "K-factor estimation based on spectrogram images by convolutional neural network," in *Proc. Int. Conf. Emerg. Technol. Commun. (ICETC)*, Nov. 2020, pp. 1–4.

[26] I. Bousnina, M. B. B. Salah, A. Samet, and I. Dayoub, "Ricean K-factor and SNR estimation for M-PSK modulated signals using the fourth-order cross-moments matrix," *IEEE Commun. Lett.*, vol. 16, no. 8, pp. 1236–1239, Aug. 2012.

[27] S. Kojima, K. Maruta, and C.-J. Ahn, "Throughput maximization by adaptive switching with modulation coding scheme and frequency symbol spreading," *J. Commun. Softw. Syst.*, vol. 14, no. 4, pp. 332–339, Nov. 2018.

[28] J. Kirchner, A. Heberle, and W. Lwe, "Classification vs. regression-machine learning approaches for service recommendation based on measured consumer experiences," in *Proc. IEEE World Congr. Serv.*, Jun. 2015, pp. 278–285.

[29] D. Ramachandram and G. W. Taylor, "Deep multimodal learning: A survey on recent advances and trends," *IEEE Signal Process. Mag.*, vol. 34, no. 6, pp. 96–108, Nov. 2017.

[30] W. Zhao, Y. Cheng, M. Xiang, M. Tang, Y. Qin, and S. Fu, "Nonlinear SNR estimation based on the data augmentation-assisted DNN with a small-scale dataset," *Opt. Exp.*, vol. 30, no. 22, pp. 39725–39735, Oct. 2022.

[31] Z. Wang, A. Yang, P. Guo, and P. He, "OSNR and nonlinear noise power estimation for optical fiber communication systems using LSTM based deep learning technique," *Opt. Exp.*, vol. 26, no. 16, pp. 21346–21357, Aug. 2018.

[32] H. Eghbal-Zadeh, B. Lehner, M. Dorfer, and G. Widmer, "CP-JKU submissions for DCASE-2016: A hybrid approach using binaural I-vectors and deep convolutional neural networks," in *Proc. DCASE Challenge*, Aug. 2016, pp. 5024–5028.

[33] Y. Yin, R. R. Shah, and R. Zimmermann, "Learning and fusing multimodal deep features for acoustic scene categorization," in *Proc. 26th ACM Int. Conf. Multimedia*, Oct. 2018, pp. 1892–1900.

[34] Y. Yang, F. Gao, C. Xing, J. An, and A. Alkhateeb, "Deep multimodal learning: Merging sensory data for massive MIMO channel prediction," *IEEE J. Sel. Areas Commun.*, vol. 39, no. 7, pp. 1885–1898, Jul. 2021.

[35] U. Challita, A. Ferdowsi, M. Chen, and W. Saad, "Machine learning for wireless connectivity and security of cellular-connected UAVs," *IEEE Wireless Commun.*, vol. 26, no. 1, pp. 28–35, Feb. 2019.

[36] P. Qi, X. Zhou, S. Zheng, and Z. Li, "Automatic modulation classification based on deep residual networks with multimodal information," *IEEE Trans. Cogn. Commun. Netw.*, vol. 7, no. 1, pp. 21–33, Mar. 2021.

[37] Z. Zhang, C. Wang, C. Gan, S. Sun, and M. Wang, "Automatic modulation classification using convolutional neural network with features fusion of SPWVD and BJD," *IEEE Trans. Signal Inf. Process. Netw.*, vol. 5, no. 3, pp. 469–478, Sep. 2019.

[38] K. Shima, S. Kojima, K. Ito, K. Maruta, and C.-J. Ahn, "Adaptive subcarrier grouping for MMSE-SMI adaptive array interference suppression," *IEEE Access*, vol. 9, pp. 18361–18372, 2021.

[39] K. Tamura, S. Kojima, K. Komatsu, J. Cha, and C.-J. Ahn, "Spectrally efficient frequency division multiplexing with QRM-MLD for Rician fading channel," in *Proc. Int. Conf. Adv. Technol. Commun. (ATC)*, Oct. 2023, pp. 128–133.

[40] S. Kojima, K. Maruta, and C.-J. Ahn, "Adaptive modulation and coding using neural network based SNR estimation," *IEEE Access*, vol. 7, pp. 183545–183553, 2019.

[41] S. Kojima, Y. Goto, K. Maruta, S. Sugiura, and C. J. Ahn, "Timing synchronization based on supervised learning of spectrogram for OFDM systems," *IEEE Trans. Cognit. Commun. Netw.*, vol. 9, no. 5, pp. 1141–1154, Oct. 2023.

[42] J. Zhang, T. Wang, Z. Feng, and S. Yang, "Towards the automatic modulation classification with adaptive wavelet network," *IEEE Trans. Cogn. Commun. Netw.*, vol. 9, no. 3, pp. 549–563, Jun. 2023.

[43] J. Vieira, E. Leitinger, M. Sarajlic, X. Li, and F. Tufvesson, "Deep convolutional neural networks for massive MIMO fingerprint-based positioning," in *Proc. IEEE 28th Annu. Int. Symp. Pers., Indoor, Mobile Radio Commun. (PIMRC)*, Oct. 2017, pp. 1–6.

[44] D. Hong et al., "More diverse means better: Multimodal deep learning meets remote-sensing imagery classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 5, pp. 4340–4354, May 2021.

[45] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: A simple way to prevent neural networks from overfitting," *J. Mach. Learn. Res.*, vol. 15, no. 1, pp. 1929–1958, 2014.

[46] M. Wang, S. Lu, D. Zhu, J. Lin, and Z. Wang, "A high-speed and low-complexity architecture for softmax function in deep learning," in *Proc. IEEE Asia–Pacific Conf. Circuits Syst. (APCCAS)*, 2018, pp. 223–226.

[47] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2014, *arXiv:1412.6980*.

[48] Z. Liu, H. Mao, C.-Y. Wu, C. Feichtenhofer, T. Darrell, and S. Xie, "A ConvNet for the 2020s," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2022, pp. 11976–11986.

[49] J. Chen et al., "Run, don't walk: Chasing higher FLOPS for faster neural networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2023, pp. 12021–12031.

KOSUKE TAMURA (Graduate Student Member, IEEE) received the B.E. degree from Chiba University, Japan, in 2023, where he is currently pursuing the M.E. degree in electrical and electronics engineering. His research interests include OFDM, SEFDM, MIMO, and machine learning-based wireless communication systems. He received the Best Paper Award at the International Conference on Electronics, Information, and Communication (ICEIC) in 2024.

SHUN KOJIMA (Member, IEEE) received the B.E., M.E., and Ph.D. degrees in electrical and electronics engineering from Chiba University, Japan, in 2017, 2018, and 2021, respectively. From 2021 to 2022, he was an Assistant Professor with the Department of Fundamental Engineering, Utsunomiya University, Tochigi, Japan. He is currently a Project Research Associate with the Institute of Industrial Science, The University of Tokyo, Tokyo, Japan. His research interests include adaptive modulation and coding, visible light communications, physical layer security, and machine learning. He received the Best Paper Award at the 26th International Conference on Software, Telecommunications and Computer Networks in 2018, the Best Poster Award at the 3rd Communication Quality Student Workshop in 2019, the IEEE VTS Tokyo/Japan Chapter 2020 Young Researcher's Encouragement Award, the RISP Best Paper Award in 2021, the Institute of Electronics, Information and Communication Engineers (IEICE) Radio Communication Systems Active Researcher Award in 2021, the IEICE Young Researchers Award in 2023, and the Takayanagi Research Encouragement Award in 2023.

PHUC V. TRINH (Member, IEEE) received the B.E. degree in electronics and telecommunications from the Posts and Telecommunications Institute of Technology (PTIT), Hanoi, Vietnam, in 2013, and the M.Sc. and Ph.D. degrees in computer science and engineering from The University of Aizu, Aizuwakamatsu, Japan, in 2015 and 2017, respectively. From 2017 to 2023, he was a Researcher with the Space Communication Systems Laboratory, Wireless Networks Research Center, National Institute of Information and Communications Technology (NICT), Tokyo, Japan. Since 2023, he has been a Project Research Associate (Specially Appointed Assistant Professor) with the Communications and Signal Processing Laboratory (Sugiura Laboratory), Institute of Industrial Science, The University of Tokyo, Tokyo. His current research interests include optical and wireless communications for space, airborne, and terrestrial networks.

SHINYA SUGIURA (Senior Member, IEEE) received the B.S. and M.S. degrees in aeronautics and astronautics from Kyoto University, Kyoto, Japan, in 2002 and 2004, respectively, and the Ph.D. degree in electronics and electrical engineering from the University of Southampton, Southampton, U.K., in 2010.

From 2004 to 2012, he was a Research Scientist with Toyota Central R&D Labs., Inc., Nagakute, Japan. From 2013 to 2018, he was an Associate Professor with the Department of Computer and Information Sciences, Tokyo University of Agriculture and Technology, Koganei, Japan. Since 2018, he has been an Associate Professor with the Institute of Industrial Science, The University of Tokyo, Tokyo, Japan, where he heads the Wireless Communications Research Group. He has authored or coauthored over 110 IEEE journal and magazine articles. His research interests include wireless communications, networking, signal processing, and antenna technology.

Dr. Sugiura was a recipient of numerous awards, including the 18th JSPS Prize in 2022, the Fifth Yasuharu Suematsu Award in 2019, the Sixth RIEC Award from the Foundation for the Promotion of Electrical Communication in 2016, the Young Scientists' Prize by the Minister of Education, Culture, Sports, Science and Technology of Japan in 2016, the 14th Funai Information Technology Award (First Prize) from the Funai Foundation in 2015, the 28th Telecom System Technology Award from the Telecommunications Advancement Foundation in 2013, the Sixth IEEE Communications Society Asia–Pacific Outstanding Young Researcher Award in 2011, the 13th Ericsson Young Scientist Award in 2011, and the 2008 IEEE Antennas and Propagation Society Japan Chapter Young Engineer Award. He has been serving as an Editor for IEEE Wireless Communications Letters since 2019 and an Associate Editor for IEEE Transactions on Communications since 2023. He served as an Editor for *Scientific Reports* (2021–2024). He was certified as the 2021 IEEE Wireless Communications Letters Exemplary Editor.

CHANG-JUN AHN (Senior Member, IEEE) received the Ph.D. degree from the Department of Information and Computer Science, Keio University, Japan, in 2003. From 2001 to 2003, he was a Research Associate with the Department of Information and Computer Science, Keio University. From 2003 to 2006, he was with the Communication Research Laboratory, Independent Administrative Institution (now the National Institute of Information and Communications Technology). In 2006, he was on assignment at ATR Wave Engineering Laboratories. In 2007, he was with the Faculty of Information Sciences, Hiroshima City University, as a Lecturer. From 2019 to 2021, he was with the Electrical and Computer Engineering, Duke University, as a Visiting Professor. He is currently with the Graduate School of Engineering, Chiba University, as a Professor. His research interests include OFDM, MIMO, digital communication, channel coding, and signal processing for telecommunications. From 2005 to 2006, he was an Expert Committee Member of the Emergence Communication Committee, Shikoku Bureau of Telecommunications, Ministry of Internal Affairs and Communications (MIC), Japan. From 2010 to 2016, he was a Technical Committee Member of IEICE. He is a Senior Member of IEICE. He received the ICF Research Award for Young Engineer in 2002, the Funai Information Science Award for Young Scientist in 2003, the Distinguished Service Award from Hiroshima City in 2010, the IEEE SoftCOM2018 Best Paper Award, the IEEE APCC2019 Best Paper Award, the IEICE ICETC2020 Best Paper Award, *Journal of Signal Processing* Best Paper Award in 2021, the IEEE ICCE-Asia2022 Best Paper Award, the IEEE ISAAC 2023 Best Paper Award, and the IEEE ICEIC2024 Best Paper Award. From 2021 to 2023, he was an Editor of IEICE Transactions.