IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 4 July 2023, accepted 24 July 2023, date of publication 2 August 2023, date of current version 9 August 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3301139

RESEARCH ARTICLE

Short-Packet URLLCs for MIMO Underlay Cognitive Multihop Relaying With Imperfect CSI

NGO HOANG TU[1], (Graduate Student Member, IEEE), TRONG-DAI HOANG[2], (Graduate Student Member, IEEE), AND KYUNGCHUN LEE[1,3], (Senior Member, IEEE)

[1]Department of Electrical and Information Engineering, Seoul National University of Science and Technology, Seoul 01811, Republic of Korea
[2]Global Big Data Technologies Centre, University of Technology Sydney, Ultimo, NSW 2007, Australia
[3]Research Center for Electrical and Information Technology, Seoul National University of Science and Technology, Seoul 01811, Republic of Korea

Corresponding author: Kyungchun Lee (kclee@seoultech.ac.kr)

This work was supported in part by the Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education under Grant NRF-2019R1A6A1A03032119, and in part by the NRF Grant funded by the Korean Government [Ministry of Science and Information and Communication Technology (ICT) (MSIT)] under Grant NRF-2022R1A2C1006566.

ABSTRACT In this work, we investigate short-packet communications for multiple-input multiple-output underlay cognitive multihop relay networks with multiple primary users, where transceivers transmit and receive short packets to provide ultra-reliable and low-latency communications (uRLLCs). For performance evaluation, the closed-form expressions of the end-to-end (E2E) block error rate (BLER) for the considered systems are derived in a practical scenario under imperfect channel state information of the interference channels, from which the E2E throughput, energy efficiency (EE), latency, reliability, and asymptotic analysis are also studied. Based on the analytical results, we adapt some state-of-the-art machine learning (ML)-aided estimators to predict the system performance in terms of the E2E throughput, EE, latency, and reliability for real-time configurations. We also obtain the closed-form expressions for the optimal power-allocation and relay-location strategies to minimize the asymptotic E2E BLER under the proportional tolerable interference power and uRLLC constraints, which require negligible computational complexity and offer significant power savings. Furthermore, the ML-based evaluation achieves equivalent performance while significantly reducing the execution time compared to conventional analytical and simulation methods. Among the ML frameworks, the extreme gradient boosting model is demonstrated to be the most efficient estimator for future practical real-time applications.

INDEX TERMS Short-packet communication (SPC), multiple-input multiple-output (MIMO), underlay cognitive radio, multihop relaying, machine learning.

I. INTRODUCTION

Because of the spectral scarcity in fifth-generation (5G) and beyond networks [2], [3], higher spectral efficiency (SE) is a key challenge attracting research attention. To improve the SE, the underlay cognitive radio (CR) enables secondary users (SUs) to simultaneously leverage the licensed frequency band of the primary network without causing any harmful interference to the primary users (PUs) [4]. In the CR paradigm, the transmit power of the secondary transmitters must be limited, which leads to performance degradation of the secondary network.

Herein, a multiple-input multiple-output (MIMO) multihop relay network along with CR has been proposed to combat the spectrum scarcity and limited transmit power problems, extend the network coverage, and improve reliability. In particular, Gao et al. [5] investigated a joint optimzation scheme of the channel assignment in a CR

The associate editor coordinating the review of this manuscript and approving it for publication was Adao Silva.

network and MIMO degree-of-freedom (DoF) allocation to maximize user throughputs in a multihop network. In [6], an optimal relay precoding scheme was developed for spectrum-sharing multihop MIMO-empowered CR networks, aiming to maximize the capacity under the constraints of the transmit power at SUs and allowable interference power at the PUs. Zaidi et al. [7] characterized the system performance in terms of area SE and energy efficiency (EE) of multihop MIMO underlay CR networks, where suboptimal performance was achieved by arbitrarily selecting one DoF and adapting the other to meet the PUs' quality-of-service requirements. Subsequently, in [8], Zaidi et al. further quantified the achievable spatial throughput of the same system model in [7], where SUs operate under slotted-ALOHA medium-access control and channel modulation. Furthermore, the average forward progress, isolation probability, maximum permissible medium-access probability, and optimal spatial throughput were also characterized. Meanwhile, [9] compared various diversity schemes for the performance of multihop MIMO underlay CR networks, including transmit antenna selection (TAS) with maximum ratio combining (MRC) and TAS with/without receive interference cancellation. In addition, the power-allocation (PA) and relay-location (RL) optimizations were also analyzed in [9]. Xu et al. [10] proposed a jointly optimal framework for multihop MIMO CR networks that optimizes the bandwidth allocation, rate control, multihop/multipath routing, and EE. In [11], the outage probability (OP), with the joint constraints of limited interference and intercept probability, was significantly decreased by exploiting MIMO CR multihop relay networks. Furthermore, the authors in [12] optimized the throughput of multihop MIMO CR networks by using energy harvesting (EH) with adaptive transmit power.

However, the techniques in the aforementioned studies may not satisfy the rigorous conditions of ultra-reliable and low-latency communications (uRLLCs) in 5G and beyond networks. Because the uRLLCs' stringent requirements for 5G and beyond require not only a very low latency of around $1-10$ ms [13], but also ultra-reliable with a block error rate (BLER) of less than 10^{-5} [14], short-packet communication (SPC), also known as finite-blocklength (FBL) communication, has been considered as an efficient enabling technology to support the uRLLCs. Specifically, in [15], the average packet latency, EE, and reliability of uplink EH underlay CR internet-of-things (IoT) networks were improved by utilizing SPCs in accordance with the uRLLC requirements. Reference [16] investigated SPCs for dual-hop relaying networks, where the system performance was evaluated via the end-to-end (E2E) BLER, latency, and throughput. The optimal PA and RL configurations were also provided in [16]. The authors in [17] investigated the BLER performance of the amplify-and-forward dual-hop relaying networks, wherein uRLLCs are conducted through SPCs under imperfect channel-state information (CSI) and hardware impairments. In addition, an optimization problem for minimizing the average BLER in the high signal-to-noise ratio (SNR) and high blocklength regimes was studied in [17]. Makki and Alouini [18] evaluated the E2E latency and throughput of multihop SPCs to support uRLLCs, where either the amplify-and-forward or decode-and-forward (DF) technique was utilized. Meanwhile, [19] performed the decoding error rate evaluation, optimal blocklength, and optimal RL allocation for the unmanned aerial vehicle-based DF three-hop relay networks. In [20], the secrecy throughput of dual-hop relaying CR SPC IoT networks was analyzed under the constraints of secrecy and the decoding error probability, where the DF relay utilizes either the MRC or zero-forcing beamforming scheme. Reference [21] investigated TAS/MRC and TAS/selection-combining (SC) MIMO systems with SPCs to gain the advantages of connectivity, SE, and uRLLCs. Subsequently, our work [22], [23] recently focused on leveraging the benefits of multihop MIMO relay networks in SPCs to facilitate uRLLCs. In addition, the PA and RL optimizations were also taken into account for the joint optimal problem in [23].

It is conceivable that taking the closed-form expressions of the performance analysis will no longer be sufficient when the complexity and heterogeneity of future wireless networks grow enormously. Numerical integration and simulation approaches can be utilized as alternative methods to evaluate the system performance, but suffer from long run times to exactly achieve the converged value. Recently, machine learning (ML) has been realized as a powerful tool to accurately evaluate system performance while dramatically reducing the execution time compared to conventional simulation methods [24]. Because ML-based applications are able to accurately estimate non-linear functions with low complexity, they enable a wide variety of networks (e.g., SPC, MIMO, CR, and multihop relaying) for real-time applications. For instance, [25] utilized a deep neural network (DNN) to efficiently predict the E2E BLER, goodput, and EE of wireless-powered CR SPC dualhop IoT systems, where an opportunistic relay selection scheme was proposed to maximize the E2E SNR. Based on the results in [25], performance prediction with high accuracy and low DNN execution time, which can facilitate real-time configurations for IoT systems, was confirmed. Nguyen et al. [26] adopted a convolution neural network (CNN) to estimate the throughput of EH full-duplex dual-hop relay-selection IoT systems using SPCs. Meanwhile, DNN was applied to predict the throughput of multihop wireless-powered networks with SPCs [27]. In addition, the system performance was evaluated in terms of E2E BLER, reliability, latency, and throughput in [27] to show the good support of the proposed networks in accordance with uRLLC constraints. Subsequently, Vu et al. [28] leveraged a DNN framework for the ergodic capacity prediction towards real-time configurations of EH CR non-orthogonal multiple-access (NOMA) dual-hop relay IoT networks. In [29], an extension of the DNN model in [28], called a deep multi-output neural network, was designed to

simultaneously predict the NOMA users' throughput and E2E BLER of wireless-powered CR NOMA IoT SPC systems. In addition to the DNN and CNN advantages, the recurrent neural network (RNN) was further utilized for performance prediction in terms of the OP and throughput, where the network model of the dual-hop coordinated direct relay transmissions and underlay CR NOMA was considered [30]. Although RNN provides the lowest root-mean-square error (RMSE), it suffers from the highest computational complexity and execution time compared to DNN and CNN.

This paper is the first to study MIMO CR multihop relay networks with multiple PUs for SPCs to satisfy uRLLC requirements, where the comprehensive analysis is conducted in a general and practical scenario with imperfect CSI. Subsequently, some state-of-the-art ML-aided estimators are designed to effectively predict the system performance. Furthermore, the practical optimization problems are also studied under the defined system constraints. The main contributions of this paper are summarized as follows:

- We propose and investigate a MIMO underlay CR multihop relay network with multiple PUs using SPCs in accordance with uRLLC requirements, where the closed-form expressions of the E2E BLER, throughput, EE, latency, and reliability are obtained to evaluate the system performance.
- To achieve more qualitative insights into performance behaviors, we analyze the asymptotic BLER in the high-SNR regime with two practical approaches, including the fixed and proportional tolerable interference power per PU. Here, the fixed interference-power scheme can be split into near- and far-CR regimes.
- To tackle two important practical problems (i.e., the optimal configurations of PA and RL) of the relaying systems, we obtain the closed-form expressions of the optimal PA and RL configurations that minimize the asymptotic E2E BLER under the system constraints. In addition, substantial power savings are observed in the numerical results.
- To develop real-time system configurations for the proposed networks, we design various state-of-the-art ML-based models, i.e., DNN, CNN, and extreme gradient boosting (XGB) algorithms, to estimate the system performance in terms of E2E throughput, EE, latency, and reliability. The numerical results confirm that our proposed ML-aided estimators exhibit highly accurate predictions with short execution times. Among the ML frameworks, the XGB model achieves the lowest RMSE while requiring the lowest computational complexity, making it the most efficient estimator for future practical real-time applications.

The remainder of this paper is organized as follows. In Section II, the networks and channel model with the conventional TAS/SC and TAS/MRC are described. Section III presents the closed-form expressions for the E2E BLER, throughput, EE, latency, and reliability in both the TAS/MRC and TAS/SC schemes. Section IV provides further insights into the network performance via the practical asymptotic analysis. In Section V, we minimize the asymptotic E2E BLER for the proportional tolerable interference power and uRLLC requirements under the PA and RL constraints. Section VI investigates the DNN, CNN, and XGB applications to predict the E2E throughput, EE, latency, and reliability performance. To validate the correctness of our analysis, Section VII shows the Monte Carlo simulations and ML-based evaluation results. Finally, some concluding remarks are presented in Section VIII.

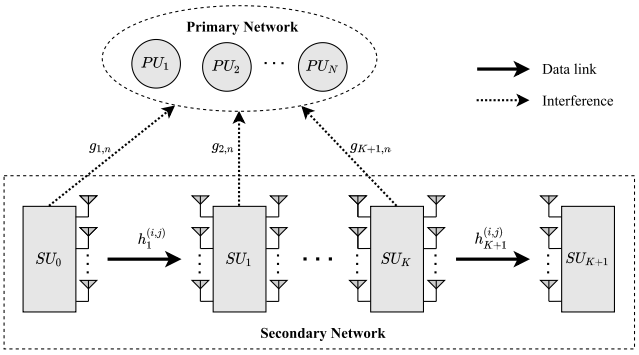


FIGURE 1. The MIMO underlay cognitive multihop relay network under SPCs.

II. NETWORKS AND CHANNEL MODEL

A MIMO underlay CR multihop network for SPC is shown in Fig. 1, where Rayleigh fading channels are assumed. In the primary network, there coexist N PUs that share the same licensed frequency band, whereas in the secondary network, the signal is sent from the secondary source SU_0 equipped with N_T transmit antennas to the secondary destination SU_{K+1} equipped with N_R receive antennas via K secondary relays, named $SU_1, SU_2, \ldots, SU_K$, that are equipped with N_T transmit and N_R receive antennas. We assume that the direct link between SU_0 and SU_{K+1} is not supported and all secondary devices' antennas are equipped with the half-duplex mode, i.e., a single secondary device cannot transmit and receive the signal simultaneously. Let $h_k^{(i,j)}$ denote the fading coefficient for the channel between transmit antenna i and receive antenna j at hop k, whereas $g_{k,n}^{(i)}$ denotes the channel coefficient between transmit antenna i of SU_{k-1} and PU_n. Throughout this paper, $i = \overline{1, N_T}$, $j = \overline{1, N_R}$, $k = \overline{1, K+1}$, and $n = \overline{1, N}$ are assumed. We note that only a single transmit antenna of node SU_{k-1} is activated for transmission by the TAS scheme, and $g_{k,n}$ represents the channel between the activated transmit antenna of SU_{k-1} and PU_n.

In the underlay CR systems, the transmit power at each SU must be lower than the maximum tolerable interference power to avoid severe interference at the PUs. Therefore, the transmit power of SU_{k-1} can be expressed as $P_{k-1} = \min\left(P_s, \frac{\mathcal{I}_p}{\max\limits_{n=1,\ldots,N} |g_{k,n}|^2}\right)$, where P_s is the maximum allowable transmit power per transmit station in the secondary network

and $\mathcal{I}_p$ is the maximum allowable interference power for a PU. However, perfect CSI of the interference channels is not available in practice because of complex channel intricacies, e.g., feedback delay, limited feedback, and channel estimation (CE) errors. Henceforth, the estimate of the channel coefficient $g_{k,n}$ can be expressed as $\hat{g}_{k,n} = \beta g_{k,n} + \sqrt{1-\beta^2} g_{k,n}^{\text{err}}$, where $g_{k,n}^{\text{err}}$ is a circular symmetric complex Gaussian random variable with zero-mean and the same variance of $g_{k,n}$, whereas $\beta \in [0, 1]$ represents the correlation coefficient between the real and estimated channels, which is utilized to measure the CSI condition.[1] As a result, the transmit power of SU_{k-1} with imperfect CSI is expressed as $\hat{P}_{k-1} = \min\left(P_s, \frac{\mathcal{I}_p}{\max\limits_{n=1,\ldots,N} |\hat{g}_{k,n}|^2}\right)$. We note that a higher value of β implies a lower CE error. As done in previous studies [33], [34], [35], we assume that the distance between the primary transmitter and the secondary network is sufficiently large so that any interference to the secondary network caused by the primary network is negligible. Accordingly, the instantaneous received SNR at SU_k with imperfect CSI for the channel between the ith transmit antenna and the jth receive antenna is determined by

$$\gamma_k^{(i,j)} = \frac{\hat{P}_{k-1}}{\mathcal{N}_0} \left|h_k^{(i,j)}\right|^2 = \min\left(\bar{\gamma}_P, \frac{\bar{\gamma}_I}{\max\limits_{n=1,\ldots,N} |\hat{g}_{k,n}|^2}\right) \left|h_k^{(i,j)}\right|^2, \quad (1)$$

where $\bar{\gamma}_P = P_s/\mathcal{N}_0$, $\bar{\gamma}_I = \mathcal{I}_p/\mathcal{N}_0$, and $\mathcal{N}_0$ is the noise variance. For the Rayleigh fading channel, $\left|h_k^{(i,j)}\right|^2$ and $|g_{k,n}|^2$ follow an exponential distribution with characteristic parameters $\lambda_k^{(i,j)}$ and $\lambda_{k,n}$. We assume the channel coefficients of each hop and interference link are independent and identically distributed, i.e., $\lambda_k^{(i,j)} = \lambda_k$, $\lambda_{k,n} = \lambda_{kp}$, $\forall i, j, n$. For the transmission schemes, the TAS technique is employed at the transmit side to achieve transmit diversity, power consumption reduction, and hardware cost reduction [36]. For TAS, only a single optimal antenna out of N_T providing the highest received SNR is selected to transmit the signal. Meanwhile, either MRC or SC is utilized at the receive side to achieve receive diversity [23], [37].

According to the MRC principle, the received signals from all branch channels are coherently combined [38]. In the kth hop, when transmit antenna i is chosen based on the TAS technique, the instantaneous output SNR after combining all the received signals can be expressed as

$$\gamma_k^{\text{TAS/MRC}} = \max_{i=1,\ldots,N_T} \sum_{j=1}^{N_R} \gamma_k^{(i,j)}. \quad (2)$$

Proposition 1: By denoting $\mathcal{X}_k = \sum\limits_{j=1}^{N_R} \left|h_k^{(i,j)}\right|^2$ *and* $\mathcal{Y}_k = \max\limits_{n=1,\ldots,N} |\hat{g}_{k,n}|^2$, *the cumulative distribution functions (CDFs) and probability density functions (PDFs) of* $\mathcal{X}_k$ *and* $\mathcal{Y}_k$ *are given, respectively, by*

$$F_{\mathcal{X}_k}(x) = 1 - \exp(-\lambda_k x) \sum_{n=0}^{N_R-1} \frac{1}{n!} (\lambda_k x)^n, \quad (3)$$

$$f_{\mathcal{X}_k}(x) = \frac{\lambda_k^{N_R} x^{N_R-1}}{(N_R-1)!} \exp(-\lambda_k x), \quad (4)$$

$$F_{\mathcal{Y}_k}(x) = 1 - \sum_{i=0}^{N-1} \frac{\phi(N,i)}{2} \exp\left(-\hat{\lambda}_{kp}^{(i)} x\right), \quad (5)$$

$$f_{\mathcal{Y}_k}(x) = \sum_{i=0}^{N-1} \frac{\phi(N,i)}{2} \hat{\lambda}_{kp}^{(i)} \exp\left(-\hat{\lambda}_{kp}^{(i)} x\right), \quad (6)$$

where $\phi(N,i) = \binom{N-1}{i} \frac{2N(-1)^i}{i+1}$ *and* $\hat{\lambda}_{kp}^{(i)} = \frac{(i+1)\lambda_{kp}}{i+1-i\beta^2}$. *By substituting (1) into (2) and utilizing the CDFs and PDFs of* $\mathcal{X}_k$ *and* $\mathcal{Y}_k$, *the CDF of* $\gamma_k^{\text{TAS/MRC}}$ *is obtained as*

$$F_{\gamma_k}^{\text{TAS/MRC}}(\gamma) = \left\{ F_{\mathcal{Y}_k}\left(\frac{\bar{\gamma}_I}{\bar{\gamma}_P}\right) F_{\mathcal{X}_k}\left(\frac{\gamma}{\bar{\gamma}_P}\right) + \sum_{i=0}^{N-1} \frac{\phi(N,i)}{2} \cdot \exp\left(-\hat{\lambda}_{kp}^{(i)} \frac{\bar{\gamma}_I}{\bar{\gamma}_P}\right) - \sum_{i=0}^{N-1} \frac{\phi(N,i)}{2} \hat{\lambda}_{kp}^{(i)} \sum_{n=0}^{N_R-1} \frac{1}{n!} \left(\frac{\lambda_k \gamma}{\bar{\gamma}_I}\right)^n \cdot \left(\hat{\lambda}_{kp}^{(i)} + \frac{\lambda_k \gamma}{\bar{\gamma}_I}\right)^{-n-1} \Gamma\left(n+1, \frac{\hat{\lambda}_{kp}^{(i)} \bar{\gamma}_I + \lambda_k \gamma}{\bar{\gamma}_P}\right) \right\}^{N_T}, \quad (7)$$

where $\Gamma(\alpha, x) = \int_x^\infty e^{-t} t^{\alpha-1} dt$ *denotes the upper incomplete gamma function [39, Eq. (8.350.2)].*

Proof: Based on (2), the CDF of $\gamma_k^{\text{TAS/MRC}}$ is calculated as

$$F_{\gamma_k}^{\text{TAS/MRC}}(\gamma) = \prod_{i=1}^{N_T} \Pr\left\{ \min\left(\bar{\gamma}_P, \frac{\bar{\gamma}_I}{\max\limits_{n=1,\ldots,N} |\hat{g}_{k,n}|^2}\right) \sum_{j=1}^{N_R} \left|h_k^{(i,j)}\right|^2 \le \gamma \right\} = \left\{ \Pr\left(\bar{\gamma}_P \le \frac{\bar{\gamma}_I}{\mathcal{Y}_k}, \bar{\gamma}_P \mathcal{X}_k \le \gamma\right) + \Pr\left(\bar{\gamma}_P > \frac{\bar{\gamma}_I}{\mathcal{Y}_k}, \frac{\bar{\gamma}_I}{\mathcal{Y}_k} \mathcal{X}_k \le \gamma\right) \right\}^{N_T}. \quad (8)$$

[1] β generally characterizes the influence of CE errors or hardware impairments on the networks. In this work, we do not specify the particular CE methods (e.g., minimum mean-square-error (MMSE), linear-MMSE, local-partial-MMSE, maximum likelihood estimator, least-square estimator, etc.). The details of these CE techniques are provided in [31, Chapter 11]. The value of β can be determined by leveraging a pilot symbol modulated parameter [32].

It is noted that, in (8), the first term is the probability of two independent events, whereas the second term is the probability of two dependent events. Therefore, (8) can be rewritten as

$$F_{\gamma_k}^{\text{TAS/MRC}}(\gamma) = \left\{ F_{\mathcal{Y}_k}\left(\frac{\bar{\gamma}_I}{\bar{\gamma}_P}\right) F_{\mathcal{X}_k}\left(\frac{\gamma}{\bar{\gamma}_P}\right) + \int_{\bar{\gamma}_I/\bar{\gamma}_P}^{\infty} f_{\mathcal{Y}_k}(y)\, F_{\mathcal{X}_k}\left(\frac{\gamma}{\bar{\gamma}_I} y\right) dy \right\}^{N_T}. \quad (9)$$

By substituting the CDFs and PDFs of $\mathcal{X}_k$ and $\mathcal{Y}_k$ into the integral in (9), after some mathematical manipulations, we derive (7), which completes the proof. □

According to the SC principle, only the link with the highest average received SNR is chosen to perform the signal detection [40]. When utilizing both the TAS and SC schemes, the instantaneous output SNR can be expressed as

$$\gamma_k^{\text{TAS/SC}} = \max_{\substack{i=1,\ldots,N_T \\ j=1,\ldots,N_R}} \gamma_k^{(i,j)}. \quad (10)$$

Proposition 2: Based on (10), the CDF of $\gamma_k^{\text{TAS/SC}}$ is obtained as

$$F_{\gamma_k}^{\text{TAS/SC}}(\gamma) = \left\{ F_{\mathcal{Y}_k}\left(\frac{\bar{\gamma}_I}{\bar{\gamma}_P}\right) F_{\left|h_k^{(i,j)}\right|^2}\left(\frac{\gamma}{\bar{\gamma}_P}\right) + \sum_{i=0}^{N-1} \frac{\phi(N,i)}{2} \exp\left(-\hat{\lambda}_{kp}^{(i)} \frac{\bar{\gamma}_I}{\bar{\gamma}_P}\right) - \sum_{i=0}^{N-1} \frac{\phi(N,i)}{2} \cdot \frac{\hat{\lambda}_{kp}^{(i)} \bar{\gamma}_I}{\hat{\lambda}_{kp}^{(i)} \bar{\gamma}_I + \lambda_k \gamma} \exp\left(-\frac{\hat{\lambda}_{kp}^{(i)} \bar{\gamma}_I + \lambda_k \gamma}{\bar{\gamma}_P}\right) \right\}^{N_T N_R}, \quad (11)$$

where $F_{\left|h_k^{(i,j)}\right|^2}(x) = 1 - \exp(-\lambda_k x)$.

Proof: The proof of (11) follows the same steps as **Proposition 1**. □

III. PERFORMANCE ANALYSIS

In this section, we analyze the E2E performance metrics, including the E2E BLER, E2E throughput, EE, latency, and reliability.

A. E2E BLER

We assume that SU_0 sends $\mathcal{B}$ information bits to SU_{K+1} with the assistance of $SU_1, \ldots, SU_K$ via quasi-static fading channels [41], where $\mathcal{B}$ information bits are encoded into a block of τ channel uses (CUs). Thus, the channel coding rate is given by $\mathcal{R} = \mathcal{B}/\tau$. Under the FBL regime with an SPC of $\tau > 100$ CUs [42], [43], the average BLER at link k is given by

$$\bar{\varepsilon}_k^{\mathcal{S}} = \mathbb{E}\left\{ Q\left(\frac{\log_2(1+\gamma_k^{\mathcal{S}}) - \mathcal{B}/\tau}{\sqrt{\mathcal{V}(\gamma_k^{\mathcal{S}})/\tau}} \right) \right\}, \quad (12)$$

where $\mathcal{S}$ denotes one type of the transmission scheme, i.e., $\mathcal{S} \in \{\text{TAS/MRC}, \text{TAS/SC}\}$; $\mathbb{E}(\cdot)$ represents the expectation operator, $\mathcal{V}(\gamma_k^{\mathcal{S}}) \triangleq \left(1 - \frac{1}{(1+\gamma_k^{\mathcal{S}})^2}\right)(\log_2 e)^2$ denotes the channel dispersion [42]; and $Q(\cdot)$ is the Gaussian Q-function. It is challenging to directly derive the closed-form expression of $\bar{\varepsilon}_k^{\mathcal{S}}$ in (12) because of the complicated Q-function. To tackle this problem, we utilize a tight approximation of the Q-function, as discussed in [43], which yields

$$\bar{\varepsilon}_k^{\mathcal{S}} = \vartheta\sqrt{\tau} \int_{\psi_L}^{\psi_H} F_{\gamma_k}^{\mathcal{S}}(\gamma)\, d\gamma, \quad (13)$$

where $\vartheta = \left[2\pi\left(2^{2\mathcal{B}/\tau} - 1\right)\right]^{-1/2}$, $\psi_L = \theta - 1/\left(2\vartheta\sqrt{\tau}\right)$, $\psi_H = \theta + 1/\left(2\vartheta\sqrt{\tau}\right)$, and $\theta = 2^{\mathcal{B}/\tau} - 1$.

It is noted that the integral in (13) is very strenuous to calculate due to the complexity of the output SNR's CDF. Motivated by this issue, we utilize two approximations to obtain a tightly bounded closed-form expression for the average BLER at each hop. These approximated frameworks can achieve not only very high accuracy, but also low complexity, as presented in the following propositions.

Proposition 3 (First-Order Riemann Integral Approximation): The tightly bounded closed-form expression for the average BLER in the kth hop can be derived by using the first-order Riemann integral approximation [44]: $\bar{\varepsilon}_k^{\mathcal{S},\text{Rie}} = F_{\gamma_k}^{\mathcal{S}}(\theta)$.

Proof: When $\tau > 100$ CUs, it is observed that $\psi_H - \psi_L = \sqrt{2\pi\left(2^{2\mathcal{B}/\tau} - 1\right)/\tau}$ is very small. Therefore, it is valid to utilize the first-order Riemann integral approximation $\int_x^y f(z)\, dz \cong (y-x) f\left(\frac{x+y}{2}\right)$ for (13). As a result, (13) is approximated as $\bar{\varepsilon}_k^{\mathcal{S},\text{Rie}} = \vartheta\sqrt{\tau}\,(\psi_H - \psi_L) F_{\gamma_k}^{\mathcal{S}}\left(\frac{\psi_H+\psi_L}{2}\right) \overset{(a)}{=} F_{\gamma_k}^{\mathcal{S}}(\theta)$, where step (a) is based on the observation that $\psi_H - \psi_L = 1/\left(\vartheta\sqrt{\tau}\right)$ and $\psi_H + \psi_L = 2\theta$, which completes the proof. □

Proposition 4 (Gauss–Chebyshev quadrature): By using the Gauss–Chebyshev quadrature integral $\int_x^y f(z)\, dz \cong \frac{(y-x)}{2} \sum_{u=1}^{T} \frac{\pi}{T}\sqrt{1-x_u^2} f\left(\frac{y-x}{2} x_u + \frac{y+x}{2}\right)$, where $x_u = \cos\left(\frac{(2u-1)\pi}{2T}\right)$ and T denotes the number of tradeoff terms, i.e., the tradeoff parameter between the complexity and accuracy [45], the tightly bounded closed-form expression for the average BLER at link k can be obtained as $\bar{\varepsilon}_k^{\mathcal{S},\text{GCheb}} = \sum_{u=1}^{T} \frac{\pi}{2T}\sqrt{1-x_u^2} F_{\gamma_k}^{\mathcal{S}}\left(\frac{x_u}{2\vartheta\sqrt{\tau}} + \theta\right)$.

*Remark 1: Based on **Proposition 3** and **Proposition 4**, the approximated BLERs are revealed to be directly dependent on the original CDF forms. We can conclusively comment that the first-order Riemann integral approximation offers lower complexity than the Gauss-Chebyshev quadrature because the latter suffers from the tradeoff constant T. In this work, we not only utilize the first-order Riemann integral*

approximation to reduce calculation complexity but also validate the correctness of the Gauss-Chebyshev quadrature approach.

According to the selective DF principle [46], the E2E BLER for the considered system is determined by

$$\bar{\varepsilon}_{\mathrm{E2E}}^{\mathcal{S},\mathcal{Z}} = \bar{\varepsilon}_{1}^{\mathcal{S},\mathcal{Z}} + \sum_{k=1}^{K} \left\{ \bar{\varepsilon}_{k+1}^{\mathcal{S},\mathcal{Z}} \cdot \prod_{m=1}^{k} \left(1 - \bar{\varepsilon}_{m}^{\mathcal{S},\mathcal{Z}}\right) \right\}, \tag{14}$$

where $\mathcal{Z} \in \{\mathrm{Rie}, \mathrm{GCheb}\}$ indicates whether the Riemann approach in **Proposition 3** or the Gauss–Chebyshev approach in **Proposition 4** is applied.

B. E2E THROUGHPUT, EE, LATENCY, AND RELIABILITY

We consider the latency-limited transmission mode, where SU_0 transmits its data with a fixed transmission rate of $\mathcal{R} = \mathcal{B}/\tau$. The system throughput is defined as the achievable effective rate of the considered network, which is measured in bits per CU (BPCU). As a result, the E2E system throughput for scheme $\mathcal{S}$ is given by [47] and [48]

$$\mathrm{TP}_{\mathrm{E2E}}^{\mathcal{S}} = \left(1 - \bar{\varepsilon}_{\mathrm{E2E}}^{\mathcal{S}}\right) \mathcal{R}/(K+1). \tag{15}$$

Furthermore, to obtain more insights into the tradeoff between the system throughput and energy consumption, we determine the EE for scheme $\mathcal{S}$, which is measured in BPCU per watt (BPCU/W) as [25]

$$\mathrm{EE}_{\mathrm{E2E}}^{\mathcal{S}} = \frac{\left(1 - \bar{\varepsilon}_{\mathrm{E2E}}^{\mathcal{S}}\right) \mathcal{R}}{(K+1) \sum_{k=1}^{K+1} \hat{P}_{k-1}}. \tag{16}$$

The efficient gains from SPCs over long-packet conventional communications are low latency and ultra-reliability, where the latency and reliability for scheme $\mathcal{S}$ are presented, respectively, as [25], [47], and [48]

$$\mathrm{Latency}_{\mathrm{E2E}}^{\mathcal{S}} = \frac{\mathcal{B}}{\mathrm{TP}_{\mathrm{E2E}}^{\mathcal{S}}} = \frac{\tau(K+1)}{1 - \bar{\varepsilon}_{\mathrm{E2E}}^{\mathcal{S}}}, \tag{17}$$

$$\mathrm{Reliability}_{\mathrm{E2E}}^{\mathcal{S}} = \left(1 - \bar{\varepsilon}_{\mathrm{E2E}}^{\mathcal{S}}\right) \cdot 100\%. \tag{18}$$

Remark 2: In the high-SNR regime, the bounded system throughput can be directly determined as $\widetilde{\mathrm{TP}}_{\mathrm{E2E}}^{\mathcal{S}} \approx \left(1 - \tilde{\varepsilon}_{\mathrm{E2E}}^{\mathcal{S},\mathrm{Near}}\right) \mathcal{R}/(K+1) \approx \left(1 - \tilde{\varepsilon}_{\mathrm{E2E}}^{\mathcal{S},\mathrm{Prop}}\right) \mathcal{R}/(K+1)$, *where* $\tilde{\varepsilon}_{\mathrm{E2E}}^{\mathcal{S},\mathrm{Near}}$ *and* $\tilde{\varepsilon}_{\mathrm{E2E}}^{\mathcal{S},\mathrm{Prop}}$ *are defined in (21) and (27), respectively. When* $\tilde{\varepsilon}_{\mathrm{E2E}}^{\mathcal{S},\mathrm{Near}}$ *and* $\tilde{\varepsilon}_{\mathrm{E2E}}^{\mathcal{S},\mathrm{Prop}}$ *are extremely small in a sufficiently high-SNR regime, we can simplify the asymptotic system throughput as* $\widetilde{\mathrm{TP}}_{\mathrm{E2E}}^{\mathcal{S}} \approx \mathcal{R}/(K+1)$. *In such a scenario,* $\mathcal{R}/(K+1)$ *is determined as the target throughput of the system.*

Remark 3 (Impact of a large number of hops): We note that the E2E BLER dramatically increases and tends to approach unity as $K \to \infty$. *The related explanation is given in the discussion of Fig. 5(a). Furthermore, based on (15)–(18), we can comment that* $\mathrm{TP}_{\mathrm{E2E}}^{\mathcal{S}} \to 0$, $\mathrm{EE}_{\mathrm{E2E}}^{\mathcal{S}} \to 0$, $\mathrm{Reliability}_{\mathrm{E2E}}^{\mathcal{S}} \to 0$, *and* $\mathrm{Latency}_{\mathrm{E2E}}^{\mathcal{S}} \to \infty$ *as* $K \to \infty$. *Consequently, we can conclude that an excessively large number of hops causes performance degradation. Alternatively, as verified by the simulation results in Section VII, employing an appropriate number of relays not only achieves optimal performance in terms of E2E BLER, throughput, EE, latency, and reliability but also reduces the implementation cost.*

Remark 4 (Impact of a large number of antennas): It is observed in (8) that $F_{\gamma_k}^{\mathrm{TAS/MRC}}$ *is a product of* N_T *probabilities. In addition, increasing* N_R *implies that the DoF of the chi-square random variable* $\mathcal{X}_k$ *also increases, leading to a lower value of* $\Pr\{c\mathcal{X}_k \le \gamma\}$ *for any constant* c. *For the TAS/SC scheme, based on (10), the calculation of* $F_{\gamma_k}^{\mathrm{TAS/SC}}$ *requires a product of* $N_T N_R$ *probabilities. Since* $\Pr\{c\mathcal{X}_k \le \gamma\}$ *is in the range* $[0, 1]$ *and* $\Pr\{c\mathcal{X}_k \le \gamma\} \to 0$ *as* $N_R \to \infty$, *we get* $F_{\gamma_k}^{\mathcal{S}} \to 0$ *as* $N_T, N_R \to \infty$. *Consequently, the E2E BLER approaches zero as* $N_T, N_R \to \infty$, *which is expected as a result of high diversity gains. Accordingly, system performance in terms of throughput, EE, latency, and reliability significantly improves for a large number of antennas.*

IV. ASYMPTOTIC ANALYSIS

In the CR environment, two practical approaches for the asymptotic analysis include the fixed and proportional tolerable interference power per PU. This analysis provides the performance behaviors in special cases, which are useful to obtain more insights into the influence of the system parameters.

A. FIXED TOLERABLE INTERFERENCE POWER CONSTRAINT

When the constraint $\mathcal{I}_P$ for the network is fixed [49], we investigate cases in which (*i*) the PUs are located close to the secondary network, which is referred to as the near-CR asymptotic regime; and (*ii*) the PUs are located very far from the secondary network, which is referred to as the far-CR asymptotic regime.

1) NEAR-CR ASYMPTOTIC REGIME

When the PUs are located close to the secondary network, i.e., $|\hat{g}_{k,n}|$ becomes very large, which yields $\frac{\mathcal{I}_P}{\max\limits_{n=1,\ldots,N} |\hat{g}_{k,n}|^2} \ll P_s$. In such a scenario, the instantaneous received SNR at SU_k with the channel between transmit antenna i and receive antenna j becomes $\gamma_k^{(i,j),\mathrm{Near}} = \frac{\bar{\gamma}_I}{\max\limits_{n=1,\ldots,N} |\hat{g}_{k,n}|^2} \left|h_k^{(i,j)}\right|^2$. Based on (2) and (10), the CDFs of $\gamma_k^{\mathrm{TAS/MRC,Near}}$ and $\gamma_k^{\mathrm{TAS/SC,Near}}$ are derived, respectively, as

$$F_{\gamma_k}^{\mathrm{TAS/MRC,Near}}(\gamma) = \left\{ \sum_{i=0}^{N-1} \frac{\phi(N,i)}{2} \cdot \left[1 - \hat{\lambda}_{kp}^{(i)} \sum_{n=0}^{N_R-1} \left(\frac{\lambda_k \gamma}{\bar{\gamma}_I}\right)^n \left(\hat{\lambda}_{kp}^{(i)} + \frac{\lambda_k \gamma}{\bar{\gamma}_I}\right)^{-n-1} \right] \right\}^{N_T}, \tag{19}$$

$$F_{\gamma_k}^{\mathrm{TAS/SC,Near}}(\gamma) = \left[\sum_{i=0}^{N-1} \frac{\phi(N,i)}{2} \frac{\lambda_k \gamma}{\hat{\lambda}_{kp}^{(i)} \bar{\gamma}_I + \lambda_k \gamma}\right]^{N_T N_R}. \quad (20)$$

Because the results from the first-order Riemann and Gauss–Chebyshev quadrature integral approximations are identical, for simplification, we only utilize the first-order Riemann approach to derive the closed-form expression of the asymptotic BLER. By applying (19) and (20) to **Proposition 3**, the closed-form expression for the asymptotic BLER at link k is given by $\tilde{\varepsilon}_k^{\mathcal{S},\mathrm{Near}} \cong F_{\gamma_k}^{\mathcal{S},\mathrm{Near}}(\theta)$. Eventually, similar to (14), the asymptotic E2E BLER for scheme $\mathcal{S}$ in the near-CR asymptotic regime is obtained as

$$\tilde{\varepsilon}_{\mathrm{E2E}}^{\mathcal{S},\mathrm{Near}} = \tilde{\varepsilon}_1^{\mathcal{S},\mathrm{Near}} + \sum_{k=1}^{K} \left\{ \tilde{\varepsilon}_{k+1}^{\mathcal{S},\mathrm{Near}} \cdot \prod_{m=1}^{k} \left(1 - \tilde{\varepsilon}_m^{\mathcal{S},\mathrm{Near}}\right) \right\}. \quad (21)$$

2) FAR-CR ASYMPTOTIC REGIME

The PUs are located very far from the secondary network in the far regime, i.e., $|\hat{g}_{k,n}|$ becomes very small, which yields $P_s \ll \frac{\mathcal{I}_p}{\max\limits_{n=1,\ldots,N} |\hat{g}_{k,n}|^2}$. In this scenario, the instantaneous received SNR for SU_k is given by $\gamma_k^{(i,j),\mathrm{Far}} = \bar{\gamma}_P \left|h_k^{(i,j)}\right|^2$. Accordingly, the asymptotic CDFs of $\gamma_k^{\mathrm{TAS/MRC,Near}}$ and $\gamma_k^{\mathrm{TAS/SC,Near}}$ are obtained, respectively, as

$$F_{\gamma_k}^{\mathrm{TAS/MRC,Far}}(\gamma) = \prod_{i=1}^{N_T} \Pr\left\{ \bar{\gamma}_P \sum_{j=1}^{N_R} \left|h_k^{(i,j)}\right|^2 \le \gamma \right\} = \left[F_{\mathcal{X}_k}(\gamma/\bar{\gamma}_P)\right]^{N_T}, \quad (22)$$

$$F_{\gamma_k}^{\mathrm{TAS/SC,Far}}(\gamma) = \prod_{i=1}^{N_T} \prod_{j=1}^{N_R} \Pr\left\{ \bar{\gamma}_P \left|h_k^{(i,j)}\right|^2 \le \gamma \right\} = \left[F_{\left|h_k^{(i,j)}\right|^2}(\gamma/\bar{\gamma}_P)\right]^{N_T N_R}. \quad (23)$$

By inserting (22) and (23) to **Proposition 3**, the closed-form expression for the asymptotic BLER at the kth hop is obtained as $\tilde{\varepsilon}_k^{\mathcal{S},\mathrm{Far}} \cong F_{\gamma_k}^{\mathcal{S},\mathrm{Far}}(\theta)$. Similarly, the asymptotic E2E BLER for scheme $\mathcal{S}$ in the far-CR asymptotic regime is

$$\tilde{\varepsilon}_{\mathrm{E2E}}^{\mathcal{S},\mathrm{Far}} = \tilde{\varepsilon}_1^{\mathcal{S},\mathrm{Far}} + \sum_{k=1}^{K} \left\{ \tilde{\varepsilon}_{k+1}^{\mathcal{S},\mathrm{Far}} \cdot \prod_{m=1}^{k} \left(1 - \tilde{\varepsilon}_m^{\mathcal{S},\mathrm{Far}}\right) \right\}. \quad (24)$$

B. PROPORTIONAL TOLERABLE INTERFERENCE POWER CONSTRAINT

When PUs can tolerate a high level of interference from SUs, the peak allowable interference power is proportional to the maximum transmit power, i.e., $\mathcal{I}_p = \mu P_s$, where μ is a proportional positive constant [50]. In the high-SNR regime, i.e., $\bar{\gamma}_P \to \infty$, the asymptotic CDFs for both the TAS/MRC and TAS/SC schemes are given by

$$F_{\gamma_k}^{\mathrm{TAS/MRC,Prop}}(\gamma) = \left\{ \frac{F_{\mathcal{Y}_k}(\mu)}{N_R!} \left(\frac{\lambda_k \gamma}{\bar{\gamma}_P}\right)^{N_R} + \sum_{i=0}^{N-1} \frac{\phi(N,i)}{2 \times N_R!} \left(\frac{\lambda_k \gamma}{\hat{\lambda}_{kp}^{(i)} \bar{\gamma}_I}\right)^{N_R} \Gamma\left(N_R + 1, \hat{\lambda}_{kp}^{(i)} \mu\right) \right\}^{N_T}, \quad (25)$$

$$F_{\gamma_k}^{\mathrm{TAS/SC,Prop}}(\gamma) = \left\{ F_{\mathcal{Y}_k}(\mu) \frac{\lambda_k \gamma}{\bar{\gamma}_P} + \sum_{i=0}^{N-1} \frac{\phi(N,i)}{2} \frac{\lambda_k \gamma}{\hat{\lambda}_{kp}^{(i)} \bar{\gamma}_I} \Gamma\left(2, \hat{\lambda}_{kp}^{(i)} \mu\right) \right\}^{N_T N_R}, \quad (26)$$

where $F_{\mathcal{X}_k}(x) \approx (\lambda_k x)^{N_R}/N_R!$ and $F_{\left|h_k^{(i,j)}\right|^2}(x) \approx \lambda_k x$ are utilized when $x \to 0$ [51], which is evidently valid for our analysis in the high-SNR regime.

By utilizing **Proposition 3** for (25) and (26), the asymptotic BLER at link k for scheme $\mathcal{S}$ is determined as $\tilde{\varepsilon}_k^{\mathcal{S},\mathrm{Prop}} \cong F_{\gamma_k}^{\mathcal{S},\mathrm{Prop}}(\theta)$. Note that $\tilde{\varepsilon}_k^{\mathcal{S},\mathrm{Prop}}$ is very small when $\bar{\gamma}_P \to \infty$, yielding $\prod\limits_{k\ge 2} \tilde{\varepsilon}_k^{\mathcal{S},\mathrm{Prop}} \approx 0$. As a result, the asymptotic E2E BLER for scheme $\mathcal{S}$ in the proportional tolerable interference power constraint scenario is derived as

$$\tilde{\varepsilon}_{\mathrm{E2E}}^{\mathcal{S},\mathrm{Prop}} \approx \sum_{k=1}^{K+1} \tilde{\varepsilon}_k^{\mathcal{S},\mathrm{Prop}}. \quad (27)$$

Corollary 1: Based on the derived asymptotic results under the proportional $\mathcal{I}_p$ constraint as $\bar{\gamma}_P \to \infty$, the diversity and array gains for scheme $\mathcal{S}$ are given respectively by $\mathcal{D}_{\mathcal{S}} = N_T N_R$ and

$$\mathcal{G}_{\mathcal{S}} = \begin{cases} \min\limits_{k=1,\ldots,K+1} (\Psi_k)^{-\frac{1}{N_R}}, & \text{if } \mathcal{S} = \mathrm{TAS/MRC}, \\ \min\limits_{k=1,\ldots,K+1} (\Xi_k)^{-1}, & \text{if } \mathcal{S} = \mathrm{TAS/SC}, \end{cases} \quad (28)$$

where $\Xi_k = F_{\mathcal{Y}_k}(\mu)\lambda_k\theta + \sum\limits_{i=0}^{N-1} \frac{\phi(N,i)}{2} \frac{\lambda_k \theta}{\hat{\lambda}_{kp}^{(i)} \mu} \Gamma\left(2, \hat{\lambda}_{kp}^{(i)} \mu\right)$ *and*

$$\Psi_k = \frac{F_{\mathcal{Y}_k}(\mu)}{N_R!} (\lambda_k \theta)^{N_R} + \sum_{i=0}^{N-1} \frac{\phi(N,i)}{2 \times N_R!} \left(\frac{\lambda_k \theta}{\hat{\lambda}_{kp}^{(i)} \mu}\right)^{N_R} \Gamma\left(N_R + 1, \hat{\lambda}_{kp}^{(i)} \mu\right). \quad (29)$$

Proof: The diversity gain is calculated by [52]

$$\mathcal{D}_{\mathcal{S}} = -\lim_{\bar{\gamma}_P \to \infty} \frac{\log \tilde{\varepsilon}_{\mathrm{E2E}}^{\mathcal{S},\mathrm{Prop}}}{\log \bar{\gamma}_P} = N_T N_R. \quad (30)$$

For the diversity gain, rewritting the asymptotic E2E BLER in (27) yields

$$\tilde{\varepsilon}_{\mathrm{E2E}}^{\mathcal{S},\mathrm{Prop}} \approx \begin{cases} \sum_{k=1}^{K+1} (\Psi_k)^{N_T} (\bar{\gamma}_P)^{-N_T N_R}, & \text{if } \mathcal{S} = \mathrm{TAS/MRC}, \\ \sum_{k=1}^{K+1} (\Xi_k)^{N_T N_R} (\bar{\gamma}_P)^{-N_T N_R}, & \text{if } \mathcal{S} = \mathrm{TAS/SC}. \end{cases} \tag{31}$$

By mapping (31) to the form of $\tilde{\varepsilon}_{\mathrm{E2E}}^{\mathcal{S},\mathrm{Prop}} \approx (\mathcal{G}_{\mathcal{S}} \bar{\gamma}_P)^{-\mathcal{D}_{\mathcal{S}}}$ *[9], we determine (28). The proof is concluded.* □

Remark 5: The maximum achievable diversity gains for both the TAS/MRC and TAS/SC schemes are only feasible under the condition that $\bar{\gamma}_I = \mu \bar{\gamma}_P$ *as* $\bar{\gamma}_P \to \infty$. *When* $\mathcal{I}_p$ *is fixed and* $\bar{\gamma}_P \to \infty$, *the resultant E2E BLER for scheme* $\mathcal{S}$ *becomes saturated, and hence the achievable diversity order cannot be guaranteed, which results in no diversity gain for scheme* $\mathcal{S}$.

Remark 6: The results in ***Corollary 1*** *have revealed that TAS/MRC and TAS/SC offer the same maximum diversity gains. Therefore, the tradeoff between them is solely characterized by their array gains. As a result, the SNR gap between TAS/MRC and TAS/SC is represented by the ratio of their respective array gains, given by* $\mathrm{SNR}_{\mathrm{Gap}} = \frac{\mathcal{G}_{\mathrm{TAS/MRC}}}{\mathcal{G}_{\mathrm{TAS/SC}}} = \min_{k=1,\ldots,K+1} \frac{\Xi_k}{(\Psi_k)^{\frac{1}{N_R}}}$. *This indicates that, for the same E2E BLER, the power savings of TAS/MRC over TAS/SC are* $10\log_{10}\mathrm{SNR}_{\mathrm{Gap}}$.

Remark 7 (Observations From the CDFs): When the approximated E2E BLERs can be represented directly via the CDFs, observing the tendency of these CDFs is beneficial. Based on ***Proposition 1*** *and* ***Proposition 2****, we can observe that the CDFs primarily depend on the parameters of* $\bar{\gamma}_I$, $\bar{\gamma}_P$, N, N_T, N_R, λ_{kp}, λ_k, *and* β, *where* $\bar{\gamma}_I$ *and* $\bar{\gamma}_P$ *are major concerns in the CR environment. Particularly, when* $\bar{\gamma}_I = \mu \bar{\gamma}_P$ *(i.e., proportional tolerable interference power constraint),* $F_{\gamma_k}^{\mathcal{S}}$ *linearly decreases to zero as* $\bar{\gamma}_P \to \infty$. *This is because increasing* $\bar{\gamma}_P$ *to infinity leads to* $F_{\gamma_k}^{\mathrm{TAS/MRC}} \to F_{\gamma_k}^{\mathrm{TAS/MRC,Prop}}$ *in (25) and* $F_{\gamma_k}^{\mathrm{TAS/SC}} \to F_{\gamma_k}^{\mathrm{TAS/SC,Prop}}$ *in (26). It is observed in (25) and (26) that* $F_{\gamma_k}^{\mathcal{S},\mathrm{Prop}} \to 0$ *when* $\bar{\gamma}_P, \bar{\gamma}_I \to \infty$. *Contradictorily, when* $\bar{\gamma}_P \to \infty$ *and all the other parameters are fixed (i.e., near-CR asymptotic regime),* $F_{\gamma_k}^{\mathrm{TAS/MRC}}$ *and* $F_{\gamma_k}^{\mathrm{TAS/SC}}$ *converge to the floor constants (19) and (20), respectively.*

V. OPTIMIZATION PROBLEMS

This section is dedicated to addressing the optimization problems for the considered system, including the PA and relay placement.

A. OPTIMAL PA

For a given relay placement, we optimally allocate the transmit power to minimize the E2E BLER. In particular, we formulate the E2E BLER as an objective function under the constraints of the total transmit power in the secondary network, denoted by P_{tot}, and the maximum allowable interference power limit per PU. For simplicity, the asymptotic E2E BLER under the proportional tolerable interference power constraint is considered. We note that our simplified consideration is valid for the uRLLC requirements when the E2E BLER falls below 10^{-5}. As a result, the PA optimization problem becomes

$$\min_{\mathbf{P}} \tilde{\varepsilon}_{\mathrm{E2E}}^{\mathcal{S},\mathrm{Prop}}(\mathbf{P}) = \begin{cases} \min_{\mathbf{P}} \sum_{k=1}^{K+1} (\Psi_k)^{N_T} (\bar{\gamma}_P)^{-N_T N_R}, & \text{if } \mathcal{S} = \mathrm{TAS/MRC}, \\ \min_{\mathbf{P}} \sum_{k=1}^{K+1} (\Xi_k)^{N_T N_R} (\bar{\gamma}_P)^{-N_T N_R}, & \text{if } \mathcal{S} = \mathrm{TAS/SC}, \end{cases} \tag{32a}$$

$$\text{subject to } \sum_{k=1}^{K+1} \hat{P}_{k-1} \le P_{\mathrm{tot}}, \tag{32b}$$

$$\hat{P}_{k-1} \le P_s, \tag{32c}$$

$$\sum_{k=1}^{K+1} \hat{\lambda}_{kp} \hat{P}_{k-1} \le I_p, \tag{32d}$$

where $\mathbf{P} = \left(\hat{P}_0, \hat{P}_1, \ldots, \hat{P}_K\right)$ denotes a set of multiple transmit power variables and $\hat{\lambda}_{kp} = \left(\beta + \sqrt{1-\beta^2}\right) \lambda_{kp}$.

It is worth noting that the asymptotic E2E BLER is convex and the constraints are linear. Hence, (32) can be solved as a convex optimization problem, in which the Lagrangian multiplier method is beneficial [53]. Besides that, it is observed that (32) can be split into two constrained cases: $P_{\mathrm{tot}} \le \sum_{k=1}^{K+1} \frac{\mathcal{I}_p}{\hat{\lambda}_{kp}}$ and otherwise [9].

1) CASE 1

When $P_{\mathrm{tot}} \le \sum_{k=1}^{K+1} \frac{\mathcal{I}_p}{\hat{\lambda}_{kp}}$ holds, (32) can be rewritten as

$$\min_{\mathbf{P}} \tilde{\varepsilon}_{\mathrm{E2E}}^{\mathcal{S},\mathrm{Prop}}(\mathbf{P}) = \begin{cases} \min_{\mathbf{P}} \sum_{k=1}^{K+1} (\Psi_k)^{N_T} \left(\frac{\mathcal{N}_0}{\hat{P}_{k-1}}\right)^{N_T N_R}, & \text{if } \mathcal{S} = \mathrm{TAS/MRC}, \\ \min_{\mathbf{P}} \sum_{k=1}^{K+1} (\Xi_k)^{N_T N_R} \left(\frac{\mathcal{N}_0}{\hat{P}_{k-1}}\right)^{N_T N_R}, & \text{if } \mathcal{S} = \mathrm{TAS/SC}, \end{cases} \tag{33a}$$

$$\text{subject to } \sum_{k=1}^{K+1} \hat{\lambda}_{kp} \hat{P}_{k-1} \Big/ \sum_{l=1}^{K+1} \hat{\lambda}_{lp} \le P_{\mathrm{tot}}. \tag{33b}$$

Based on (33), the Lagrangian functions with the Karush–Kuhn–Tucker conditions for the TAS/MRC and TAS/SC schemes are given, respectively, by

$$\mathcal{L}_{1,\mathrm{TAS/MRC}}\left(\mathbf{P}^*, \varsigma_1\right) = \sum_{k=1}^{K+1} (\Psi_k)^{N_T} \left(\frac{N_0}{\hat{P}_{k-1}^*}\right)^{N_T N_R}$$

$$+\varsigma_1\left(\sum_{k=1}^{K+1}\hat{\lambda}_{kp}\hat{P}^*_{k-1}\Big/\sum_{l=1}^{K+1}\hat{\lambda}_{lp}-P_{\text{tot}}\right), \tag{34}$$

$$\mathcal{L}_{1,\text{TAS/SC}}\left(\mathbf{P}^*,\varsigma_1\right) = \sum_{k=1}^{K+1}(\Xi_k)^{N_TN_R}\left(\frac{N_0}{\hat{P}^*_{k-1}}\right)^{N_TN_R} + \varsigma_1\left(\sum_{k=1}^{K+1}\hat{\lambda}_{kp}\hat{P}^*_{k-1}\Big/\sum_{l=1}^{K+1}\hat{\lambda}_{lp}-P_{\text{tot}}\right), \tag{35}$$

where $\mathbf{P}^* = \left(\hat{P}^*_0, \hat{P}^*_1, \ldots, \hat{P}^*_K\right)$ is the optimal solution and ς_1 denotes a nonzero Lagrange multiplier constant.

By setting the derivatives of $\mathcal{L}_{1,\mathcal{S}}(\mathbf{P}^*,\varsigma_1)$ with respect to $\hat{P}^*_{k-1}$ and ς_1 to zero and subsequently solving a set of $K+1$ equations, we obtain the optimal PA solutions for the TAS/MRC and TAS/SC schemes in *Case 1*, respectively, in the forms of

$$\hat{P}^{*\text{TAS/MRC}}_{k-1} = \frac{P_{\text{tot}}\sum_{l=1}^{K+1}\hat{\lambda}_{lp}}{\hat{\lambda}_{kp}\sum_{u=1}^{K+1}\left(\frac{\Psi_u}{\Psi_k}\right)^{\frac{N_T}{N_TN_R+1}}\left(\frac{\hat{\lambda}_{up}}{\hat{\lambda}_{kp}}\right)^{\frac{N_TN_R}{N_TN_R+1}}}, \tag{36}$$

$$\hat{P}^{*\text{TAS/SC}}_{k-1} = \frac{P_{\text{tot}}\sum_{l=1}^{K+1}\hat{\lambda}_{lp}}{\hat{\lambda}_{kp}\sum_{u=1}^{K+1}\left(\frac{\Xi_u\hat{\lambda}_{up}}{\Xi_k\hat{\lambda}_{kp}}\right)^{\frac{N_TN_R}{N_TN_R+1}}}. \tag{37}$$

2) CASE 2

Otherwise, $P_{\text{tot}} > \sum_{k=1}^{K+1}\frac{\mathcal{I}_p}{\hat{\lambda}_{kp}}$, and (32) can be formulated as

$$\min_{\mathbf{P}}\tilde{\varepsilon}^{\mathcal{S},\text{Prop}}_{\text{E2E}}(\mathbf{P}) = \begin{cases}\min\limits_{\mathbf{P}}\sum\limits_{k=1}^{K+1}(\Psi_k)^{N_T}\left(\frac{\mathcal{N}_0}{\hat{P}_{k-1}}\right)^{N_TN_R}, & \text{if } \mathcal{S}=\text{TAS/MRC},\\ \min\limits_{\mathbf{P}}\sum\limits_{k=1}^{K+1}(\Xi_k)^{N_TN_R}\left(\frac{\mathcal{N}_0}{\hat{P}_{k-1}}\right)^{N_TN_R}, & \text{if } \mathcal{S}=\text{TAS/SC},\end{cases} \tag{38a}$$

$$\text{subject to}\ \sum_{k=1}^{K+1}\hat{\lambda}_{kp}\hat{P}_{k-1}\le\mathcal{I}_p. \tag{38b}$$

Similarly, by utilizing the Lagrangian multiplier method [53] to solve (38), we obtain the optimal PA solutions for the TAS/MRC and TAS/SC schemes in *Case 2*, respectively, as

$$\hat{P}^{*\text{TAS/MRC}}_{k-1} = \frac{\mathcal{I}_p}{\hat{\lambda}_{kp}\sum_{u=1}^{K+1}\left(\frac{\Psi_u}{\Psi_k}\right)^{\frac{N_T}{N_TN_R+1}}\left(\frac{\hat{\lambda}_{up}}{\hat{\lambda}_{kp}}\right)^{\frac{N_TN_R}{N_TN_R+1}}}, \tag{39}$$

$$\hat{P}^{*\text{TAS/SC}}_{k-1} = \frac{\mathcal{I}_p}{\hat{\lambda}_{kp}\sum_{u=1}^{K+1}\left(\frac{\Xi_u\hat{\lambda}_{up}}{\Xi_k\hat{\lambda}_{kp}}\right)^{\frac{N_TN_R}{N_TN_R+1}}}. \tag{40}$$

B. OPTIMAL RL

This subsection aims to optimally allocate the RL to minimize the E2E BLER. When the PA is defined, we consider the constraint on the RL such that $\sum_{k=1}^{K+1}d_k = D$, where a linear topology is considered, D denotes the normalized transmission distance, and d_k represents the distance between two adjacent nodes. Similar to Section V-A, we utilize the asymptotic E2E BLER under the proportional tolerable interference power constraint as an objective function, which is valid for uRLLCs to simplify our optimization problems. Hence, the RL optimization problem is formulated as

$$\min_{\mathbf{d}}\tilde{\varepsilon}^{\mathcal{S},\text{Prop}}_{\text{E2E}}(\mathbf{d}) = \begin{cases}\min\limits_{\mathbf{d}}\sum\limits_{k=1}^{K+1}\frac{\left(\tilde{\Psi}_k\right)^{N_T}(\mathcal{N}_0)^{N_TN_R}}{\hat{P}_{k-1}^{N_TN_R}d_k^{\eta N_TN_R}}, & \text{if } \mathcal{S}=\text{TAS/MRC},\\ \min\limits_{\mathbf{d}}\sum\limits_{k=1}^{K+1}\frac{\left(\tilde{\Xi}_k\right)^{N_TN_R}(\mathcal{N}_0)^{N_TN_R}}{\hat{P}_{k-1}^{N_TN_R}d_k^{\eta N_TN_R}}, & \text{if } \mathcal{S}=\text{TAS/SC},\end{cases} \tag{41a}$$

$$\text{subject to}\ \sum_{k=1}^{K+1}d_k = D, \tag{41b}$$

where the simplified path-loss model with $\lambda_k = d_k^{-\eta}$ is employed [31], η represents the path-loss exponent, $\tilde{\Psi}_k = \frac{\Psi_k}{\lambda_k^{N_R}}$, $\tilde{\Xi} = \frac{\Xi}{\lambda_k}$, and $\mathbf{d} = (d_1, d_2, \ldots, d_{K+1})$ is a set of multiple distance variables.

When the objective function of (41) is convex and the constraint is linear, (41) can be solved as a convex optimization problem by using the Lagrangian multiplier method [53]. As a result, we obtain the optimal RL configuration for both the TAS/MRC and TAS/SC schemes as

$$d^*_k = \frac{D}{1+\sum_{u=1,u\neq k}^{K+1}\left(\frac{\hat{P}_{k-1}}{\hat{P}_{u-1}}\right)^{\frac{N_TN_R}{\eta N_TN_R+1}}}. \tag{42}$$

Remark 8: In practice, the optimal relay coordinates can be determined easily based on the derived optimal distances established in (42) because of the linear topology consideration.

Remark 9: The given closed-form solutions to the optimization problems require negligible computational complexity, which is beneficial compared with search methods that require much higher complexity.

VI. MACHINE-LEARNING APPLICATIONS

In this section, we employ some state-of-the-art ML models, including multilayer perceptrons (i.e., DNN), CNN, and a decision-tree-based ensemble ML algorithm (i.e., XGB), to estimate the E2E performance in terms of the system throughput, EE, latency, and reliability. For notational simplicity, we define $\mathcal{Q}^{\mathcal{S}}_{\text{E2E}}$ as the output of the training data, i.e, $\mathcal{Q}^{\mathcal{S}}_{\text{E2E}} \in \left\{\text{TP}^{\mathcal{S}}_{\text{E2E}}, \text{EE}^{\mathcal{S}}_{\text{E2E}}, \text{Latency}^{\mathcal{S}}_{\text{E2E}}, \text{Reliability}^{\mathcal{S}}_{\text{E2E}}\right\}$.

A. MOTIVATION AND ML PROCESS DESCRIPTION

Unlike simulation and numerical integration methods that are model-based approaches requiring long run times to converge to the exact results, ML frameworks are data-driven approaches that do not require mathematical derivations. As a result, ML frameworks are rationally designed to enable real-time performance evaluations with considerably high predicted accuracies and short execution times. In addition, when future wireless networks become more complex and heterogeneous, closed-form expressions can become intractable; this will potentially necessitate data-driven approaches for real-time configurations of such future wireless networks.

ML-based performance evaluation comprises two phases: training and prediction. During the training phase, the ML model is trained using historical data or simulated scenarios. The training phase typically requires significant computational resources and can be conducted offline on any computing device. It does not necessarily need to be performed on the network nodes. Conversely, the prediction phase can be conducted on the network nodes where the ML model is deployed because it typically requires less computational resources compared to the training phase. The trained model can deliver high accuracy and real-time estimation of performance metrics $\mathcal{Q}_{\text{E2E}}^{\mathcal{S}}$ whenever any new information is available at the input.

B. DATASET GENERATION

The data is generated for training based on the first-order Riemann integral approximation, as stated in **Proposition 3**, and equations (15)–(18). The related system parameters are chosen via an artificial stochastic process within their range. Specifically, $N \in [1, 5]$, $K \in [0, 10]$, $N_T \in [1, 5]$, $N_R \in [1, 5]$, $\overline{\gamma}_P \in [0, 35]$ dB, $\overline{\gamma}_I \in [15, 20]$ dB, and the correlation coefficient $\beta \in [0, 1]$. Suppose that the information length of $\mathcal{B} \in [64, 8192]$ bits are encoded into the blocklength with the number of $\tau \in \left[10^2, 10^5\right]$ CUs. Then, the transmission rate $\mathcal{R} \in \left[6.4 \times 10^{-4}, 81.92\right]$ can be used as an input variable. In short, the input features for the ML-based estimation is vectorized as $\boldsymbol{x} = [\bar{\gamma}_P, \bar{\gamma}_I, \mathcal{R}, N_T, N_R, K, N, \beta]^T$. A dataset with one million samples is generated based on these assumptions and 80% of the data points are used for training.

C. DESCRIPTION OF MODELS

1) MULTILAYER PERCEPTRONS

As shown in Fig. 2(a), the proposed methodology with multilayer perceptions, also known as a DNN or feedforward neural network, comprises L_1 layers, including an input layer, $(L_1 - 2)$ hidden layers, and a single node in the output layer. Let Λ_0 and Λ_l be the number of input features and the number of neurons in the lth layer of the DNN, respectively. Then, the estimated output is obtained via consecutive nonlinear transformations as

$$\widehat{\mathcal{Q}}_{\text{E2E}}^{\mathcal{S}} = \mathcal{T}_{\text{DNN}}^{[L_1]}\left(\ldots\left(\mathcal{T}_{\text{DNN}}^{[1]}\left(\boldsymbol{x}^{[0]}; \boldsymbol{\eta}_1^{[1]}\right); \ldots\right), \boldsymbol{\eta}_1^{[L_1]}\right), \quad (43)$$

where $\boldsymbol{\eta}_1^{[l]} = \left\{\boldsymbol{W}_1^{[l]}, \boldsymbol{b}_1^{[l]}\right\}$ contains the weight matrix $\boldsymbol{W}_1^{[l]} \in \mathbb{R}^{\Lambda_l \times \Lambda_{l-1}}$ and bias vector $\boldsymbol{b}_1^{[l]} \in \mathbb{R}^{\Lambda_l \times 1}$ between the $(l-1)$th and lth layers of the model, whereas $\boldsymbol{x}^{[l]} \in \mathbb{R}^{\Lambda_l \times 1}$ represents the output of the lth layer. In (43), the transformation function $\mathcal{T}_{\text{DNN}}{}^{[l]}(\cdot)$ is defined as

$$\mathcal{T}_{\text{DNN}}^{[l]}\left(\boldsymbol{x}^{[l-1]}; \boldsymbol{\eta}_1^{[l]}\right) = \mathcal{A}^{[l]}\left(\boldsymbol{W}_1^{[l]}\boldsymbol{x}^{[l-1]} + \boldsymbol{b}_1^{[l]}\right), \quad (44)$$

where $l \in [1, L_1]$ and $\mathcal{A}^{[l]}(\cdot)$ denotes an activation function at layer l's output, which is determined by a scaled exponential linear unit for its self-normalizing property and fast training convergence [54]. We note that the proposed DNN model is designed to estimate the system performance $\mathcal{Q}_{\text{E2E}}^{\mathcal{S}}$ as a regression problem. As a result, the mean squared error (MSE) is employed as the loss function, which is given by $\mathcal{J}\left(\left\{\boldsymbol{\eta}_1^{[l]}\right\}_{l=1}^{L_1}\right) = \frac{1}{\Upsilon}\sum_{b=1}^{\Upsilon}\left[\mathcal{Q}_{\text{E2E},b}^{\mathcal{S}} - \widehat{\mathcal{Q}}_{\text{E2E},b}^{\mathcal{S}}\right]^2$, where Υ is the number of data points that are stochastically selected from the training set in each training iteration, whereas $\mathcal{Q}_{\text{E2E},b}^{\mathcal{S}}$ and $\widehat{\mathcal{Q}}_{\text{E2E},b}^{\mathcal{S}}$ are the expected and estimated $\mathcal{Q}_{\text{E2E}}^{\mathcal{S}}$ corresponding to the bth sample, respectively.

2) CONVOLUTIONAL NEURAL NETWORK

As can be seen in Fig. 2(b), the structure of our proposed CNN comprises multiple 1-dimensional convolution layers, a flattening layer, and fully connected layers. In the lth convolution, $\mathcal{C}$ filters with size $h \times 1$ are employed to create the feature map that summarizes the presence of specific features of the input. The cth channel of this convolution result is calculated by

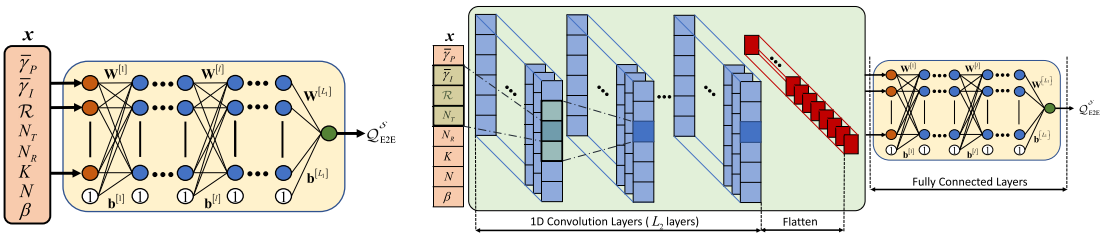
$$T_{\text{CNN}}^{[l,c]}\left(\boldsymbol{x}^{[l-1]}; \left\{\boldsymbol{W}_2^{[l]}, \boldsymbol{b}_2^{[l]}\right\}\right) = \mathcal{A}^{[l]}\left(\sum_{c'=1}^{\mathcal{C}} \boldsymbol{W}_2^{[l,c,c']} * \boldsymbol{x}^{[l-1,c']} + \boldsymbol{b}_2^{[l,c]}\right), \quad (45)$$

where $l \in [1, L_2]$, $c \in [1, \mathcal{C}]$, $(*)$ is the convolutional operation, L_2 is the number of convolution layers, and $\boldsymbol{W}_2^{[l]}$ and $\boldsymbol{b}_2^{[l]}$ are the tensors of the weighting kernels and biases, respectively. Unlike other CNN applications (e.g., image and video processing), the amount of input information in this work is small; therefore, we utilize a padding method and single-stride convolutions to avoid rapid downsampling on the feature spatial sizes. After feature extraction, the data is flattened for conversion into a 1-dimensional array to feed the subsequent layers. From this stage, the adjacent fully connected block estimating $\mathcal{Q}_{\text{E2E}}^{\mathcal{S}}$ is similar to the DNN in the previous subsection.

3) XGB ALGORITHM

The XGB model employs L_3 consecutive trees to predict $\mathcal{Q}_{\text{E2E}}^{\mathcal{S}}$ by summing the output values of all estimators as

$$\widehat{\mathcal{Q}}_{\text{E2E}}^{\mathcal{S}} = \sum_{l=1}^{L_3} \alpha_l \mathcal{T}_{\text{XGB}}^{[l]}(\boldsymbol{x}), \quad (46)$$



(a) Deep neural network (b) Convolution neural network

FIGURE 2. Description of the deep learning (DL)-aided estimators.

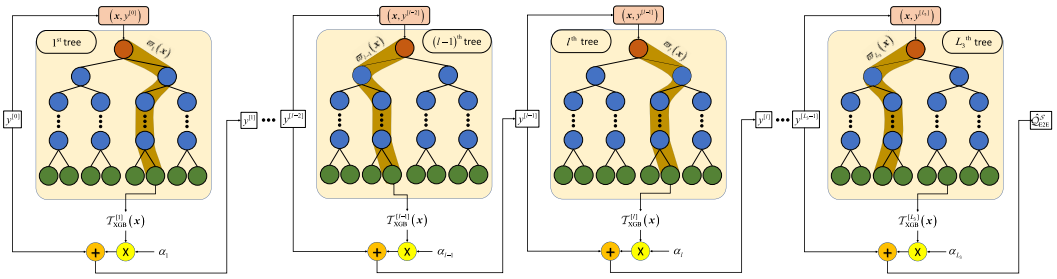


FIGURE 3. Description of the decision trees in the XGB algorithm.

where $\mathcal{T}_{\text{XGB}}^{[l]}(\boldsymbol{x})$ and α_l represent a nonlinear transformation and a learning rate chosen at the lth iteration, respectively. As shown in Fig. 3, the numbers of leaves are equally configured as ℓ for all trees. Each leaf of the lth regression tree is determined as a continuous value, denoted by $w_q^{[l]}$, for $q \in [1, \ell]$. For a given sample $\boldsymbol{x}$, a set of decision rules $\boldsymbol{\varpi}$ assigns an input toward a target leaf. In other words, we have $\mathcal{T}_{\text{XGB}}^{[l]}(\boldsymbol{x}) = w_q^{[l]}$ for $q = \boldsymbol{\varpi}_l(\boldsymbol{x})$. Furthermore, we assume that the predicted value at the lth tree's output is modeled as $y^{[l]} = y^{[l-1]} + \alpha_l \mathcal{T}_{\text{XGB}}^{[l]}(\boldsymbol{x})$, where $y^{[0]}$ equals the mean value of labels in the training dataset, which yields $y^{[L_3]} = \widehat{\mathcal{Q}}_{\text{E2E}}^{\mathcal{S}}$. According to the additive training, the regularized MSE loss function at the lth tree is given by [55]

$$\mathcal{J}\left(\left\{w_q^{[l]}\right\}_{q=1}^{\ell}\right) = \sum_{q=1}^{\ell}\left\{\sum_{b\in I_q^{[l]}} \frac{2}{\Upsilon}\left[\mathcal{Q}_{\text{E2E},b}^{\mathcal{S}} - y_b^{[l-1]}\right] w_q^{[l]} + \left(\sum_{b\in I_q^{[l]}} \frac{1}{\Upsilon} + \chi\right)\left(w_q^{[l]}\right)^2\right\}, \tag{47}$$

where χ represents a regularization parameter to reduce the overfitting problem during the training phase and $I_q^{[l]} = \left\{b|\pi^{[l]}(\boldsymbol{x}_b) = q\right\}$ contains the indices of the data points oriented to the qth leaf of the lth tree with input $\boldsymbol{x}_b$, whereas $y_b^{[l]}$ and $\mathcal{Q}_{\text{E2E},b}^{\mathcal{S}}$ are the estimated residual value at step l and the expected value of $\mathcal{Q}_{\text{E2E}}^{\mathcal{S}}$ corresponding to the bth sample point, respectively.

D. ML–AIDED PERFORMANCE PREDICTION

When the offline training process is finished, the resulting ML model comprising weights and biases can be utilized to obtain performance prediction through a compact mapping function, i.e.,

$$\hat{\mathcal{Q}}_{\text{E2E}}^{\mathcal{S}} = \mathcal{M}(\boldsymbol{x}). \tag{48}$$

The trained network is used for the online prediction of performance metrics $\mathcal{Q}_{\text{E2E}}^{\mathcal{S}}$ whenever any new information is available at the input. In addition, (48) indicates that the predicted output is achieved without the requirement of complex operations, implying a significant reduction in the execution time that enables real-time performance prediction. We note that the prediction phase can be conducted on the network nodes where the ML model is deployed because it typically necessitates requires computational resources in contrast to the training phase.

E. COMPLEXITY ANALYSIS

The computational complexity of ML models is measured based on the number of floating-point operations (FLOPs). According to (44), the total number of FLOPs of the proposed DNN is given by [56]

$$\mathcal{O}_{\text{DNN}}\left(\{\Lambda_l\}_{l=0}^{L_1}\right) = 2\sum_{l=1}^{L_1}\Lambda_{l-1}\Lambda_l. \tag{49}$$

In the CNN architecture, the padding method guarantees that the output size in each channel remains unchanged; therefore, we have $\boldsymbol{x}^{[l]} \in \mathbb{R}^{\Lambda_0\times 1\times \mathcal{C}}$, for $l \in [1, L_2]$. According to (45), the total number of FLOPs of the lth convolution

layer with $l \in [2, L_2]$ is $2\Lambda_0 h\mathcal{C}^2$. When there is only a single channel in the input layer, the total number of FLOPs of the first convolution layer is $2\Lambda_0 h\mathcal{C}$. Let $\widetilde{L}_1$ and $\widetilde{\Lambda}_l$ be the number of fully-connected layers and the hidden neurons in the lth layer with $l \in \left[0, \widetilde{L}_1\right]$, respectively. Then, the total complexity of the proposed CNN model is given by

$$\mathcal{O}_{\text{CNN}}\left(\Lambda_0, h, \mathcal{C}, L_2, \{\widetilde{\Lambda}_l\}_{l=0}^{\widetilde{L}_1}\right) = 2\Lambda_0 h\mathcal{C}\left(L_2\mathcal{C}+1\right) + \mathcal{O}_{\text{DNN}}\left(\{\widetilde{\Lambda}_l\}_{l=0}^{\widetilde{L}_1}\right). \quad (50)$$

The number of FLOPs associated with employing a trained XGB system is mainly a comparison of the feature values in all nodes. In practical applications, the lth decision tree with a depth of $\mathcal{D}_l$ infers the result of $\mathcal{T}_{\text{XGB}}^{[l]}(\boldsymbol{x})$ after $(\mathcal{D}_l - 1)$ FLOPs. Based on (46), the required operations of the analysis based on the tree boosting algorithm are given by

$$\mathcal{O}_{\text{XGB}}\left(\{\mathcal{D}_l\}_{l=0}^{L_3}\right) = \sum_{l=1}^{L_3}\left(\mathcal{D}_l - 1\right) + 2L_3 = \sum_{l=1}^{L_3}\mathcal{D}_l + L_3. \quad (51)$$

VII. NUMERICAL RESULTS AND DISCUSSION

In this section, Monte Carlo simulations have been performed to verify our theoretical analysis. Furthermore, the state-of-the-art ML-based applications, including DNN, CNN, and XGB, are utilized to predict the E2E throughput, EE, latency, and reliability.

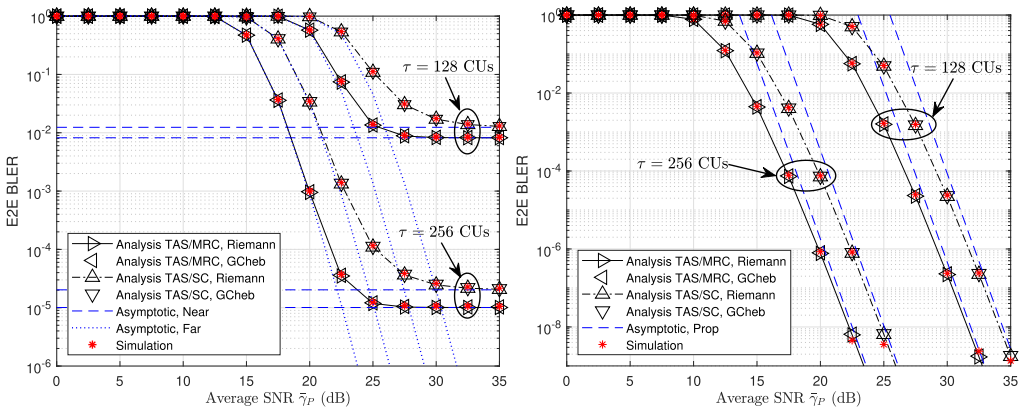
A. PERFORMANCE EVALUATION

This subsection presents the numerical results for the achievable E2E performance of our proposed system with the short-packet uRLLCs. For channel settings, we consider a two-dimensional plane where PU_n, SU_0, SU_{K+1}, and SU_k are located at coordinates (0.5, 0.5), (0, 0), (1, 0), and $\left(\frac{k}{K+1}, 0\right)$, respectively. Unless otherwise stated, we assume equal PA, $D = 1$, $\eta = 3$, $T = 50$, $\beta = 0.8$, $K = 3$, $N = 2$, $N_T = N_R = 3$, $\bar{\gamma}_P = \bar{\gamma}_I = 20$ dB, $\mathcal{B} = 1280$ bits, and $\tau = 128$ CUs. In terms of training the DL models, the DNN and CNN models are E2E trained with 200 epochs with weights randomly initialized by utilizing the Adam optimizer with a gradient decay factor of 0.95, whereas the learning rate is initialized with 10^{-3} for the first 20 epochs and decreased to 90% every 10 epochs. In the XBG model, the learning factor is kept constant as $\alpha_l = 0.3$ in all trees, and we assume that all trees have the same depth as $\mathcal{D}_l = \mathcal{D}$ for $l \in [1, L_3]$.

Fig. 4 compares the E2E BLER of the proposed TAS/SC and TAS/MRC schemes for the underlay cognitive multihop relay SPC system, where two practical approaches, namely, the fixed $\bar{\gamma}_I = 20$ dB and proportional $\bar{\gamma}_I = \bar{\gamma}_P$, are shown in Fig. 4(a) and Fig. 4(b), respectively. First, in both Fig. 4(a) and Fig. 4(b), it is readily observed that the TAS/MRC scheme achieves better performance than the TAS/SC scheme over the entire SNR range. We note that, as stated in **Remark 6**, there is no diversity gain in the fixed $\mathcal{I}_p$ scenario in Fig. 4(a), whereas the TAS/MRC scheme outperforms the TAS/SC scheme by a factor of $10\log_{10}\text{SNR}_{\text{Gap}}$ in Fig. 4(b). Second, in Fig. 4(a), the E2E BLER becomes saturated in the high-SNR regime as a result of the fixed tolerable interference power constraint, whereas in Fig. 4(b), the E2E performance is continuously improved with increasing average SNR $\bar{\gamma}_P$. We note that the E2E BLER floor in the fixed $\bar{\gamma}_I$ scenario can be improved significantly by adopting a longer blocklength. Third, because the Riemann integral analysis perfectly matches the Gauss–Chebyshev method, we only implement the Riemann approach in the subsequent figures for simplification. In Fig. 4(a), the near- and far-CR asymptotic results agree well with the analysis results in the high- and low-SNR regimes, respectively. Similarly, the proportional asymptotic results in Fig. 4(b) also converge to the analysis results in the high-SNR regime. More importantly, the Monte Carlo simulation results agree well with all analytical results, which confirms of our analysis.

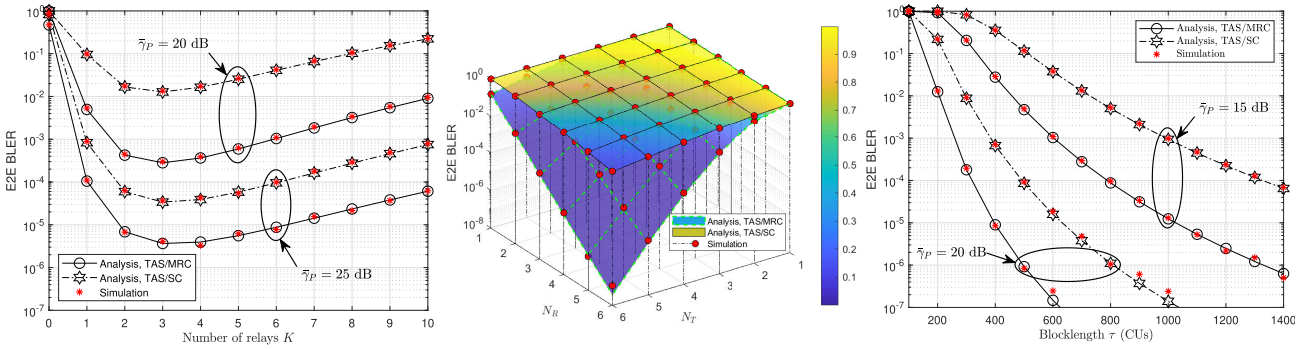
Fig. 5(a) presents the influence of the number of relays K on the system performance. It is clear that the E2E BLER as a function of K has convex formality, which yields the existence of an optimal K. In particular, $K = 3$ is the optimal choice in this simulation environment, which minimizes the E2E BLER. This behavior can be explained by increasing K leading to a smaller d_k, which reduces the BLER at each hop. As a result, the E2E BLER drops dramatically in the low-K range. Nevertheless, when K becomes too large, the number of time slots increases proportionally. In such a scenario, under the total transmit power budget P_{tot}, P_k becomes lower, which can result in performance degradation. In other words, when $K \to \infty$, the BLER at each hop approaches unity, i.e., $\bar{\varepsilon}_k^{\mathcal{S}} \to 1$. Based on (14), the E2E BLER tends to approach unity when $K \to \infty$. In Fig. 5(a), we also observe that (*i*) the TAS/MRC scheme offers better performance than the TAS/SC scheme and (*ii*) the system performance is substantially improved by increasing the average transmit SNR, which is consistent with the observations in Fig. 4.

The influences of MIMO implementation and blocklength are depicted in Fig. 5(b) and Fig. 5(c), respectively. As shown in Fig. 5(b), increasing N_T and N_R leads to substantially lower E2E BLERs, which is precise to the discussion in **Remark 4**. In addition, it also shows that the TAS/MRC scheme, represented by the lower layer, provides better performance than the TAS/SC scheme, corresponding to the upper layer. In Fig. 5(c), the longer blocklength offers higher E2E performance, which is consistent with the observation in Fig. 4. However, a short blocklength is expected in SPC systems, as it is selected to achieve low latency. Motivated by the tradeoff between the ultra-reliability and low-latency requirements, the value of τ should be chosen carefully. For instance, when a practical system with two relay nodes is assumed with maximum transmit power of 20 dB for each, to satisfy both the required E2E BLER on the order of 10^{-5} for ultra-reliability and minimal latency, the optimal



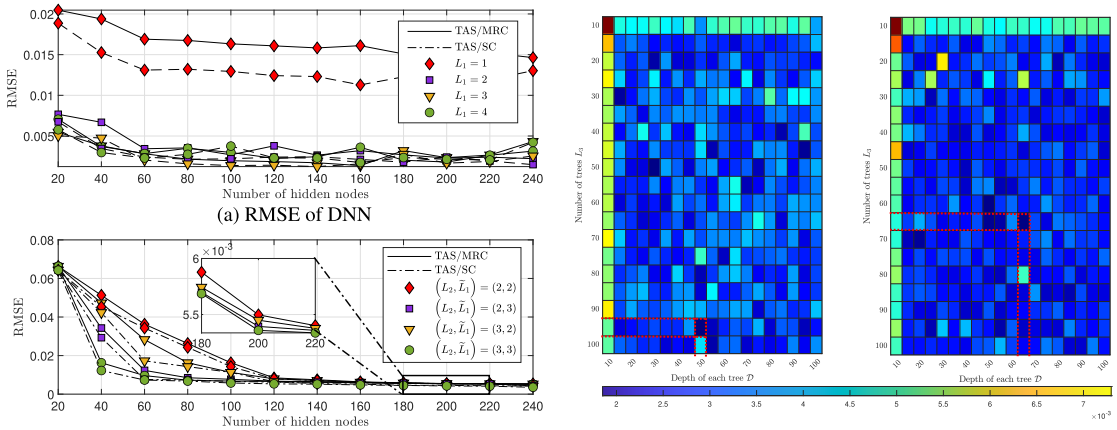
(a) Fixed $\bar{\gamma}_I = 20$ dB (b) Proportional $\bar{\gamma}_I = \bar{\gamma}_P$

FIGURE 4. E2E BLER of the MIMO underlay CR multihop relay network under SPCs in spectrum sharing.



(a) E2E BLER versus K (b) E2E BLER versus N_T and N_R (c) E2E BLER versus τ

FIGURE 5. The impact of K, N_T, N_R, and τ on the E2E BLER.



(a) RMSE of DNN (b) RMSE of CNN (c) RMSE of XGB on TAS/MRC (left) and TAS/SC (right) testset

FIGURE 6. RMSE validation of (a) DNN, (b) CNN with different numbers of hidden layers and neurons, and (c) XGB with different numbers of trees and varying tree depth.

blocklengths of $\tau = 400$ CUs and $\tau = 650$ CUs can be chosen for the TAS/MRC and TAS/SC schemes, respectively.

Fig. 6 compares the RMSE in the validation stage with different configurations of ML-aided estimators for both the TAS/MRC and TAS/SC schemes. In Fig. 6(a), the DNN models with a higher number of hidden layers achieve significantly lower RMSE values than those of the single-layer networks. Because the DNN model with $L_1 = 3$ at 140 hidden neurons offers the lowest validating RMSE, these parameters are chosen for later DNN experiments. Whereas increasing

TABLE 1. Average run time (ms) versus K of the ML-based and analytical computations.

(K, N)	XGB		DNN		CNN		Riemann		GCheb	
	TAS/MRC	TAS/SC	TAS/MRC	TAS/SC	TAS/MRC	TAS/SC	TAS/MRC	TAS/SC	TAS/MRC	TAS/SC
$(2, 4)$	0.30	0.29	45.10	48.21	50.11	48.57	0.80	0.66	32.10	22.16
$(4, 6)$	0.31	0.36	43.12	43.22	48.12	48.31	2.12	1.21	90.12	53.51
$(6, 8)$	0.29	0.31	43.22	40.12	46.44	46.67	3.50	2.41	131.22	97.53
$(8, 10)$	0.34	0.34	45.12	42.87	48.88	46.14	6.12	3.58	180.46	114.88

the number of hidden neurons causes a fluctuation in the RMSE of the DNN system, the RMSE values in the CNN model converge to a floor value, as illustrated in Fig. 6(b). Our aim is to design ML models that not only satisfy the lowest RMSE, but also that have low numbers of hidden neurons and hidden layers to achieve low complexity and short run times for real-time configurations. As a result, $L_2 = 2$ and $\widetilde{L}_1 = 3$ with a total of 180 hidden nodes are applied to the CNN estimator for later evaluations, which can balance the performance, execution time, and complexity. In the case of the XBG model, $(\mathcal{D}, L_3) = (50, 95)$ and $(\mathcal{D}, L_3) = (65, 65)$ are the configurations that provide the best RMSE performance on the TAS/MRC and TAS/SC testing datasets, respectively, as shown in Fig. 6(c). Therefore, these parameters are utilized for the XGB designs in subsequent simulations.

Table 1 shows the execution time of the ML-aided estimators and the computation based on the closed-form expressions. As can be observed, the amount of time spent on computing based on the analytical derivations significantly increases as K and N increase, whereas the ML approaches remain fixed regardless of execution time. In other words, the numbers of elements of the primary and secondary networks in the underlay CR multihop relay systems have no impact on the performance prediction. In Table 1, the XGB run time remains stable at around 0.3 ms. In contrast, for the TAS/MRC scheme, as K and N increase to $(K, N) = (8, 10)$, the Riemann and Gauss–Chebyshev approximations gradually increase to 6.12 ms and 180.46 ms, which are approximately 20 and 600 times higher than the XGB prediction, respectively. Furthermore, the estimation based on the XGB algorithm always requires the shortest prediction time. This is because the DL models need remarkably many layers and neurons to achieve the lowest RMSE, as shown in Fig. 6. This results in substantial computation time, as shown in (49) and (50), whereas the XGB model applies comparison operators, which have a lower complexity of $\mathcal{O}(\mathcal{D}L_3)$. Subsequently, Fig. 7 illustrates the performance of different ML-based estimators on the prediction ability of the system throughput, latency, EE, and reliability. It can be seen that, the estimated results provided by the XGB algorithm tightly fit the curve of the analytical calculation. In contrast, when SNR is less than 18 dB, the predictions based on the DL models (i.e., CNN and DNN) have larger errors, which can be observed in Fig. 7(a) and Fig. 7(b). According to the evidence in Table 1 and Fig. 7, we can conclude that the XGB model provides superior performance in predicting the E2E performance values. In later simulations, we only show the estimated results of the XGB model for comparison of the system performance.

The benefits of SPCs compared to long-packet communications (LPCs) are ultra-reliability and low latency, as shown in Fig. 8. We can observe that short-message scenarios (e.g., 128 bytes) offer higher reliability and lower latency than long-message scenarios (e.g., 512 and 1024 bytes). Encapsulating long messages into a long blocklength to guarantee ultra-reliability also causes higher latency. For example, 512 information bytes are encapsulated into packets with over 5500 CUs by using channel coding techniques (e.g., low-density parity check, polar codes, and Turbo codes) to satisfy the ultra-reliability requirement, as shown in Fig. 8(a). However, for a CU duration of 3 μs [57], Fig. 8(b) shows that the E2E latency is 22000 CUs = 66 ms, which is too high to serve uRLLC applications that require lower latency of 10 ms [57]. Meanwhile, 128 information bytes encapsulated into packets with 1000 CUs provide not only ultra-reliability, but also low latency.

Fig. 9(a) and Fig. 9(b) plot the E2E throughput and EE versus $\bar{\gamma}_P$ for different MIMO settings, whereas Fig. 9(c) and Fig. 9(d) display the E2E throughput and EE versus K. In Fig. 9(a), we observe that the E2E throughput significantly improves when $\bar{\gamma}_P$ is increased. Apparently, in the region of high $\bar{\gamma}_P$, the E2E throughput converges to the coding rate, as stated in **Remark 2**. In Fig. 9(b), it can be seen that there exists an optimal value of $\bar{\gamma}_P$ to achieve the highest EE. We also note that EE represents the tradeoff between the system throughput and energy consumption. In the low-SNR region, EE monotonically increases as $\bar{\gamma}_P$ increases. Meanwhile, in the high-SNR region, EE is reduced. This is because at a sufficiently high SNR value, the system throughput achieves a target throughput and remains saturated, whereas $\bar{\gamma}_P$ still increases, leading to EE degradation. Furthermore, it can be further observed that higher numbers of transmit and receive antennas provide better performance in terms of E2E throughput and EE, which is precise to the discussion in **Remark 4**. Fig. 9(c) and Fig. 9(d) reveal that there exists an optimal K to maximize the E2E throughput and EE. Notably, the system throughput and EE approach zero when $K \to \infty$, as indicated in **Remark 3**. We note that choosing the value of K is important in multihop networks to balance the tradeoffs among the implementation cost, transmit power, and system performance (e.g., reliability, latency, throughput, and EE).

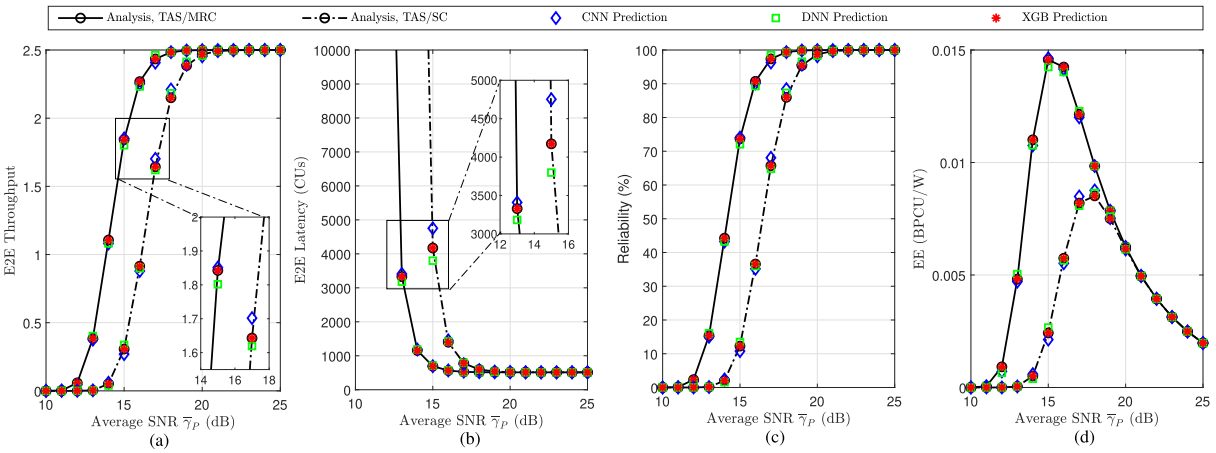


FIGURE 7. ML-based estimation of E2E throughput, latency, EE and reliability versus $\overline{\gamma}_P$.

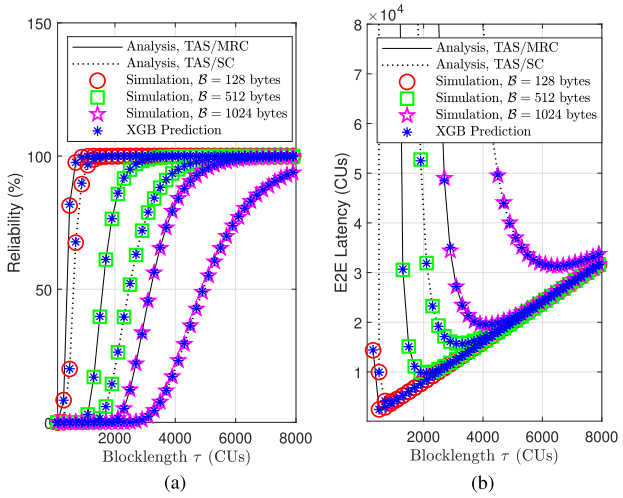


FIGURE 8. The influence of blocklength τ and message length $\mathcal{B}$ on the performance in terms of reliability and E2E latency.

Figs. 8 and 9 also show that the analysis, simulation, and XGB prediction results are in excellent agreement, which confirms our designed ML application and theoretical analysis. Furthermore, the results reveal that the TAS/MRC scheme achieves better performance than TAS/SC in terms of reliability, latency, throughput, and EE.

B. OPTIMAL PERFORMANCE EVALUATION

To provide a general performance comparison between the optimal and nonoptimal strategies, we adopt nonoptimal linear functions for the transmit PA and RL in [23], which are determined, respectively, by

$$\hat{P}_{k-1}(k) = \left\{ (k-1)\mathcal{M} - \frac{K}{2}\mathcal{M} + \frac{P_{\text{tot}}}{K+1} \,\middle|\, \sum_{k=1}^{K+1} \hat{P}_{k-1} = P_{\text{tot}} \right\},$$

$$d_k(k) = \left\{ (k-1)\mathcal{N} - \frac{K}{2}\mathcal{N} + \frac{D}{K+1} \,\middle|\, \sum_{k=1}^{K+1} d_k = D \right\},$$

where $\mathcal{M}$ and $\mathcal{N}$ are the slopes of the linear functions. In such a scenario, the non-optimal case studies can be defined by the slopes only, which are compared to the optimal configurations given in (36), (37), (39), (40), and (42). An illustration of different non-optimal linear configurations for PA and RL is depicted in Figs. 10(a) and 10(b). It is worth noting that the equal PA and RL configurations occur when the uniform configurations are assumed, i.e., $\mathcal{M} = 0$ and $\mathcal{N} = 0$, respectively. For a fair comparison, the parameters $K = 3$, $N = 2$, $N_T = 3$, $N_R = 2$, $\mathcal{B} = 1280$ bits, $\tau = 128$ CUs, $\beta = 0.8$, $D = 1$, $\eta = 3$, and $\bar{\gamma}_I = \bar{\gamma}_{\text{tot}}/(K+1)$ are fixed in this subsection, where $\bar{\gamma}_{\text{tot}} = P_{\text{tot}}/\mathcal{N}_0$.

Fig. 10 compares the performance among various PA and RL configurations for both the TAS/MRC and TAS/SC schemes. To compare the performance for PA strategies in Fig. 10(c) and Fig. 10(d), we fix the relay distance constraint as

$$d_k = \left\{ k \cdot D \Big/ \sum_{i=1}^{K+1} i \,\middle|\, \sum_{k=1}^{K+1} d_k = D \right\}.$$

Otherwise, the fixed PA

$$\hat{P}_{k-1} = \left\{ (K+2-k) \cdot P_{\text{tot}} \Big/ \sum_{i=1}^{K+1} i \,\middle|\, \sum_{k=1}^{K+1} \hat{P}_{k-1} = P_{\text{tot}} \right\}$$

is considered for Fig. 10(e) and Fig. 10(f). In Fig. 10, it is clear that the optimal PA and RL strategies provide significantly lower E2E BLERs than the nonoptimal linear schemes. As a result, the power savings from employing optimal PA/RL configurations are also observed. Specifically, at an E2E BLER of 10^{-5}, the optimal PA strategies in Fig. 10(c) and Fig. 10(d) achieve substantial SNR gains of up to 9.75 dB over the worst nonoptimal scheme (i.e., $\mathcal{M}_5$). Meanwhile, at an E2E BLER of 10^{-5}, the optimal RL strategies in Fig. 10(e) and Fig. 10(f) offer up to 14.25 dB power savings compared with the worst nonoptimal case (i.e., $\mathcal{N}_1$).

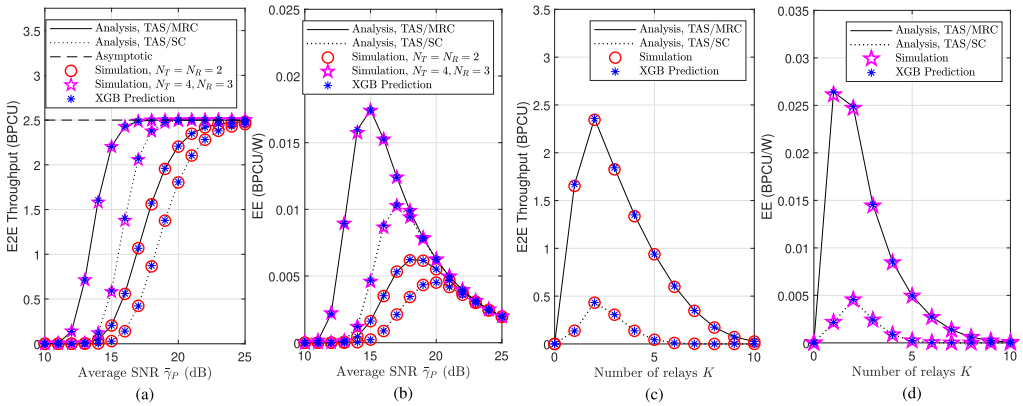


FIGURE 9. E2E throughput and EE versus the average SNR $\bar{\gamma}_P$ and the number of relays K.

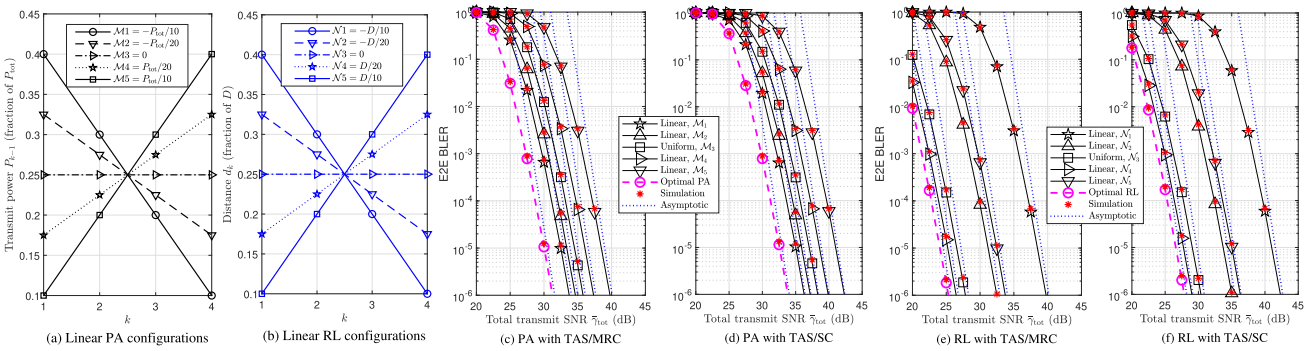


FIGURE 10. The performance of the optimal PA and RL strategies compared to different non-optimal linear configurations.

VIII. CONCLUSION

In this paper, we have investigated MIMO underlay CR multihop relay networks in the context of SPCs, where general and practical scenarios are considered with multiple PUs in the primary network and imperfect CSI of the interference channels. In our proposed framework, SPCs aim to support uRLLCs, underlay CR networks aim to tackle the spectrum-scarcity problem, and multihop relay networks aim to gain benefits from the limited-transmit-power problem. The approximated E2E BLER closed-form expressions of both the TAS/MRC and TAS/SC schemes are derived under the consideration of quasi-static Rayleigh fading channels and the FBL regime, from which their E2E throughput, EE, latency, and reliability are analyzed. We have studied the optimal PA and RL problems, which are two important practical problems of relaying systems. Subsequently, DNN-based, CNN-based, and XGB-based estimators have been designed to predict the system performance in terms of E2E throughput, EE, latency, and reliability, which can assist the real-time configurations for the proposed systems. The results from the simulation, analysis, and ML-aided prediction approaches are consistent, which validates the correctness of our analysis. The asymptotic curves in the contexts of the near-CR regime, the far-CR regime, and the proportional tolerable interference power regime completely converge to the analytical results, which yield insights into the behaviors of the system performance in the high-SNR regime. The strengths of the MIMO implementation, SPC, and multihop relay networks in accordance with uRLLC requirements are confirmed, and the impacts of the average SNR and information length of bits on the system performance are also shown. In addition, the optimal PA and RL solutions offer negligible computational complexity and substantial power savings, which minimize the asymptotic E2E BLER under the system constraints. The employed ML-aided estimators provide both highly accurate predictions and a very short execution time, which demonstrates the possibility of realizing real-time configurations. Among the ML frameworks, the XGB algorithm gains the most significant enhancement in prediction accuracy for performance evaluation and reduction in execution time, which makes XGB an efficient tool to estimate the system performance in real-time performance evaluations.

ACKNOWLEDGMENT

An earlier version of this paper was presented in part at the 14th International Conference on ICT Convergence (ICTC), 2023 [1].

REFERENCES

[1] N. H. Tu and K. Lee, "Performance analysis of MIMO CR multihop relaying with imperfect CSI in short-packet URLLCs," in *Proc. Int. Conf. ICT Converg. (ICTC)*, Jun. 2023.

[2] J. Lee, E. Tejedor, K. Ranta-aho, H. Wang, K.-T. Lee, E. Semaan, E. Mohyeldin, J. Song, C. Bergljung, and S. Jung, "Spectrum for 5G: Global status, challenges, and enabling technologies," *IEEE Commun. Mag.*, vol. 56, no. 3, pp. 12–18, Mar. 2018.

[3] N.-N. Dao, Q.-V. Pham, N. H. Tu, T. T. Thanh, V. N. Q. Bao, D. S. Lakew, and S. Cho, "Survey on aerial radio access networks: Toward a comprehensive 6G access infrastructure," *IEEE Commun. Surveys Tuts.*, vol. 23, no. 2, pp. 1193–1225, 2nd Quart., 2021.

[4] A. Sharmila and P. Dananjayan, "Spectrum sharing techniques in cognitive radio networks—A survey," in *Proc. IEEE Int. Conf. Syst., Comput., Autom. Netw. (ICSCAN)*, Mar. 2019, pp. 1–4.

[5] C. Gao, Y. Shi, Y. T. Hou, and S. Kompella, "On the throughput of MIMO-empowered multihop cognitive radio networks," *IEEE Trans. Mobile Comput.*, vol. 10, no. 11, pp. 1505–1519, Nov. 2011.

[6] Q. Li, J. Qin, and R. Feng, "Optimal relay precoding for spectrum sharing multi-hop MIMO cognitive radio networks," *Electron. Lett.*, vol. 48, no. 24, pp. 1562–1564, Nov. 2012.

[7] S. A. R. Zaidi, M. Ghogho, D. C. McLernon, A. Swami, and S. Govindasamy, "Area spectral efficiency and energy efficiency of multi-hop multi-antenna cognitive underlay networks," in *Proc. Asilomar Conf. Signals, Syst. Comput.*, Nov. 2013, pp. 1330–1335.

[8] S. A. R. Zaidi, M. Ghogho, D. C. McLernon, and A. Swami, "Achievable spatial throughput in multi-antenna cognitive underlay networks with multi-hop relaying," *IEEE J. Sel. Areas Commun.*, vol. 31, no. 8, pp. 1543–1558, Aug. 2013.

[9] F. S. Al-Qahtani, R. M. Radaydeh, S. Hessien, T. Q. Duong, and H. Alnuweiri, "Underlay cognitive multihop MIMO networks with and without receive interference cancellation," *IEEE Trans. Commun.*, vol. 65, no. 4, pp. 1477–1493, Apr. 2017.

[10] W. Xu, W. Yuan, Q. Shi, X. Wang, and Y. Zhang, "Distributed energy-efficient cross-layer optimization for multihop MIMO cognitive radio networks with primary user rate protection," *IEEE Trans. Veh. Technol.*, vol. 66, no. 1, pp. 785–797, Jan. 2017.

[11] P. T. Tin, D.-H. Ha, P. M. Quang, N. T. Binh, and N. L. Nhat, "Performance of multi-hop cognitive MIMO relaying networks with joint constraint of intercept probability and limited interference," *Telkomnika*, vol. 19, no. 1, pp. 44–50, Feb. 2021.

[12] N. Ben Halima and H. Boujemâa, "Energy harvesting with adaptive transmit power for multi-antenna multihop cognitive radio networks," *Sustain. Comput., Informat. Syst.*, vol. 31, Sep. 2021, Art. no. 100567.

[13] J. Navarro-Ortiz, P. Romero-Diaz, S. Sendra, P. Ameigeiras, J. J. Ramos-Munoz, and J. M. Lopez-Soler, "A survey on 5G usage scenarios and traffic models," *IEEE Commun. Surveys Tuts.*, vol. 22, no. 2, pp. 905–929, 2nd Quart., 2020.

[14] K. I. Pedersen, F. Frederiksen, G. Berardinelli, and P. E. Mogensen, "The coverage-latency-capacity dilemma for TDD wide area operation and related 5G solutions," in *Proc. IEEE 83rd Veh. Technol. Conf. (VTC Spring)*, May 2016, pp. 1–5.

[15] M. R. Amini and M. W. Baidas, "Performance analysis of URLL energy-harvesting cognitive-radio IoT networks with short packet and diversity transmissions," *IEEE Access*, vol. 9, pp. 79293–79306, 2021.

[16] N. T. Y. Linh, N. H. Tu, P. N. Son, and V. N. Q. Bao, "Dual-hop relaying networks for short-packet URLLCs: Performance analysis and optimization," *J. Commun. Netw.*, vol. 24, no. 4, pp. 408–418, Aug. 2022.

[17] V. Shahiri, A. Kuhestani, and L. Hanzo, "Short-packet amplify-and-forward relaying for the Internet-of-Things in the face of imperfect channel estimation and hardware impairments," *IEEE Trans. Green Commun. Netw.*, vol. 6, no. 1, pp. 20–36, Mar. 2022.

[18] B. Makki and M.-S. Alouini, "End-to-end performance analysis of delay-sensitive multi-relay networks," *IEEE Commun. Lett.*, vol. 23, no. 12, pp. 2159–2163, Dec. 2019.

[19] A. Ranjha and G. Kaddoum, "Quasi-optimization of distance and block-length in URLLC aided multi-hop UAV relay links," *IEEE Wireless Commun. Lett.*, vol. 9, no. 3, pp. 306–310, Mar. 2020.

[20] Y. Chen, T. Zhang, Y. Zhang, B. Yu, and Y. Cai, "Relay-assisted secure short-packet communications in cognitive Internet of Things," in *Proc. IEEE Int. Conf. Commun. Workshops (ICC Workshops)*, Jun. 2021, pp. 1–6.

[21] D.-D. Tran, S. K. Sharma, S. Chatzinotas, I. Woungang, and B. Ottersten, "Short-packet communications for MIMO NOMA systems over Nakagami-*m* fading: BLER and minimum blocklength analysis," *IEEE Trans. Veh. Technol.*, vol. 70, no. 4, pp. 3583–3598, Apr. 2021.

[22] N. H. Tu and K. Lee, "Ultra-reliable and low-latency short-packet communications for multihop MIMO relaying," 2021, *arXiv:2112.10529*.

[23] N. H. Tu and K. Lee, "Performance analysis and optimization of multihop MIMO relay networks in short-packet communications," *IEEE Trans. Wireless Commun.*, vol. 21, no. 6, pp. 4549–4562, Jun. 2022.

[24] W. Liu, Z. Wang, X. Liu, N. Zeng, Y. Liu, and F. E. Alsaadi, "A survey of deep neural network architectures and their applications," *Neurocomputing*, vol. 234, pp. 11–26, Apr. 2017.

[25] C. D. Ho, T.-V. Nguyen, T. Huynh-The, T.-T. Nguyen, D. B. da Costa, and B. An, "Short-packet communications in wireless-powered cognitive IoT networks: Performance analysis and deep learning evaluation," *IEEE Trans. Veh. Technol.*, vol. 70, no. 3, pp. 2894–2899, Mar. 2021.

[26] T.-V. Nguyen, T. Huynh-The, and B. An, "A deep CNN-based relay selection in EH full-duplex IoT networks with short-packet communications," in *Proc. IEEE Int. Conf. Commun.*, Jun. 2021, pp. 1–6.

[27] T.-V. Nguyen, V.-D. Nguyen, D. B. da Costa, and B. An, "Short-packet communications in multi-hop WPINs: Performance analysis and deep learning design," in *Proc. IEEE Global Commun. Conf. (GLOBECOM)*, Dec. 2021, pp. 1–6.

[28] T.-H. Vu, T.-V. Nguyen, and S. Kim, "Wireless powered cognitive NOMA-based IoT relay networks: Performance analysis and deep learning evaluation," *IEEE Internet Things J.*, vol. 9, no. 5, pp. 3913–3929, Mar. 2022.

[29] T.-H. Vu, T.-V. Nguyen, T.-T. Nguyen, and S. Kim, "Performance analysis and deep learning design of wireless powered cognitive NOMA IoT short-packet communications with imperfect CSI and SIC," *IEEE Internet Things J.*, vol. 9, no. 13, pp. 10464–10479, Jul. 2022.

[30] T.-H. Vu, T.-V. Nguyen, D. B. da Costa, and S. Kim, "Performance analysis and deep learning design of underlay cognitive NOMA-based CDRT networks with imperfect SIC and co-channel interference," *IEEE Trans. Commun.*, vol. 69, no. 12, pp. 8159–8174, Dec. 2021.

[31] A. Goldsmith, *Wireless Communications*. Cambridge, U.K.: Cambridge Univ. Press, 2005.

[32] X. Tang, M.-S. Alouini, and A. J. Goldsmith, "Effect of channel estimation error on M-QAM BER performance in Rayleigh fading," *IEEE Trans. Commun.*, vol. 47, no. 12, pp. 1856–1864, Dec. 1999.

[33] T. V. Nguyen, T.-N. Do, V. N. Q. Bao, D. B. d. Costa, and B. An, "On the performance of multihop cognitive wireless powered D2D communications in WSNs," *IEEE Trans. Veh. Technol.*, vol. 69, no. 3, pp. 2684–2699, Mar. 2020.

[34] H. Van Toan, V. Q. Bao, and H. Nguyen-Le, "Cognitive two-way relay systems with multiple primary receivers: Exact and asymptotic outage formulation," *IET Commun.*, vol. 11, no. 16, pp. 2490–2497, Nov. 2017.

[35] A. Afana, I. A. Mahady, and S. Ikki, "Quadrature spatial modulation in MIMO cognitive radio systems with imperfect channel estimation and limited feedback," *IEEE Trans. Commun.*, vol. 65, no. 3, pp. 981–991, Mar. 2017.

[36] L. Li, S. A. Vorobyov, and A. B. Gershman, "Transmit antenna selection based strategies in MISO communication systems with low-rate channel state feedback," *IEEE Trans. Wireless Commun.*, vol. 8, no. 4, pp. 1660–1666, Apr. 2009.

[37] T. M. Hoang, X. N. Tran, B. C. Nguyen, and L. T. Dung, "On the performance of MIMO full-duplex relaying system with SWIPT under outdated CSI," *IEEE Trans. Veh. Technol.*, vol. 69, no. 12, pp. 15580–15593, Dec. 2020.

[38] Z. Chen, J. Yuan, and B. Vucetic, "Analysis of transmit antenna selection/maximal-ratio combining in Rayleigh fading channels," *IEEE Trans. Veh. Technol.*, vol. 54, no. 4, pp. 1312–1321, Jul. 2005.

[39] I. S. Gradshteyn and I. M. Ryzhik, *Table of Integrals, Series, and Products*. New York, NY, USA: Academic, 2014.

[40] A. F. Coskun and O. Kucur, "Performance of joint transmit and receive antenna selection in Nakagami-*m* fading channels," in *Proc. Eur. Wireless Conf. (EW)*, Apr. 2010, pp. 1–4.

[41] W. Yang, G. Durisi, T. Koch, and Y. Polyanskiy, "Quasi-static multiple-antenna fading channels at finite blocklength," *IEEE Trans. Inf. Theory*, vol. 60, no. 7, pp. 4232–4265, Jul. 2014.

[42] Y. Polyanskiy, H. V. Poor, and S. Verdu, "Channel coding rate in the finite blocklength regime," *IEEE Trans. Inf. Theory*, vol. 56, no. 5, pp. 2307–2359, May 2010.

[43] B. Makki, T. Svensson, and M. Zorzi, "Finite block-length analysis of the incremental redundancy HARQ," *IEEE Wireless Commun. Lett.*, vol. 3, no. 5, pp. 529–532, Oct. 2014.

[44] R. M. McLeod, *The Generalized Riemann Integral*, vol. 20. Providence, RI, USA: American Mathematical Society, 1980.

[45] C. K. Chui, "Concerning Gaussian–Chebyshev quadrature errors," *SIAM J. Numer. Anal.*, vol. 9, no. 2, pp. 237–240, Jun. 1972.
[46] M. Khafagy, A. Ismail, M.-S. Alouini, and S. Aissa, "On the outage performance of full-duplex selective decode-and-forward relaying," *IEEE Commun. Lett.*, vol. 17, no. 6, pp. 1180–1183, Jun. 2013.
[47] R. Devassy, G. Durisi, P. Popovski, and E. G. Ström, "Finite-blocklength analysis of the ARQ-protocol throughput over the Gaussian collision channel," in *Proc. 6th Int. Symp. Commun., Control Signal Process. (ISCCSP)*, May 2014, pp. 173–177.
[48] O. L. Alcaraz López, R. D. Souza, H. Alves, and E. M. G. Fernández, "Ultra reliable short message relaying with wireless power transfer," in *Proc. IEEE Int. Conf. Commun. (ICC)*, May 2017, pp. 1–6.
[49] J. Lee, H. Wang, J. G. Andrews, and D. Hong, "Outage probability of cognitive relay networks with interference constraints," *IEEE Trans. Wireless Commun.*, vol. 10, no. 2, pp. 390–395, Feb. 2011.
[50] C. Zhong, T. Ratnarajah, and K.-K. Wong, "Outage analysis of decode-and-forward cognitive dual-hop systems with the interference constraint in Nakagami-*m* fading channels," *IEEE Trans. Veh. Technol.*, vol. 60, no. 6, pp. 2875–2879, Jul. 2011.
[51] I. Bahceci, T. M. Duman, and Y. Altunbasak, "Antenna selection for multiple-antenna transmission systems: Performance analysis and code construction," *IEEE Trans. Inf. Theory*, vol. 49, no. 10, pp. 2669–2681, Oct. 2003.
[52] D. Gunduz, M. A. Khojastepour, A. Goldsmith, and H. V. Poor, "Multi-hop MIMO relay networks: Diversity-multiplexing trade-off analysis," *IEEE Trans. Wireless Commun.*, vol. 9, no. 5, pp. 1738–1747, May 2010.
[53] R. T. Rockafellar, "Lagrange multipliers and optimality," *SIAM Rev.*, vol. 35, no. 2, pp. 183–238, Jun. 1993.
[54] G. Klambauer, T. Unterthiner, A. Mayr, and S. Hochreiter, "Self-normalizing neural networks," in *Proc. NeurIPS*, vol. 30, Dec. 2017, pp. 971–980.
[55] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in *Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, Aug. 2016, pp. 785–794.
[56] D. T. Hoang and K. Lee, "Deep learning-aided coherent direction-of-arrival estimation with the FTMR algorithm," *IEEE Trans. Signal Process.*, vol. 70, pp. 1118–1130, 2022.
[57] O. L. A. López, H. Alves, R. D. Souza, and E. M. G. Fernández, "Ultra-reliable short-packet communications with wireless energy transfer," *IEEE Signal Process. Lett.*, vol. 24, no. 4, pp. 387–391, Apr. 2017.

NGO HOANG TU (Graduate Student Member, IEEE) received the B.S. degree in computer networking and data communications from the Ho Chi Minh City University of Transport (UT-HCMC), Vietnam, in 2020, and the M.S. degree from the Department of Smart Energy System Engineering, Seoul National University of Science and Technology (SeoulTech), South Korea, in 2022. He is currently pursuing the Ph.D. degree with the Department of Electrical and Information Engineering, SeoulTech. From January 2019 to January 2020, he was an Assistant Researcher with the Wireless Communication Laboratory, Posts and Telecommunications Institute of Technology (PTIT), Vietnam. From February 2020 to August 2020, he was a Lecturer with the Department of Computer Engineering, UT-HCMC. His research interests include wireless communications, signal processing, intelligent reflecting surface, 6G infrastructures, and applied machine learning. He serves as an Editor for the *EAI Endorsed Transactions on Industrial Networks and Intelligent Systems*.

TRONG-DAI HOANG (Graduate Student Member, IEEE) received the B.S. degree (Hons.) in control engineering and automation from the Ho Chi Minh City University of Technology, Ho Chi Minh City, Vietnam, in 2019, and the M.S. degree in electrical and information engineering from the Seoul National University of Science and Technology (SeoulTech), Seoul, South Korea, in 2021. He is currently pursuing the Ph.D. degree with the School of Electrical and Data Engineering, University of Technology Sydney, Australia. From September 2021 to July 2022, he was a Research Scientist with the Research Center for Electrical and Information Technology, SeoulTech. His research interests include applied machine learning, wireless communication, signal processing, and optimization.

KYUNGCHUN LEE (Senior Member, IEEE) received the B.S., M.S., and Ph.D. degrees in electrical engineering from the Korea Advanced Institute of Science and Technology (KAIST), Daejeon, in 2000, 2002, and 2007, respectively. From April 2007 to June 2008, he was a Postdoctoral Researcher with the University of Southampton, U.K. From July 2008 to August 2010, he was with Samsung Electronics, Suwon, South Korea. Since September 2010, he has been with the Seoul National University of Science and Technology, South Korea. In 2017, he was a Visiting Assistant Professor with North Carolina State University, Raleigh, NC, USA. His research interests include wireless communications and applied machine learning. He received the Best Paper Awards from the IEEE International Conference on Communications (ICC), in 2009, and the IEEE Wireless Communications and Networking Conference (WCNC), in 2020.