WILEY | Hindawi

Review Article

A Survey on Transmission Schemes on Large-Scale Internet of Things with Nonorthogonal Multiple Access

Wenyu Zhou,[1] Rui Zhao [ID],[2] Fusheng Zhu [ID],[3] Lijia Lai [ID],[2] and Xutao Li[4]

[1]*School of Management, Hefei University of Technology, Hefei, Anhui, China*
[2]*School of Computer Science, Guangzhou University, Guangzhou, China*
[3]*Guangdong New Generation Communication and Network Innovative Institute (GDCNi), Guangzhou, China*
[4]*Department of Electronic Engineering, Shantou University, Shantou, Guangdong, China*

Correspondence should be addressed to Rui Zhao; 2111806073@e.gzhu.edu.cn and Fusheng Zhu; zhufusheng@gdcni.cn

This paper performs a comprehensive survey on transmission schemes for the large-scale Internet of things (IoT) networks with nonorthogonal multiple access (NOMA). By solving the interference among users, NOMA can significantly improve the frequency reuse efficiency and support multiple users to use the same frequency resources. It is considered to be one of the most effective technologies for the next-generation wireless communication. However, there are still many challenges on the transmission schemes for the large-scale NOMA system, including the short-data packet transmission, active user detection, channel estimation, and data detection. In order to meet these challenges, this paper first reviews the short-packet transmission in the large-scale NOMA systems and then reviews the active user detection and channel estimation technologies of the considered systems.

1. Introduction

The application of Internet of things (IoT) has promoted a significant increase in data traffic of wireless networks [1–3]. For machine-type communication (MTC) or mission-critical communication (autonomous driving, drones, and augmented/virtual reality), all of them have proposed new development requirements on capacity, latency, reliability, and scalability. In particular, the massive machine-type communication (mMTC) and mass access are considered as two very important scenarios of the fifth-generation (5G) and beyond 5G (B5G) wireless communications [4–8]. As a typical application of mMTC, large-scale IoT-enabled cellular networks are widely used in intelligent buildings, intelligent cities, intelligent medical treatment, factory automation, automatic driving, intelligent meters, and other fields. It has been recognized that large-scale IoT can significantly reduce the system operation cost [9, 10].

As shown in Figure 1, only a short number of unknown users are active at any given time in the large-scale IoT model. This communication mode is mainly designed with the purpose of energy saving. Generally, developing future cellular IoT faces many challenges, such as adapting to more users with low latency requirements, dynamically identifying active users, and reliably receiving their information [11]. For the traditional authorization-based random access scheme, uncoordinated users can compete for physical layer resource blocks for data transmission. Each active user randomly selects one from a set of predefined orthogonal preamble sequences to notify the base station (BS) that the user is active [12, 13]. As to each activation preamble, the BS responds to the corresponding user as permission for the subsequent transmission. Each user who receives a response to its preamble transmission sends a connection request for the required resources of the subsequent data transmission [14–16]. When only one user in the system
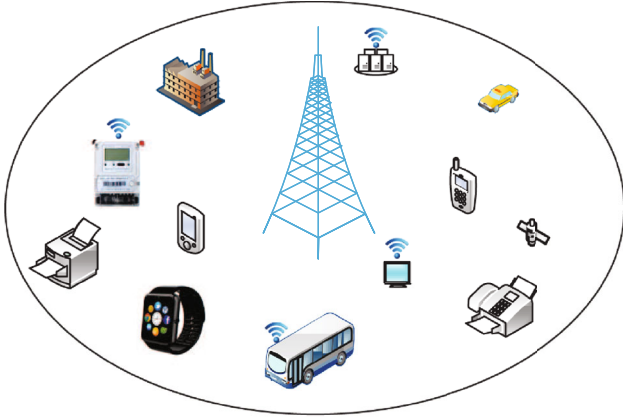
📶 Denoting active users, and others are idle users.

FIGURE 1: System model of large-scale IoT sporadic communications.

selects the current preamble sequence, the BS authorizes the user's connection request and sends a contention resolution message to notify the reserved resource user to send suspended data. If two or more users select the same preamble sequence, the connection request will send conflicts [17–19].

When the BS detects a conflict, it will not reply to the contention resolution message. The affected user restarts the random access process after the timer expires. For the authorized access scheme, the user needs to handshake four times with the BS, which increases the control signaling overhead and latency. On the contrary, users can start data transmission in the unauthorized protocol without shaking hands with the BS, which can meet the low latency requirements of large-scale IoT. In addition, the nonorthogonal multiple access (NOMA) technology allows multiple users to share the same resource block (including time and spectrum resources) to establish more connections.

At present, there are still some challenges, such as the nonorthogonal transmission scheme for short data packets (or short packets), large-scale pilot sequences, active user detection, channel estimation, and multiple optimal design and analysis of user data detection. Recently, it has been recognized that a reliable nonorthogonal transmission scheme and effective multiuser detection can improve the spectrum efficiency of the system. Therefore, the license-free NOMA can effectively combine the low access latency of the license-free protocol with the advantages of large numbers of connections of NOMA and the high spectrum efficiency, which can meet the communication requirements of the large-scale IoT networks that realize the low latency and reliable communication. The cochannel interference of NOMA, especially in the IoT networks with large-scale multiple access, dramatically limits the system performance. Hence, designing an appropriate license-free NOMA scheme is very critical.

Due to the requirements of low transmission latency, the data packet length and pilot sequence of IoT users are limited, the scale of users is enormous, and it is challenging to meet the needs of all users for the orthogonal spectrum

resources and pilot sequence. Therefore, nonorthogonal transmission schemes and nonorthogonal pilot sequences, such as Gaussian and Reed Muller (RM) sequences, have been widely studied. In addition, signal detection plays a vital role in the design of a license-free NOMA system. This is because with no coordination, multiple users can transmit information to BS in parallel and BS needs to detect these signals efficiently. In particular, BS needs to detect signals from some unknown users among a large number of potential users. The signal detection problem of large-scale IoT networks involves three essential parts: identifying active users, channel estimation, and multiuser data detection [20, 21]. The signal detection of large-scale IoT can be divided into two successive stages: channel training and data transmission, in which active user detection, channel estimation, and data decoding will be performed. The paper implementation plan includes the following steps: (1) performing the channel estimation and then performing the joint detection of the data of active users and multiuser and (2) performing the channel estimation and active users jointly and then detecting the multiuser data. For the single-phase transmission scheme, data information is embedded in a pilot sequence or pilot symbols are embedded in a data packet and the BS can jointly detect active users and their data in a phase. This scheme does not require particular channel estimation and can reduce the system overhead and transmission latency. However, the signal transmission and reception scheme need to be carefully designed.

Due to the broadcast characteristics of wireless transmission, ten large-scale IoT networks are vulnerable to interference and attack by illegal users, so physical layer security is very important. In [22, 23], the authors analyze the physical layer security of the cooperative NOMA system and a two-stage relay scheme is proposed. The closed expression of outage probability and the asymptotic expression of outage probability under signal-to-noise ratios (SNRs) are derived. In [24], the authors proposes a security-enhanced user pairing scheme under weak security constraints to achieve high-frequency spectral efficiency and weak system confidentiality. In [25], the authors study the outage probability and secret outage probability in two-user NOMA systems. Based on the concept of NOMA, the interruption behavior of trusted users under the constraint of security outage probability is studied.

Most researches on IoT data transmission are based on the Shannon theory [26–29]. In practice, the data packets of IoT users are tiny and different users have different business requirements. In addition, signal detection mainly uses compressed sensing (CS) or linear inversion technology, which ignores the changing characteristics of sparse signals [30–33]. Therefore, researches based on the large-scale IoT transmission characteristics and different business requirements are not systematic and complete. This paper studies the nonorthogonal transmission and signal detection of large-scale IoT from the following three aspects: (1) from the perspective of short-data packet transmission, we design a reliable nonorthogonal transmission scheme for the cellular IoT uplink and downlink and analyze the system error performance and its influencing factors, (2) we study the

characteristics of fast fading of nonzero elements of sparse signals and design efficient and reliable active user detection algorithms to reduce the false alarm rate of IoT user activity detection, thereby improving the utilization of spectrum resources, and (3) from the structural characteristics of double signal sparsity, we use Bayesian inference and optimization theory to design estimation algorithms for the user activity, channel jointly, and data, reducing the pilot overhead for reliable information transmission. However, for the large-scale NOMA, a series of topics such as the design of nonorthogonal transmission scheme with a short data packet, reliable detection algorithm with low complexity, and the reduction of system signaling overhead and transmission latency are very challenging. The problems mentioned above have essential theory and application value for the planning and deploying the large-scale multiple access technology in IoT networks.

2. Recent Progress and Challenges

It is essential to design the large-scale IoT signal transmission and detection solutions to meet a series of requirements such as large-scale connections, reliable transmission of short data packets, high spectrum efficiency, and low signal overhead. In particular, the reliable transmission design and performance analysis of short data packets and the design of large-scale signal detection algorithms are the keys to IoT researches. To fully understand the existing problems and limitations of the IoT research, we will analyze the relevant research status and development of the trend in the following.

2.1. Short-Data Packet Transmission in Large-Scale Multiple Access. In the third-generation partnership project (3GPP) IoT standard, the uplink of the narrow-band IoT (NB-IoT) uses single-carrier frequency division multiple access (SC-FDMA) and the downlink uses orthogonal frequency division multiple access (OFDMA). Although the transmission and reception schemes of IoT can be simplified, the spectrum efficiency of the system is low. Therefore, it is necessary to design new multiple-access schemes to improve IoT network spectrum efficiency further. In fact, the performance of the large-scale IoT networks is determined by the channel estimation, uplink data transmission, and downlink data transmission. At present, some researchers have optimized designs for a system that couples two or three parts of channel estimation, uplink data transmission, and downlink data transmission. In the literature, the researchers considered the channel estimation and uplink data transmission of a single-cell uplink in the large-scale multiantenna system, jointly optimized the pilot length and transmission power, and then obtained an optimized design with an optimal pilot length equal to the user number [34]. In addition, for the three-stage transmission, a cellular IoT transmission protocol can be designed for large-scale access by optimizing the transmission time of three stages to maximize the rate and speed. They proposed a reliable transmission scheme based on NOMA, analyzed the relationship between packet rate and interruption of the performance, and optimized the

transmission power of the system [35]. Specifically, the current optimized design for IoT uplink and downlink data transmission is based on the Shannon theory which is given by [36],

$$C = \log_2(1 + \gamma), \tag{1}$$

where γ represents the received SNR. Although the Shannon theory defines the theoretical upper bound of the communication rate without transmission error, it theoretically requires an infinite amount of code to implement. In practice, the amount of data transmitted by IoT users is short and the short-packet transmission requires a low transmission latency. Therefore, the Shannon capacity cannot accurately characterize the reliability and latency of the large-scale IoT short-data packet transmission.

To accurately describe the channel capacity under the finite code length, the finite coding system can be designed for the additive white Gaussian noise (AWGN) channel. The approximate relationship between the achievable rate r, the code length L, and the packet error rate ε can be derived as

$$r \approx \log_2(1 + \gamma) - \sqrt{\frac{1}{L}\left(1 - \frac{1}{(1+\gamma)^2}\right)}\frac{Q^{-1}(\varepsilon)}{\ln 2}, \tag{2}$$

where $Q^{-1}(\bullet)$ is the inverse function of function $Q(x) = \int_x^{+\infty} 1/\sqrt{2\pi}e^{t^2/2}dt$. Although equation (2) is an approximation of the upper bound of the achievable rate under a finite code length, most of existing researches have verified that equation (2) is very accurate.

For the research on the nonorthogonal transmission system of short packets of large-scale IoT, there is still no complete theoretical framework and the design and performance analysis of the large-scale IoT short-data packet nonorthogonal transmission scheme based on the finite coding has not been studied. The influence of channel estimation error propagation, data encoding length, and the number of message bits on system performance are also to be further studied.

2.2. Active User Detection and Channel Estimation for Large-Scale Multiple Access. In order to estimate the active users and their channel information, a practical approach is to use the "sporadic" nature of the large-scale IoT communication to model the received signal in the training phase as a sparse signal recovery problem and then solve the problem based on the CS technology or a Bayesian inference framework. In [37], Du et al. considered a joint design scheme for the active user detection and channel estimation for a single antenna system, where the received pilot signal is portrayed as a single measurement vector problem. They proposed a joint iterative algorithm based on the message passing and block sparse Bayesian learning to design high-precision and low-complexity algorithms to solve the sparse signal recovery problem. In addition, for a system with analog-to-digital converters (ADC), He et al. [38] proposed an active user detection and channel estimation based on a

generalized expectation-consistent signal recovery algorithm, which allows the output mapping of the signal to be of the arbitrary form. The above findings show that the single antenna system can serve thousands of IoT users and requires long lead sequences. Therefore, the large-scale antenna system is considered to improve system connectivity. The joint active user detection and channel estimation problem for the large-scale receive antenna system is described as a multiple measurement vector problem [39]. The authors use the approximate message passing (AMP) algorithm with multiple measurement vectors to solve the problem and use the approximate message passing algorithm based on state evolution to analyze the limiting performance of the system.

However, most of the existing algorithms for active user detection are based on the derivative frequency matrix, which satisfies the finite isometric property or has low intercorrelation. In addition, the active user detection algorithm is based on the prior distribution of the large-scale fading of the user or the exact large-scale fading information. In practice, the AMP and the CS-based active user detection algorithms are not suitable for highly correlated derivative matrices and the oversampled discrete cosine transform matrix has very high intercorrelation under individual parameters. In contrast, the AMP algorithm performs poorly in the same setting. In addition, during the active user detection and channel estimation phase, it is not practical to assume that the statistical characteristics of the channel are known to the receiver.

Based on the in-depth analysis of the sparse structure of the received signal in a large-scale multiple access system, the design of active user detection and channel estimation scheme is of great significance for further improving the accuracy of active user detection. However, there is little research on this aspect.

2.3. Data Detection Algorithms for Large-Scale Multi Access. In the large-scale IoT network based on unauthorized NOMA, multiuser data detection plays a vital role, as multiple users can transmit data simultaneously without coordination [40]. Although only a tiny fraction of users is active, the potential user base is large. Hence, the BS needs to obtain a priori information of the data detection of multiple users with the unknown active user. Moreover, two factors need to be considered when designing a data detection scheme for the large-scale multiple access. On the one hand, the impact of the active user detection and the accuracy of the channel estimation on the data detection give a large number of potential users. On the other hand, the computational complexity of the algorithm is reduced while ensuring the accuracy of the algorithm.

In the large-scale IoT networks, the multiuser data detection involves two key components, the active user detection and the channel estimation. In [41], the authors jointly designed the algorithm which can detect the low-complexity active user and data under the assumption that the complete channel information is known [42]. To characterize the impact of incomplete channel information on the multiuser data detection, Liu et al. [43] first jointly designed

the active user detection and channel estimation algorithm, based on the estimated active user and channel information, using the traditional minimum mean square error (MMSE) and least squares (LS) for data detection. However, for the two design schemes mentioned above, the accuracy of active user detection and channel estimation affects the performance of multiuser data detection to a great extent. In addition, when the number of active users is large, the computational complexity that MMSE and LS detection required is high because they are designed to find the inverse of the matrix. Moreover, the existing works focus on the packet synchronization transmission system that users can only change their activity status at the beginning of each coherent time. During the packet transmission, the packet alignment among a large number of users will result in significant overhead and the user who fails to align their packets can disrupt the entire multiuser data detection process.

Currently, in order to further reduce the system overhead and detection error rate and improve the spectral efficiency of the system, Ding et al. [44] concentrated the three parts of active user detection, channel estimation, and data detection in a phase and used blind detection and semiblind detection algorithms to solve the multiuser data detection problem. However, the dual sparsity of the signal and channel has not been fully exploited in this algorithm. Moreover, it will be more challenging as it can be seen from the three-part joint design scheme that the BS needs only a phase-based data signal to identify the active users, estimate the channel, and decode the user data. In this sense, symbolic synchronous transmission is more suitable for the unlicensed multiple access system to minimize the user coordination cost and improve reliability. However, depending on the structure of the symbol synchronous system, it is more challenging to design for the active user identification, channel estimation, and data decoding using the dual-block sparsity performance.

In summary, the current researches on the multiaccess scheme for the large-scale IoT networks mainly focus on active user detection and channel estimation. Although many effective mechanisms and algorithms have been proposed, the current researches rely more on traditional signal detection methods for multiuser data detection. In particular, in the large-scale NOMA system with two-phase or three-phase signal transmission, although there have been many results on the optimal design of active user detection, channel estimation, and data transmission, a complete theory and technology system has not yet been formed and there are still many key issues to be solved:

(1) Firstly, for the nonorthogonal transmission scheme of short packets for the large-scale IoT networks, most of the existing researches are based on the Shannon theory for optimal design. In practice, the quality of service requirements of large-scale IoT users is different. In the smart home scenario, most of the data packets received by the user equipment are very short, which requires low latency and high-reliability transmission. For the fax service

scenario, it requires a lot of data transmission. Therefore, the design of a signal transmission and reception scheme that can meet the needs of large-scale IoT services and realize short packets, low latency, and reliable transmission is worthy of further research

(2) Secondly, for the NOMA-based large-scale IoT research, most of the existing signal processing schemes mostly focus on active user detection and channel estimation. Regarding the false-positive rate and false-negative rate of active users, the relevant research only finds a compromise between them and does not consider handling the false-positive rate and false-negative rate separately based on the characteristics of active user signals. Meanwhile, the traditional MMSE detection is more in consideration for multiuser data detection. However, this detection scheme is only applicable to the case of a short number of active users. If the number of the potential user base and active users is still large, the detection requires a low-complexity algorithm. This prompts us to study a more reliable and more widely applicable detection scheme, which considers signal characteristics, user's prior information, and each user's guide frequency sequence to propose effective active users and data detection algorithms

(3) Finally, for the data joint detection scheme and active users of the large-scale IoT networks, the related literature only uses the CS or linear model inversion technology to configure the system without considering the specific signal transmission, reception scheme, and the structure of the signal processing problem. The relevant algorithm or mechanism has much room for improvement. Therefore, how to study the signal detection model of the large-scale multiple access based on the characteristics of the received signal and reduce the system detection error rate, the transmission latency, and the system overhead also needs further investigation

3. Research Plan

3.1. Research Plan to Be Adopted. This paper studies how to design an effective nonorthogonal transmission and reception scheme under the condition of limited spectrum and energy resources. The objective is to realize low latency and reliable transmission in large-scale IoT networks, improve the system user access rate and spectrum efficiency, and save the system cost. Drawing on the research progress and thought methods in related fields such as NOMA, large-scale IoT, and signal processing and combining the characteristics of large-scale multiple access and IoT, the technical solution shown in Figure 2 has been proposed and the performance of the adopted solution and its impact factors are comprehensively analyzed and simulated to verify the system platform. In the following, the specific technical lines and the associated research schemes will be

presented for each research component. In particular, the problem description, system modeling, and the research plans will be provided.

3.2. Nonorthogonal Transmission and Reception Scheme of Large-Scale IoT Short Data Packets Based on Finite Coding. In large-scale IoT networks, especially the cellular IoT, the system performance is mainly determined by the channel estimation and uplink and downlink data transmission. Based on the traditional time division duplex (TDD) mode of cellular IoT, the uplink and downlink of the system are transmitted on different time slots of the same frequency resource. The system only needs to estimate the uplink channel information to take advantage of the channel reciprocity in the coherent time. However, since the BS and IoT users have different processing capabilities and higher requirements for the implementation of the NOMA scheme, a hybrid multiple-access scheme is needed for the large-scale IoT uplink and downlink communication modes and characteristics. The specific technical solution is shown in Figure 3.

A cellular IoT model, as shown in Figure 4, includes N short-packet transmission users and a central BS where all user terminals are equipped with a single antenna and a low-resolution ADC.

The BS preassigns a dedicated nonorthogonal spreading sequence $S_n \in \mathbb{C}^{L_p \times 1}$ to each user, where the length of the sequence L_p is shorter than the total number of users N and each element of the sequence is independently and equally probabilistically selected from the set $\{-1, 1\}$. In the uplink, an authorization-free NOMA scheme is used, i.e., each user can access the network and transmit data without being authorized. Therefore, the BS needs to estimate the active users and their channel information before detecting the uplink data. In addition, due to the high computational power of the BS, algorithms with higher accuracy can be used to jointly detect multiuser information. However, users in large-scale IoT networks have very limited computational power. In the downlink nonorthogonal transmission, if the information of all active users is superimposed and broadcast, the performance of each user will decrease and the detection complexity will increase. Therefore, it is necessary to consider the active user pairing and grouping strategies and use a nonorthogonal transmission scheme within a group of users and orthogonal resources between groups, such as subcarriers. Moreover, in the nonorthogonal transmission scheme for downlink packet data, the obtainable capacity can be expressed as a function of code length and error block rate using the finite code length capacity. For user n, the signals of other users are usually considered as interference. Thus, in the group with N users, the signal-to-interference-plus-noise ratio (SINR) of the nth user when decoding its own signal is

$$\gamma_{n \longrightarrow n} = \frac{\alpha_n |h\wedge_n|^2}{|h\wedge_n|^2 \sum_{k=1,2,\cdots,N_i, k \neq n} \alpha_k + \sigma_\varepsilon^2 + 1/\rho}, \qquad (3)$$

where $\alpha_k = p_k/P$, $P = \sum_{k=1,2,\cdots,N_i} pk$, and $\rho = P/\sigma_d^2$ denote the
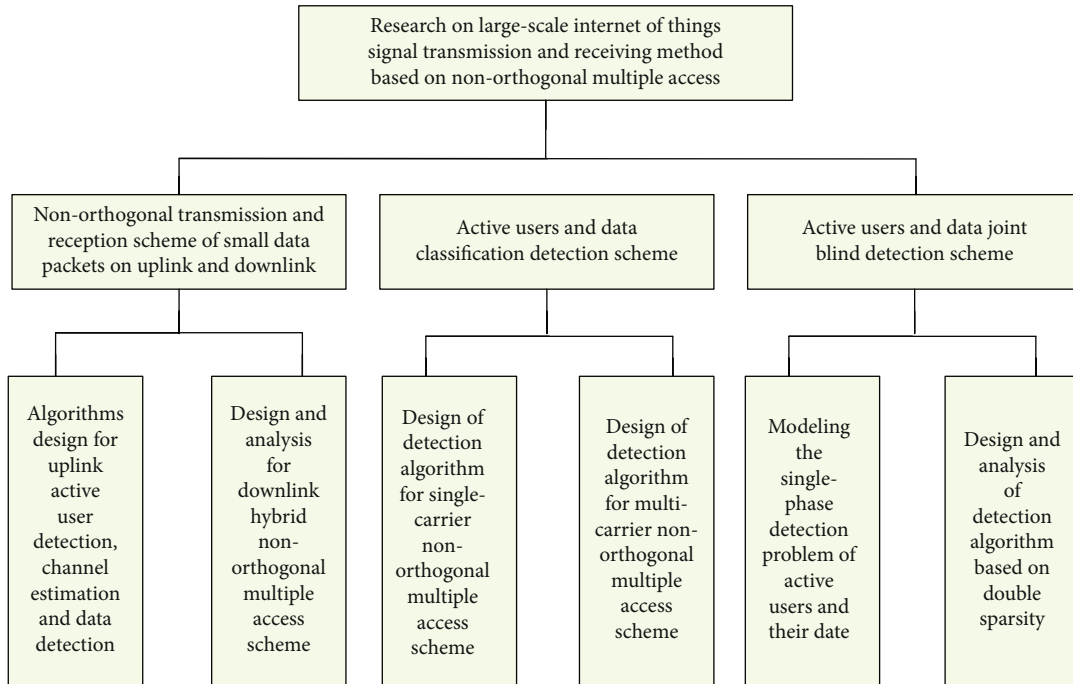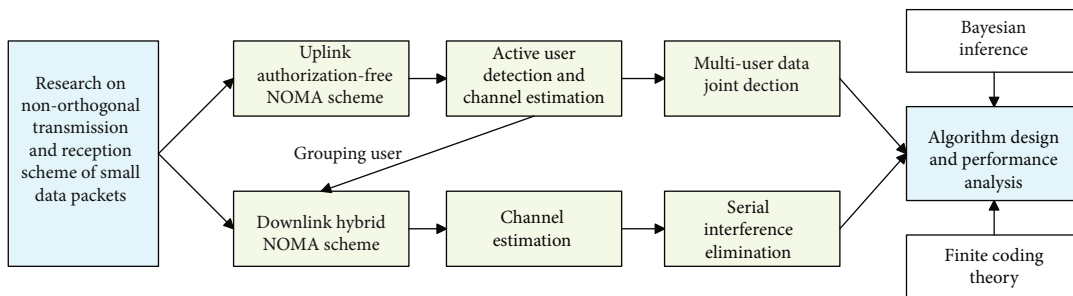
FIGURE 2: Proposed research framework.

FIGURE 3: Technical solution of nonorthogonal transmission and reception scheme for large-scale IoT short data packets.
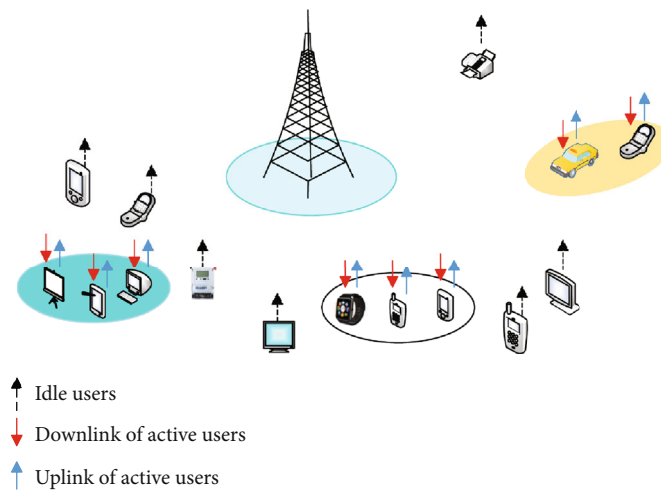

FIGURE 4: Nonorthogonal transmission model for large-scale IoT uplink and downlink in cellular networks.

transmission signal-to-noise ratio (SNR), $\hat{h}_n$ is the estimated channel, and σ_ε^2 is the variance of the channel estimation error. Accordingly, the instantaneous error block rate of user n is approximated by

$$\varepsilon_n \approx Q\left(\frac{c(\gamma_{n \longrightarrow n}) - (B_n/L^d)}{\sqrt{V(\gamma_{n \longrightarrow n})/L^d}}\right), \qquad (4)$$

where $c(\gamma_{n \longrightarrow n}) = \log_2(1 + \gamma_{n \longrightarrow n})$ is the Shannon capacity and $V(\gamma_{n \longrightarrow n}) = (1 - (1/(1 + \gamma_{n \longrightarrow n})^2))(\log_2 e)^2$ denotes the channel dispersion. When the probability density function of $\gamma_{n \longrightarrow n}$ is $f_{\gamma_{n \longrightarrow n}}(x)$, the average error block rate of user n can be written as

$$\bar{\varepsilon}_n \approx \int_0^\infty Q\left(\frac{c(\gamma_{n \longrightarrow n}) - (B_n/L_n)}{\sqrt{V(\gamma_{n \longrightarrow n})/L_n}}\right) f_{\gamma_{n \longrightarrow n}}(x)dx. \qquad (5)$$

In the uplink pilot frequency and data transmission phase, since the receiver with low-resolution ADC is considered, the received signal is quantized, which cannot provide good statistical information. At present, most of the approximate information transmission algorithms are not suitable for the model's active user detection, channel estimation, and data detection. The algorithms for this part need to be redesigned. In addition, for this model, the impact of analog-to-digital conversion bits on the system performance needs to be studied. On the other hand, in the downlink nonorthogonal transmission, the downlink channel needs to be estimated before analyzing the error block rate performance of the system. Meanwhile, the performance of the user grouping strategy for the nonorthogonal transmission should be discussed and analyzed to obtain a more explicit user grouping criterion, which will be the key issue to be investigated. Firstly, we will model the short-packet nonorthogonal transmission system for the cellular IoT uplink and downlink; design the uplink active user, channel estimation, and multiuser detection algorithms; and analyze the performance of uplink based on the short packet transmission and its impact factors. For the downlink, the downlink channel estimation method should be proposed to derive the error block rate performance of the short-packet nonorthogonal transmission and the impact of channel estimation error, information bits, coding length, and other factors on the performance of short-packet nonorthogonal transmission in the downlink is analyzed. Meanwhile, using the power domain NOMA theory, we analyze the short-packet nonorthogonal transmission performance under different cellular IoT transmission service requirements and discuss the variation of system performance under different user groupings. This guides the design of short-packet nonorthogonal schemes for the cellular IoT.

3.3. Large-Scale IoT's Active Users and Data Grading Detection Scheme Based on Reliable Information Transmission. In the signal coherent detection of large-scale IoT uplink, i.e., license-free NOMA, active users, and

data detection are the two key parts, in particular, data detection can be performed through estimation of active users and channels first or active users and data detection can be combined. In the coherent detection, if a joint detection scheme of active users and data is used, the channel information needs to be estimated in advance. However, in the case of unknown active users, the channel of active users should be estimated, which increases the complexity of the detection algorithm and leads to a degradation in the performance of active users and data detection. Therefore, for active user and hierarchical coherent detection schemes, the characteristics of sparse signals should be exploited to obtain the reliable iterative first, and then, the information iterative steps can be optimized and higher detection accuracy can be obtained at the cost of more computational complexity. The specific technical solution is shown in Figure 5.

Assuming that the summarized users are synchronized in a frame structure and are active or dormant for the entire frame, a single-cell IoT uplink license-free single/multicarrier NOMA system is considered. Moreover, a block fading channel is considered, i.e., the channel coefficients remain unchanged within a frame. Since only a short part of users is active in the coherent time, without loss of generality, it is assumed that the BS knows the distribution of each user activity indicator α_n or the total number of active users, where $K(K \ll N)$, α_n obeys Bernoulli distribution, i.e., $\Pr(\alpha_n = 1) = \epsilon$ and $\Pr(\alpha_n = 0) = 1 - \epsilon$, for a certain frame $K/N \approx \epsilon$. In a certain frame, the active user sends a pilot symbol followed by L_d data symbols, while the inactive user stays dormant throughout the frame.

In the training phase, at the BS, the pilot measurement of the lth $(l = 1, 2, \cdots, L)$ subcarrier is

$$\mathbf{y}_l^p = \sum_{n=1}^N \zeta_p \mathbf{s}_n \alpha_n h_{nl} + \mathbf{z}_l^p = \mathbf{S}(\text{diag}(\mathbf{a})\mathbf{h}_l) + \mathbf{z}_l^p, \qquad (6)$$

where ζ_p is the energy normalization factor of the pilot in the spread sequence, $\mathbf{S} = \zeta p[\mathbf{s}_1, \mathbf{s}_2, \cdots, \mathbf{s}_n]$ is the pilot observation matrix, and $\mathbf{a} = (\alpha_1, \alpha_2, \cdots, \alpha_n)T \in \{0, 1\}^N$ is the user activity indicator vector. Notation $\mathbf{h} = (h_{1l}, h_{2l}, \cdots, h_{nl})T$ is the channel vector, where $h_{nl} \sim \mathscr{CN}(0, \sigma_h^2)$, and $\mathbf{z}_l^p \sim CN(0, \sigma_p^2 \mathbf{I}_{L_p})$ represents the additive white Gaussian noise vector.

In the uplink transmission phase, at the BS, the jth data symbol received by the lth subcarrier is

$$y_{lj}^d = \sum_{n=1}^N x_{n,j} \alpha_n h_{nl} + z_{lj}^d. \qquad (7)$$

Thus, the signals received by all subcarriers can be expressed as

$$\mathbf{y}_j^d = \mathbf{H}(\mathbf{a})\mathbf{x}_j + \mathbf{z}_j^d, \qquad (8)$$

where $y_j^d \triangleq (y_{1j}^d, y_{2j}^d, \cdots, y_{L_j}^d)^T$, $\mathbf{H} \triangleq [\mathbf{h}_1, \mathbf{h}_2, \cdots, \mathbf{h}_L]^T$, and $\mathbf{x}_i = (x_{1,j}, x_{2,j}, \cdots, x_{N,j})^T$. The license-free single/multicarrier NOMA system can provide a large number of links and
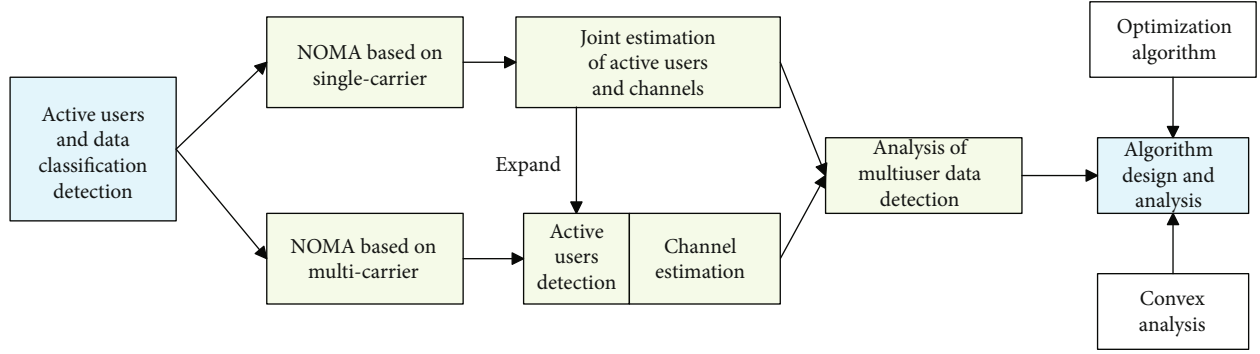
FIGURE 5: Technical solution of the activity user and data classification detection solution in the large-scale IoT.

low access latency and has a wide range of applications in the cellular IoT. However, due to a large number of potential users and that these users use nonorthogonal pilot sequences, the key challenges existing in the unlicensed connection scenarios for large-scale IoT users are channel estimation, active user detection, and data decoding. A fundamental question is, in addition to active user detection, how to design a receiver to perform the channel estimation and/or multiuser data detection effectively.

We intend to solve how to exploit the nonzero element characteristics of sparse signals and the sparse structure of system blocks to effectively improve the system's channel estimation, active user detection, and data decoding. To take advantage of the fast attenuation characteristics of nonzero components in the sparse signals, the traditional least absolute shrinkage and selection operator (LASSO) problem can be transformed into an iterative weighting problem

$$
\min_{\bar{h}_l \in \mathbb{C}^J} \quad \left\| \bar{\mathbf{h}}_l \right\|_{1,\mathrm{w}}
$$
$$
\text{s.t.} \quad \left\| \mathbf{y}_l^p - \mathbf{S}\bar{\mathbf{h}}_l \right\|_2 \le \varepsilon, \tag{9}
$$

where $\left\| \bar{\mathbf{h}}_l \right\|_{1,\mathrm{w}} = \sum_{n=1}^J \omega_n |\bar{h}_{nl}| |\bar{h}_{nl} \triangleq \alpha_n h_{nl}$, $\omega_n = 0, n \in \Lambda$, $\omega_n = 1$, $n \notin \Lambda$, ε is the tolerable estimation error and the set Λ is defined as $\Lambda = \{n : \bar{h}_{nl}\} \ne 0$. Since the information of the nonzero element part of the sparse signals is more reliable, the estimation of some nonzero elements can be stopped after several iterations. In addition, the active user set is updated based on the symbol energy of unreliable information and the accurate or approximate total number of active users known by the BS. This can not only reduce the false alarm rate but also reduce the missed detection rate and even achieve a complete detection of active users. For the uplink transmission stage, the multiple user signals received by the BS are decomposed into linear combinations of different signal vectors and the maximum likelihood (ML) detection problem is transformed into two separable subproblems, using alternate minimization algorithm. By analyzing the information of each iteration, the inversion of large-scale matrices is avoided, thereby reducing the computational complexity of multiuser data detection.

3.4. Joint Blind Detection Scheme of Large-Scale IoT Activity Users and Data Based on Double-Sparse Learning. Due to the transmission latency requirement, it is necessary to further reduce the pilot symbol transmission. Different from the active user and data classification detection scheme, the active user and data joint blind detection scheme based on the double-sparsity learning needs to devised into the data packet as the user's identity, and at the same time, the activity can be completed in one transmission stage for user and data detection. We need not to estimate the channel information firstly, which can help reduce the transmission latency further. The specific technical solution is shown in Figure 6. Therefore, considering the symbol synchronization and authorization-free uplink NOMA system, users are allowed to initiate data packet transmission at the beginning of any symbol interval. Whenever the user wants to transmit a message, it generates a data packet that carries the message and the identity of the sending user. Once activated, the user transmits the data packet to the BS at L consecutive symbol intervals. Assume that the jth data packet of user n is denote by $\mathbf{c}_n^{(j)} \in \mathbb{C}^{L \times 1}$ and $s_{n,t}$ represents the transmission symbol of user n in the tth symbol interval. As shown in Figure 7, $t_n^{(t)}$ represents the symbol interval at which user n starts to transmit the jth data packet, $s_{n,t_n^{(j)}+k-1} = c_n^{(j)}(k), k = 1, 2, \cdots, L$. The symbol of the dormant user can be represented by zero. Therefore, when $t_n^{(j)} \le t \le t_n^{(j)} + L$ holds, the transmission symbol of user n is $s_{n,t} = c_n^{(j)}(t - t_n^{(j)} + 1)$; otherwise $s_{n,t} = 0$. For the data packet $c_n^{(j)}$, the channel from user n to the BS can be expressed by $\mathbf{h}_n^{(j)} = \sqrt{\beta_n}\mathbf{g}_n^{(j)}$, where β_n and $\mathbf{g}_n^{(j)}$ represent large-scale and short-scale fading, respectively. The channel state remains unchanged during the entire transmission period of the data packet $c_n^{(j)}$, which is the block fading channel.

In the tth symbol interval, the signal received by the BS can be expressed as

$$
\mathbf{y}_t = \mathbf{H}_t \mathbf{s}_t + \mathbf{w}_t \in \mathbb{C}^{M \times 1}, \tag{10}
$$

where $\mathbf{H}_t = [\mathbf{h}_{1,t}, \mathbf{h}_{2,t}, \cdots, \mathbf{h}_{N,t}]$ and $\mathbf{s}_t = (s_{1,t}, s_{2,t}, \cdots, s_{N,t})^T$. If $t_n^{(j)} \le t \le t_n^{(j)}$ holds, then, we have $\mathbf{h}_{n,t} = \mathbf{h}_n^{(j)}$. Otherwise, $\mathbf{h}_{n,t}$
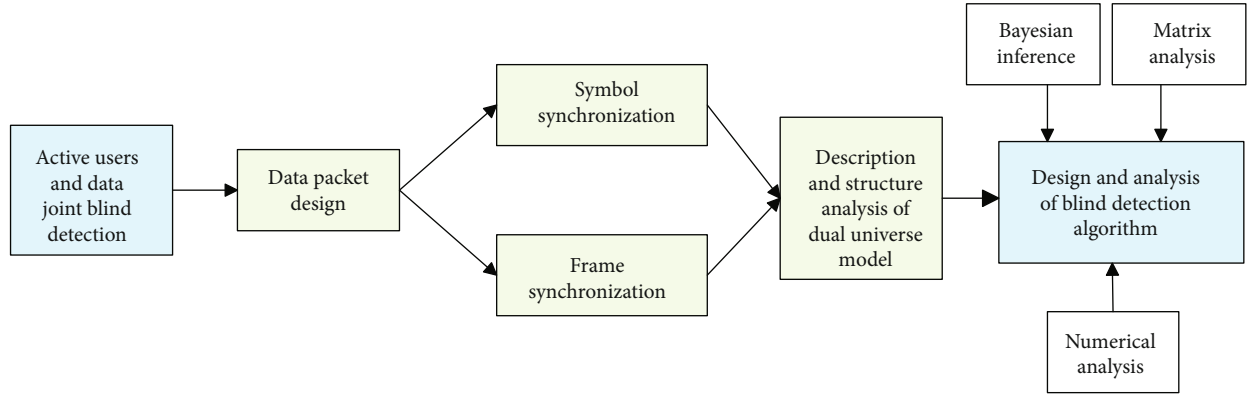
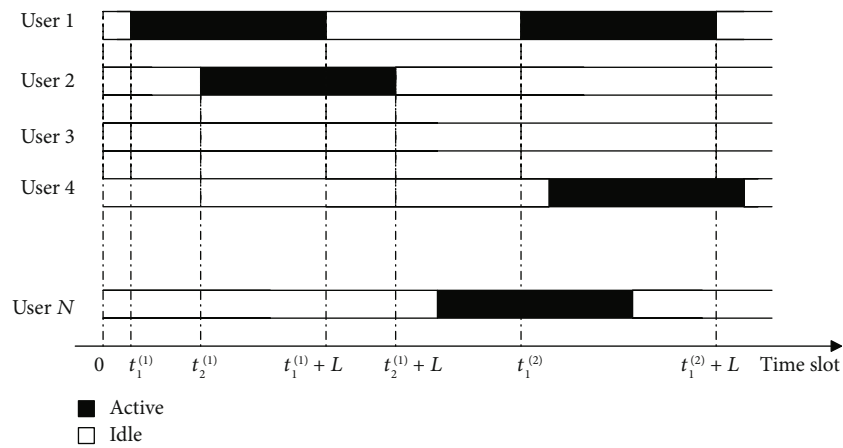Figure 6: Technical design of joint blind detection solution for large-scale IoT activity users and data.

Figure 7: User activity status within the symbol interval.

= 0. This problem combines active user detection, channel estimation, and data detection, which can help reduce the system's transmission latency. Note that the representations of the channel and the signal are both sparse, so that this problem is also called as the double-sparsity problem. The performance analysis of the model and the use of the double sparsity of the received signal model to design detection algorithms should be developed.

We firstly use the double sparsity of the received signal model to describe the iterative information of the estimated channel and data estimation as a conditionally independent random distribution and use a blind detection algorithm to estimate the signal in each sparse block. Secondly, a low-complexity information iterative algorithm is used between the independent sparse blocks, which can make good use of the sparseness of the signals to improve the system performance and reduce system transmission latency. Finally, numerical results should be provided to verify the reliability of the above theoretical results and correctness.

4. Conclusions

In this paper, we performed a comprehensive survey on the transmission schemes for large-scale IoT with NOMA, which could support multiple users to use the same fre-

quency resources as long as the interference among users could be addressed. The NOMA technique could help improve the frequency reuse efficiency significantly and was viewed as one of the most efficient candidate of the techniques for the next-generation wireless communications. However, there still existed a lot of challenges on the transmission schemes for large-scale NOMA systems, including the short-data packet transmission, active user detection, channel estimation, and data detection. To address these challenges, we firstly reviewed the short-data packet transmission in the large-scale NOMA system and then reviewed the active user detection and channel estimation techniques for the considered system. We further described the data detection algorithms for large-scale NOMA systems. In future works, we will incorporate some intelligent algorithms, such as the deep learning-based algorithm [45–47], deep Q-network-based algorithms [48], federated algorithms [49, 50], and transfer learning-based algorithms [51], to the considered system to enhance the system transmission performance.

Data Availability

The data in this work can be available through email to the authors of this paper.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

References

[1] J. Xia, "Secure cache-aided multi-relay networks in the presence of multiple eavesdroppers," *IEEE Transactions on Communications*, vol. 67, no. 11, pp. 7672–7685, 2019.

[2] B. Wang, F. Gao, S. Jin, H. Lin, and G. Y. Li, "Spatial- and frequency-wideband effects in millimeter-wave massive MIMO systems," *IEEE Trans. Signal Processing*, vol. 66, no. 13, pp. 3393–3406, 2018.

[3] X. Hu, C. Zhong, Y. Zhu, X. Chen, and Z. Zhang, "Programmable metasurface-based multicast systems: design and analysis," *IEEE Journal on Selected Areas in Communications*, vol. 38, no. 8, pp. 1763–1776, 2020.

[4] K. He, "Ultra-reliable MU-MIMO detector based on deep learning for 5G/B5G-enabled IoT," *Phys. Commun.*, vol. 43, pp. 1–7, 2020.

[5] X. Hu, J. Wang, and C. Zhong, "Statistical CSI based design for intelligent reflecting surface assisted MISO systems," *Science China: Information Science*, vol. 63, no. 12, 2020.

[6] J. Zhao, X. Sun, Q. Li, and X. Ma, "Edge caching and computation management for real-time Internet of vehicles: an online and distributed approach," *IEEE Transactions on Intelligent Transportation Systems*, vol. 22, no. 4, pp. 2183–2197, 2021.

[7] Y. Xu, C. Shen, D. Cai, and G. Zhu, "Latency constrained non-orthogonal packets scheduling with finite blocklength codes," *IEEE Transactions on Vehicular Technology*, vol. 69, no. 10, pp. 12312–12316, 2020.

[8] Z. Na, Y. Liu, J. Shi, C. Liu, and Z. Gao, "Uav-supported clustered noma for 6g-enabled Internet of things: trajectory planning and resource allocation," *IEEE Internet of Things Journal*, vol. 12, pp. 1–1, 2020.

[9] C. Li, "Dynamic offloading for multiuser muti-CAP MEC networks: a deep reinforcement learning approach," *IEEE Transactions on Vehicular Technology*, vol. 70, no. 3, pp. 2922–2927, 2021.

[10] X. Pang, J. Tang, N. Zhao, X. Zhang, and Y. Qian, "Energy-efficient design for mmwave-enabled noma-uav networks," *Science China Information Sciences*, vol. 64, no. 4, 2021.

[11] L. Yang, F. Meng, J. Zhang, M. O. Hasna, and M. D. Renzo, "On the performance of ris-assisted dual-hop UAV communication systems," *IEEE Transactions on Vehicular Technology*, vol. 69, no. 9, pp. 10385–10390, 2020.

[12] Z. Wang, "An adaptive deep learning-based UAV receiver design for coded MIMO with correlated noise," *Physical Communication*, vol. 45, no. 101295, pp. 1–8, 2021.

[13] Y. Guo and S. Lai, "Emergency semantic feature vector extraction from WiFi signals for in-home monitoring of elderly," *IEEE Journal of Selected Topics in Signal Processing*, vol. 23, pp. 1–12, 2021.

[14] X. Lai, "Secure mobile edge computing networks in the presence of multiple eavesdroppers," *IEEE Transactions on Communications*, vol. 65, pp. 1–12, 2021.

[15] Y. Zhang and J. Wang, "CV-3DCNN: complex-valued deep learning for CSI prediction in FDD massive MIMO systems," *IEEE Wireless Communications Letters*, vol. 10, no. 2, pp. 266–270, 2021.

[16] X. Li, H. Mengyan, Y. Liu, V. G. Menon, A. Paul, and Z. Ding, "I/Q imbalance aware nonlinear wireless-powered relaying of B5G networks: security and reliability analysis," *IEEE Transactions on Network Science and Engineering*, vol. 5, no. 33, pp. 1–1, 2020.

[17] W. Zhou, "PSO based offloading strategy for cache-enabled mobile edge computing UAV networks," *Cluster Computing*, vol. 53, pp. 1–8, 2021.

[18] M. Liu and F. Tang, "6G: opening new horizons for integration of comfort, security, and intelligence," *IEEE Wireless Communications*, vol. 27, no. 5, pp. 126–132, 2020.

[19] X. Li, Q. Wang, and M. Liu, "Cooperative wireless-powered NOMA relaying for B5G IoT networks with hardware impairments and channel estimation errors," *IEEE Internet of Things Journal*, vol. 8, no. 7, pp. 5453–5467, 2021.

[20] S. Tang, "Dilated convolution based CSI feedback compression for massive MIMO systems," *IEEE Transactions on Vehicular Technology*, vol. 61, pp. 1–5, 2021.

[21] X. Li, M. Zhao, M. Zeng et al., "Hardware impaired ambient backscatter NOMA systems: reliability and security," *IEEE Transactions on Communications*, vol. 69, no. 4, pp. 2723–2736, 2021.

[22] A. Li, "Enhancing the physical layer security of cooperative NOMA system," in *2019 IEEE 3rd ITNEC*, vol. 343no. 34, pp. 2194–2198, 2019.

[23] Y. Cao, N. Zhao, F. R. Yu et al., "Optimization or alignment: secure primary transmission assisted by secondary networks," *IEEE Journal on Selected Areas in Communications*, vol. 36, no. 4, pp. 905–917, 2018.

[24] T. Zhao, G. Li, G. Zhang, and C.-X. Zhang, "Security-enhanced user pairing for MISO-NOMA downlink transmission," *2018 IEEE GLOBECOM*, vol. 68, no. 36, pp. 1–6, 2018.

[25] B. M. ElHalawany and K. Wu, "Physical-layer security of NOMA systems under untrusted users," in *2018 IEEE Global Communications Conference (GLOBECOM)*, vol. 534no. 43, pp. 1–6, 2018.

[26] J. Xia, D. Deng, and D. Fan, "A note on implementation methodologies of deep learning-based signal detection for conventional MIMO transmitters," *IEEE Transactions on Broadcasting*, vol. 66, no. 3, pp. 744-745, 2020.

[27] S. Lai, "Intelligent secure mobile edge computing for beyond 5G wireless networks," *Physical Communication*, vol. 45, no. 101283, pp. 1–8, 2021.

[28] G. Gui, F. Liu, J. Sun, J. Yang, Z. Zhou, and D. Zhao, "Flight delay prediction based on aviation big data and machine learning," *IEEE Transactions on Vehicular Technology*, vol. 69, no. 1, pp. 140–150, 2020.

[29] X. Li, M. Zhao, Y. Liu, L. Li, Z. Ding, and A. Nallanathan, "Secrecy analysis of ambient backscatter NOMA systems under I/Q imbalance," *IEEE Transactions on Vehicular Technology*, vol. 69, no. 10, pp. 12286–12290, 2020.

[30] R. Zhao, "Deep reinforcement learning based mobile edge computing for intelligent Internet of things," *Physical Communication*, vol. 43, pp. 1–7, 2020.

[31] X. Hu, C. Zhong, Y. Zhang, X. Chen, and Z. Zhang, "Location information aided multiple intelligent reflecting surface systems," *IEEE Transactions on Communications*, vol. 68, no. 12, pp. 7948–7962, 2020.

[32] J. Zhao and S. Ni, "Multiband cooperation for 5G hetnets: a promising network paradigm," *IEEE Vehicular Technology Magazine*, vol. 14, no. 4, pp. 85–93, 2019.

[33] J. Yang, D. Ruan, J. Huang, X. Kang, and Y. Shi, "An embedding cost learning framework using GAN," *IEEE Transactions on Information Forensics and Security*, vol. 15, pp. 839–851, 2020.

[34] Z. Zhou, H. Yu, S. Mumtaz, S. Al-Rubaye, A. Tsourdos, and R. Q. Hu, "Power control optimization for large-scale multi-antenna systems," *IEEE Transactions on Wireless Communications*, vol. 19, no. 11, pp. 7339–7352, 2020.

[35] M. D. Renzo and P. Guan, "Stochastic geometry modeling and systemlevel analysis of uplink heterogeneous cellular networks with multiantenna base stations," *IEEE Transactions on Communications*, vol. 64, no. 6, pp. 2453–2476, 2016.

[36] C. E. Shannon, "A mathematical theory of communication," *Bell Systems Technical Journal*, vol. 27, no. 4, pp. 623–656, 1948.

[37] Y. Du, B. Dong, W. Zhu et al., "Joint channel estimation and multiuser detection for uplink grant-free NOMA," *IEEE Wireless Communications Letters*, vol. 7, no. 4, pp. 682–685, 2018.

[38] H. He, C. Wen, and S. Jin, "Bayesian optimal data detector for hybrid mmwave MIMO-OFDM systems with low-resolution adcs," *IEEE Journal of Selected Topics in Signal Processing*, vol. 12, no. 3, pp. 469–483, 2018.

[39] T. Hara and K. Ishibashi, "Grant-free non-orthogonal multiple access with multiple-antenna base station and its efficient receiver design," *IEEE Access*, vol. 7, pp. 175717–175726, 2019.

[40] D. Cai, "Active user and data detection for uplink grant-free NOMA systems," *China Communications*, vol. 17, no. 11, pp. 22–38, 2020.

[41] B. Wang, L. Dai, Y. Zhang, T. Mir, and J. Li, "Dynamic compressive sensing-based multi-user detection for uplink grant-free NOMA," *IEEE Wireless Communications Letters*, vol. 20, no. 11, pp. 2320–2323, 2016.

[42] Q. Zou, H. Zhang, D. Cai, and H. Yang, "A low-complexity joint user activity, channel and data estimation for grant-free massive mimo systems," *IEEE Signal Processing Letters*, vol. 27, no. 33, 2020.

[43] T. Liu, S. Jin, C.-K. Wen, M. Matthaiou, and X. You, "Generalized channel estimation and user detection for massive connectivity with mixedadc massive MIMO," *IEEE Transactions on Wireless Communications*, vol. 18, no. 6, pp. 3236–3250, 2019.

[44] T. Ding, X. Yuan, and S. C. Liew, "Sparsity learning-based multiuser detection in grant-free massive-device multiple access," *IEEE Transactions on Wireless Communications*, vol. 18, no. 7, pp. 3569–3582, 2019.

[45] J. Xia and L. Fan, "Computational intelligence and deep reinforcement learning for next-generation industrial IoT," *IEEE Transactions on Network Science and Engineering*, vol. 99, pp. 1–12, 2021.

[46] L. He and K. He, "Towards optimally efficient search with deep learning for large-scale mimo systems," *IEEE Transactions on Communications*, vol. 76, pp. 1–12, 2022.

[47] K. He, "Learning based signal detection for MIMO systems with unknown noise statistics," *IEEE Transactions on Communications*, vol. 69, no. 5, pp. 3025–3038, 2021.

[48] S. Tang, "Battery-constrained federated edge learning in UAV-enabled IoT for B5G/6G networks," *Physical Communication*, vol. 47, no. 101381, pp. 1–9, 2021.

[49] Y. Guo, "Efficient and flexible management for industrial Internet of things: a federated learning approach," *Computer Networks*, vol. 192, pp. 1–9, 2021.

[50] Z. Zhao, "System optimization of federated learning networks with a constrained latency," *IEEE Transactions on Vehicular Technology*, vol. 21, pp. 1–5, 2021.

[51] L. Chen, "Intelligent ubiquitous computing for future UAV-enabled MEC network systems," *Cluster Computing*, vol. 48, no. 99, pp. 1–8, 2021.