Math.Comput.Sci. (2023) 17:10
https://doi.org/10.1007/s11786-023-00561-y

Mathematics in Computer Science

Selected Aspects of Non orthogonal Multiple Access for Future Wireless Communications

Adam Flizikowski · Tomasz Marciniak · Tadeusz A. Wysocki · Olutayo Oyerinde

Received: 9 September 2022 / Revised: 19 March 2023 / Accepted: 5 April 2023 / Published online: 25 May 2023

Abstract In this paper overview of recent selected works that deal with novel directions in which Non orthogonal multiple access (NOMA) research is progressing is presented. These include the cell-free NOMA, deep learning extensions and optimizations of NOMA, energy optimization and task offloading with mobile-edge computing, NOMA and physical layer security, as well as virtualization, centralized-RAN aspects. All these are hot issues towards deployments of NOMA in the designs of beyond 5G and 6th generation (6G) wireless communication networks. Even though 3rd Generation Partnership Project (3GPP) has not yet made the decision regarding which NOMA techniques should be adopted, it seems like researchers already indicate clearly that NOMA has important place in the future network deployments based on ultra-density, novel 5G use-cases (massive machine type communications, ultra-reliable low latency communications). This paper highlights the most promising directions for NOMA research. The paper is summarized with necessary steps that are required to get NOMA into practical usage.

Keywords NOMA · Energy optimization · Physical layer security

Mathematics Subject Classification 68M10

1 Introduction

This paper provides an overview of NOMA as an important candidate for multiple-access scheme for beyond 5G wireless networks. Our contribution addresses NOMA in connection with of Things (IoT)/Massive Machine

A. Flizikowski (✉) · T. Marciniak · T. A. Wysocki
Bydgoszcz University of Science and Technology, Kaliskiego 7, 85-796 Bydgoszcz, Poland
e-mail: Adam.Flizikowski@pbs.edu.pl

T. Marciniak
e-mail: Tomasz.Marciniak@pbs.edu.pl

T. A. Wysocki
e-mail: Tadeusz.Wysocki@pbs.edu.pl

O. Oyerinde
School of Electrical and Information Engineering, University of the Witwatersrand, Johannesburg 2050, South Africa
e-mail: Olutayo.Oyerinde@wits.ac.za

Type Communications (mMTC) traffic, forecasted to grow largely in the coming years. Moreover, NOMA is a promising technique to help address aspects of task offloading (with Mobile edge computing (MEC)), incorporate some physical layer security, to mention just a few benefits of using it.

The main aim of this paper is to identify most interesting solutions beyond state-of-the-art available nowadays and indicate directions for NOMA that make it even more attractive in combination with edge computing, cloud-RAN and especially AI/ML to improve future 6G networks. In performing the research for relevant papers the key considerations were:

- superficial role of NOMA to address the IoT use-cases and networks
- the required novel resource management and allocation aspects for NOMA (largely relying on the AI/ML techniques) to assure futher improvements in spectral efficiency and lowering complexity of SIC receivers
- the ultra-dense network triggered modifications of the legacy cellular RAN paradigm towards the direction of cell-free, and cell-free mMIMO techniques
- the role of NOMA in combination with the MEC as it is becoming especially important due to (i) ubiquitous access techniques in hybrid terrestrial and non-terrestrial networking, (ii) its importance regarding fulfillment of 6G sustainability goals where energy-efficiency becomes essential driver of design and deployment decisions
- perspective of smooth alignment with existing OFDMA based 4G and 5G networks.

There are various types of NOMA (as already studied by 3GPP in technical reports) and to date, there is no consensus on which particular type of NOMA should be supported by next releases of standards (Rel.17 and beyond). However, existing technology evaluations by the 3GPP show that the different techniques do not differ much while validated in the system or link level simulations [3,4]. The authors of this paper foresee valuable usage of NOMA together with novel paradigms for mobile systems including for example a Cell-free (CF) network. Additionally, we believe that concepts like cloud/edge, big data, and virtualization are of same importance for both domains i.e. IoT platforms as well as the NOMA. Both topics studied in this paper have strong and direct influence on the capacities and capabilities of NOMA assisted IoT/mMTC future use-cases. From the technical perspective the multi-access signatures e.g. superposition coding "allows exploiting the high channel disparities among users, for enhancing the weak users' rates without incurring much degradation of strong users' rates. It has been reported in various works that NOMA is particularly vulnerable to inter-cell interference—therefore true impact of integrating NOMA in future systems would strongly depend on the capability of the global network to address issued related to efficient mitigation and handling of inter-cell interference [50]". According to [105] there will be four pillars of 6G: global coverage considering both terrestrial and non-terrestrial networks, wide range of electromagnetic spectrum will be utilized (mmWave, Terahertz (THz) and optical), big data sets generated using extremely heterogeneous networks will be processed by artificial intelligence and be exploited, and eventually, security in 6G will need to be strengthened. NOMA can facilitate addressing at least a few of the pillars. In addition, the future B5G and 6G networks will be based on ultra-dense paradigm, where massiveness, latency, and computing will be important factors of successful applications in different scenarios. To address the capabilities of the future networks novel network architectures are needed e.g. "software defined network/network functions virtualization (SDN/NFV), dynamic network slicing, Software Based Architecture (SBA), Cognitive service architecture (CSA), and CF architectures" [105]. Moreover, the current 3GPP standardization includes multiuser superposition transmission (MUST) NOMA as optional mode in Release 15. This is due to a substantial variety of NOMA proposals for downlink (15 proposals have bhaveeen delivered to date). Similarly, for Release 16 where NOMA for uplink was a study item, 20+ NOMA proposals submitted prohibited consensus building and effectively forced a removal of NOMA from 5G New Radio (NR) specification. Hence, the unified NOMA framework is necessary for 5G+/6G in contrast to currently purpose-build NOMA versions: sparse code multiple access (SCMA) for mMTC or power-domain NOMA for enhanced Multimedia Broadcast (eMBB). Additionally, different error control codes are more suitable for specific 5G use-cases: e.g. Low-Density Parity Check (LDPC) is better suited for eMBB while polar codes provide more value in case of low-latency requirement. Moreover, channel coding is different whether downlink or uplink is considered. NOMA can be also helpful in case of massive connectivity (with its grant-free access), reducing the delay and energy consumption of MEC offloading. The composition of this paper contents has been summarized in Fig. 1, and

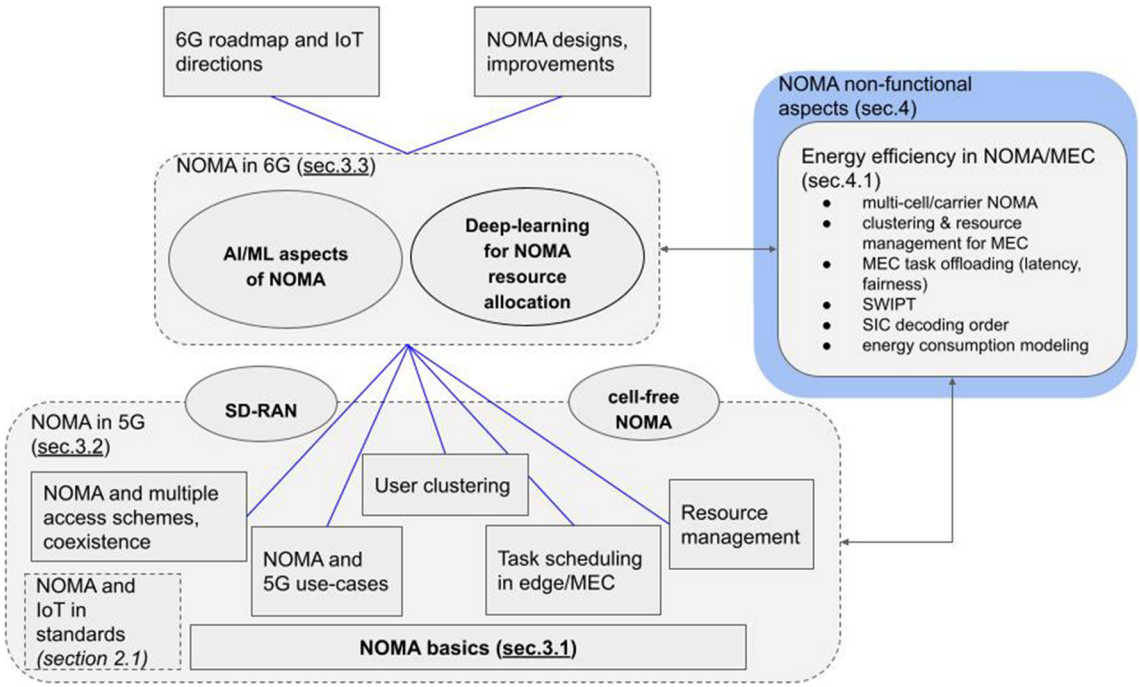

Fig. 1 Mapping of this paper contents to sections

the remainder of this paper is arranged as follows. In Sect. 2 general information on NOMA from the perspective of standardization and in relation to NOMA key use-case mMTC communication is discussed. In Sect. 3, various aspects of NOMA for 5G and 6G are presented, highlighting key research works that deal with NOMA schemes and techniques. In addition, Sect. 3 also covers main applications for NOMA such as software-defined software-defined RAN (SD-RAN) and cell-free or multi-cell approach, which is the key candidate for beyond 5G. Section 3 is concluded with the role of machine learning and deep learning from the view point of NOMA optimization and design updates. Section 4 deals with research on the use of NOMA with MEC for task offloading and energy efficiency optimizations. Finally, the summary and conclusions of the paper are provided in Sect. 5. The list of various abbreviations/acronyms used in this paper are provided at the end of the paper.

2 NOMA and IoT in Standards

It is essential to have better understanding of the standardization perspective behind NOMA and IoT, and especially what are the main overlaps (common points) for the future planning of IoT deployments vis-a-vis the use of NOMA access schemes. In general, NOMA is considered for inclusion into main standardization of the 3GPP activities, as it has been extensively studied as a candidate for providing additional capacity for future (6G) wireless systems. The release 15 of the 5G NR, has specified Grant free (GF) transmission in NR to reduce signaling overhead and latency, which is suitable for both Ultra low latency communications (URLLC) and mMTC, especially in the uplink. In the next 5G NR release standards (Rel.16, Rel.17) in addition to connectivity through the cellular infrastructure, side link connectivity with another IoT device or smartphone will be introduced. Moreover, GF transmission will be extended to support side link transmissions and be enhanced with NOMA. Furthermore, NarrowBand IoT (NB-IoT) and LTE-MTC [Machine Type Communication] (LTE-M) will be integrated with 5G NR to provide dedicated MTC services [62]. The summary of Rel.16 work items suggested enhancements to mMTC in the recently published specification [2]. These enhancements include improved coverage RAN feature, control of user data rate sent to/from user-equipment (UE), control plane congestion control, inter-UE QoS for NB-IoT, etc. There are also “5G lite”

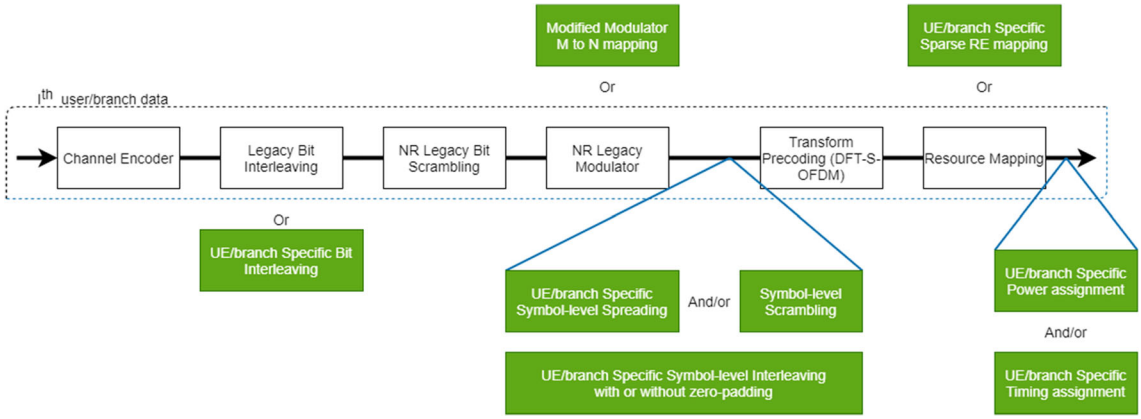

Fig. 2 NOMA transmitter extensions options (based on 3GPP)

solutions for Rel.16/17 that may be an interesting point in the delivery of NOMA as they are targeting some "balance" between 5G use-case and deliver solution that resembles Long Term Evolution (LTE) from performance perspective but already uses the 5G signaling and architecture, which allows for complementing or even replacing the NB-IoT and LTE-M in the future. A NOMA background, especially for the uplink direction together with comprehensive simulation results of NOMA (including link and system levels), has been presented in the 3GPP document [4]. The generic NOMA transmitter side processing diagram as well as receiver processing diagram are both presented in the document (Fig. 2). Receiver complexity is analyzed and compared for different receiver types. Moreover, the paper provides 35 different test scenarios with detailed settings for Link level simulation (LL), including carrier frequency, 5G use-case, Signal to noise ratio (SNR) distribution, waveform, channel model, Transport block size (TBS) size, and number of UEs. In the results from the simulation, results show more differences between NOMA schemes for larger TBS sizes. As for smaller TBS performance, differences are small for all NOMA schemes. Performance degradation is also identified for realistic channel (in the range of 2–5 dB). In a system level simulation, all three use-cases are considered (together with detailed parameters): eMBB, URLLC, and mMTC. Similar suitability and performance considerations for multiple-access technologies, including NOMA, are also included in LL/System level Simulation (SLS) results documented in the report for the 3GPP Release 14 [3]. The green elements in Fig. 2 depict modifications required by NOMA in the physical layer transmitter side.

In addition, [1] captures various features of mMTC terminals and their modifications to reduce costs and improve coverage along with various hardware simplifications that will enable production of low-cost MTC user equipment (UE). New IoT and edge computing capabilities drive the decentralization of architectures and topologies according to [7]. Alliance of IoT innovation (AIOTI) report highlights the important direction that "decentralization triggers a structural and regulatory change in main industrial sectors and intelligent infrastructures to achieve greater flexibility, agility to match demand/supply and responsiveness, while optimizing resource consumption. These new capabilities support energy efficiency and sustainability of edge applications deployment for implementing green and circular economy principles in and across various industrial sectors". More information about IoT as one of the main use-cases for utilizing NOMA is presented in [97].

3 Overview of NOMA

This paper considers NOMA techniques and solutions so far described in literature from the view point of 5G and beyond, concentrating in particular on two different scenarios of NOMA deployment:

- NOMA in 5G and beyond—this sub-section deals with current research in the framework of 5G or targeting beyond 5G aspects, e.g. cell-free concept. Concerning 5G, the sub-section focuses on NOMA/Orthogonal

multiple access (OMA) coexistence, NOMA in combination with legacy 5G architectures like SD-RAN, as well as gives some indications of similarities and differences in NOMA when combined with the various 5G use-cases.
- NOMA in 6G—this sub-section provides an overview of mainly the 6G research directions and main components of 6G and especially the Artificial Intelligence / Machine learning (AI/ML)/Deep Learning (DL) aspects in combination with the NOMA techniques

4 NOMA in 5G and Beyond

4.1 NOMA 5G Basics

In [9] two approaches in the context of 5G networks were proposed. The first is OMA and the second is NOMA. In the case of NOMA, there are two ways to implement it, i.e. power-domain multiplexing and code-domain multiplexing. Multiplexing in the power-domain is simple to implement as no significant changes are required in the networks, while code-domain multiplexing has a potential to increase spectral efficiency, but it requires a large transmission bandwidth and is not readily applicable to current systems.

In [46], fundamentals of power-domain NOMA with single and multiple antennas in both uplink and downlink settings are described. In addition, the basic principles of code-domain NOMA as well as various resource allocation techniques such as user pairing and power allocation for NOMA systems are described there. Thus, it discusses the basic form of cooperative NOMA and its variants. The authors in [48] proposed the same two approaches as in [46], focusing only on NOMA in the power domain, which imposes multiple users in the power domain and takes advantage of the channel gain difference between multiplexed users. While the mentioned article is primarily concerned with the power-domain superposition coding (SC)-based NOMA, authors have also presented a section about other NOMA classes. In addition to the basic principles and theoretical analysis, [27] presents prototype evaluations and field test results to verify actual NOMA efficiency gains. The authors have presented the underlying principles, advantages and disadvantages of NOMA, design fundamentals and key features of these dominant NOMA solutions, as well as user grouping and resource allocation. The paper also includes a systematic comparison of the NOMA schemes in terms of their spectral efficiency, system performance, receiver complexity, etc. Another comprehensive overview of the latest NOMA research results and innovations is provided in [28]. NOMA can be integrated with existing and future wireless systems due to its compatibility with other communication technologies. For example, NOMA has been shown to be compatible with conventional OMAs such as Time Division Multiple Access (TDMA) and OFDMA. Moreover, NOMA has been recently included in the forthcoming digital TV standard (ATSC 3.0), where it is referred to as layered division multiplexing (LDM). In particular, the spectral efficiency of TV broadcasts is improved by applying the NOMA principle and overlapping multiple data streams

4.1.1 Coexistence of NOMA and Other Access Techniques

Hybrid NOMA/OMA mode selection is considered by various authors in some other publications. The authors in [33] consider a multi-cell scenario where OMA/NOMA access is implemented in downlink. The proposed scheme focuses on maximizing the frequency reuse and bandwidth efficiency, by carefully selecting the NOMA/OMA mode, which is performed jointly with resource allocation. Hybrid NOMA/OMA scheme selection is also proposed in [11]. In [65] performance evaluation of a hybrid multiple access system combined with a novel pairing algorithm based on Modulation coding scheme (MCS) adjustment and extra transmission power (Tx power) allocation is introduced. The purpose of such pairing approach is to help the NOMA UEs reach the needed SINR while improving the overall system capacity. Moreover, the authors use mmWaves frequency range for transmission to further increase system capacity (due to larger available bandwidth). The MCS assignment and resources allocation in the model is done in two steps. In the first step, a preliminary assignment is done and only the Channel quality index (CQI)) reported by each UE is used as a reference. Then, the UEs map the wideband SINR calculated through simulation

into a CQI value. The minimum SINR for each CQI was estimated in the model under constraints of a 10 percent BLER and assuming an OMA transmission (without considering co-channel interference). In the second step, the system evaluates which UEs, if any, are candidates for NOMA. To determine this, a pairing method based on MCS adjustment and extra Tx power allocation is implemented. Such pairing guarantees that the throughput of each UE using NOMA either remains the same as if OMA were used or increases. A flexible Hybrid ARQ (HARQ) strategy is designed to make full use of each transmission slot, where the new signal for the user with QoS requirement (i.e. non best-effort) is transmitted in every transmission without waiting for the security user (SRU) to detect its messages correctly. Additionally, inspired by the randomize-and-forward relaying strategy [51], a randomized retransmission NOMA (RR-NOMA) scheme is designed to reduce the leakage of confidential messages, where the signals in every transmission are generated from independent randomized codebooks. RR-NOMA scheme outperforms the traditional cognitive NOMA scheme in terms of security-reliability trade-off.

Moreover, the RR-NOMA scheme achieves a better security-efficiency trade-off than the fixed retransmission NOMA (FR-NOMA) scheme in the low Service outage probability (SOP) region and outperforms the RR-OMA scheme when system parameters are determined properly. In [38], the authors analyzed and proved that the use of cooperative NOMA reduced the disconnection compared to cooperative OMA, and NOMA's performance improved for faster data rates compared to OMA. However, if the parameters were not properly matched, NOMA's performance was worse. Performance analysis of cooperative NOMA at intersections for vehicular communications in the presence of interference is considered in [38] as well. Cross Layer NOMA for femtocell users can be useful in such a scenario, where we require minimum delay without interference during transmissions. The authors combined Cognitive radio (CR) with NOMA (by multiplexing the two approaches under the heterogeneous networks scenario) and called it CR-NOMA. In CR-NOMA the data rate is guaranteed and the probability to connect two—weak and strong users—is high. In this scenario, the user with a poor channel can achieve a guaranteed Quality of service (QoS) without degrading performance of a user with a good channel. [19].

In [63] NOMA was used together with 3GPP-inspired user ranking technique. The authors study two resource allocation techniques for downlink NOMA for two users. They compare results for both systems OMA and NOMA to maximize cell-sum rate (CSR), as well as to show improved transmission rate region and improved SEC for the higher user density. The results obtained show that NOMA is better than OMA. Further research will be required to expand this model for N-users in cellular networks. A new approach for the access technology was made by the authors in [51]. They proposed a system that can dynamically choose a suitable technology between OMA and NOMA schemes for downlink communication. The scheme is valuable, because in some cases NOMA is not the best option and a system can dynamically select the best option when both technologies are available.

4.1.2 NOMA in Software Defined RAN Networks

Cloud-RAN, centralized-RAN, and virtual-RAN are different deployment options that appear on the horizon as a consequence of standards supporting virtualization, as well as prevailing trend in moving applications and services into a cloud. Different workloads in the cloud span from user-facing services (like computation services for IoT, various utility services, etc.), security services (like processing of Deep packet inspection (DPI), signal processing for attack identification and prevention), as well as infrastructure services (e.g. virtual RAN, 5G RAN). There is an essential difference between the various RAN deployment options concerning requirements for QoS and computation resources. The virtual 5G RAN deployment already challenges existing cloud data centers. The reason for this is because it requires special support, like access to acceleration technologies (Graphical processing unit (GPU), Field programmable array (FPGA)), smart Network interface card (NIC)), and the real-time processing for the 3GPP protocol layers like e.g. Physical layer (PHY) (especially Low-PHY considering the functional splits of 3GPP [52]). According to [17] the "exploiting interference that affects UL users can significantly improve their QoS and spectral efficiency. In C-RAN, multi-cell NOMA allows such interference exploitation." Fig. 3 provides conceptual view on C-RAN based multi-cell (or cell free) NOMA. In that case the "C" prefix refers to the centralization of processing, that can happen with cloud or edge resources. The main standard body influencing such developments is European telecommunications standards institute (ETSI) with its Network Function Virtualization

(NFV) group of specifications. The NFV architecture enables a unified approach to different workloads classifying them into: Virtual network function (VNF) and Physical network function (PNF). The first group refers to any workload that can be turned into virtual machines or containers. By doing so, they become hardware independent and thus multitude of general-purpose processors can be handling such a workload. The network function virtualization infrastructure Network function virtualization infrastructure (NFVI) is performing the role of a runtime environment in this scenario The functions that cannot be moved, or are provided as a legacy footprint (e.g. embedded HW) are also considered by the NFV specifications for completeness (this is refered to as PNF). However, not all of the 3GPP radio stack functions (whether VNF or PNF) can meet their performances' targets without support of accelerators of various kinds in order to meet stringent demands for the processing. The use of NOMA in combination with Fog RAN (FRAN) is studied in [50], where FRAN is described as an alternative to the C-RAN deployments. The main difference between FRAN and CRAN is reduction of requirements for fronthaul in FRAN as processing is shifted to the edge. The drawback of cloud processing where Baseband unit (BBU) are shared in the cloud to manage interference and resource allocation are the capacity limited fronthaul links, that may introduce transport delays. To overcome CRAN limitations FRAN is partially moving network intelligence, i.e., cloud computing and storage capabilities, closer to the network edge. The access points in FRAN - Fog Access Point (FAP) can perform distributed signal processing and radio resource allocation.

Utilization of FRAN/FAPs where RRM and interference mitigation can be performed with the use of NOMA is considered in [74,108,109], showing that the use of NOMA in downlink transmission of FAPs increases user fairness without sacrificing data rates. Additionally, combination of D2D and NOMA under FRAN is studied in [31,50] for jointly providing high data rates and low latency in eMBB. In the mMTC scenario, NOMA can boost uplink transmission by increasing a number of devices (per Resource block (RB)) that can be connected to the network in grant-free access. "In the downlink, NOMA multiplexes multiple user messages on the same basic RB unit. In the uplink, NOMA allows several users to simultaneously access the same basic RB unit without collisions. Moreover, cloud-based optimized packet scheduling, Radio resoure allocation (RRA) and Interference mitigation (IM) can jointly enhance the network reliability metrics of the users [50]". Unlike the conventional cloud computing operated in the remote cloud that suffers severe transmission latency via the Internet, MEC offers cloud computing capabilities at the edge of radio access network (e.g., at small-cell BSs) in close proximity to NB-IoT devices [71]. Through bringing intensive computation tasks from NB-IoT devices to MEC units, the low-latency as well as reliable computing services can be implemented for NB-IoT devices.

4.1.3 NOMA Resource Management and 5G Use-Cases

In the prior state-of-art, NOMA techniques are considered suitable for all the "5G triangle" use-cases (i.e. eMBB, URLLC, mMTC). However, use cases related to IoT seem to be the most relevant usage of NOMA. Grant-free NOMA is a generic technology that can bring benefits to mMTC, URLLC, eMBB small packet and two-step random-access channel scenarios [88], which leads to a collision if two or more users select the same resource for transmission. In such a situation, the receiver is unable to decode the data of users sharing the same RB. There are two ways of performing grant-free access: 1) UE's resources are pre-configured and periodically allocated, and each time when a packet arrives, the UE would choose the nearest allowable time-frequency resource for the uplink transmission, which is called Semi-persistent scheduling (SPS) based grant-free; 2) UE can randomly select a resource at any time for uplink transmission, leading to contention-based transmission. Stringent URLLC like requirements are also considered in [98] where authors deal with power control for delay-bounded IoT applications. Typical emerging IoT applications require a latency from 0.25 ms to 10 ms and an outage probability (or packet loss rate) in the order of 10^{-3} to 10^{-9} [78]. Short packet transmission in absence of a closed-loop control in a GF NOMA is addressed by deep learning in [37]. A remedy here is an open-loop selection of transmit power by users from the predefined pool of power values—solely based on their distance from the base station. Each IoT user acts as an agent and learns policy by interacting with wireless environment. To prevent Q-learning overestimation problem a double Deep Q-networks (DQN) based GF-NOMA is proposed. DQN converges faster than Q-learning under changing environments due to limiting action space based on previous learning. Information about channel can be

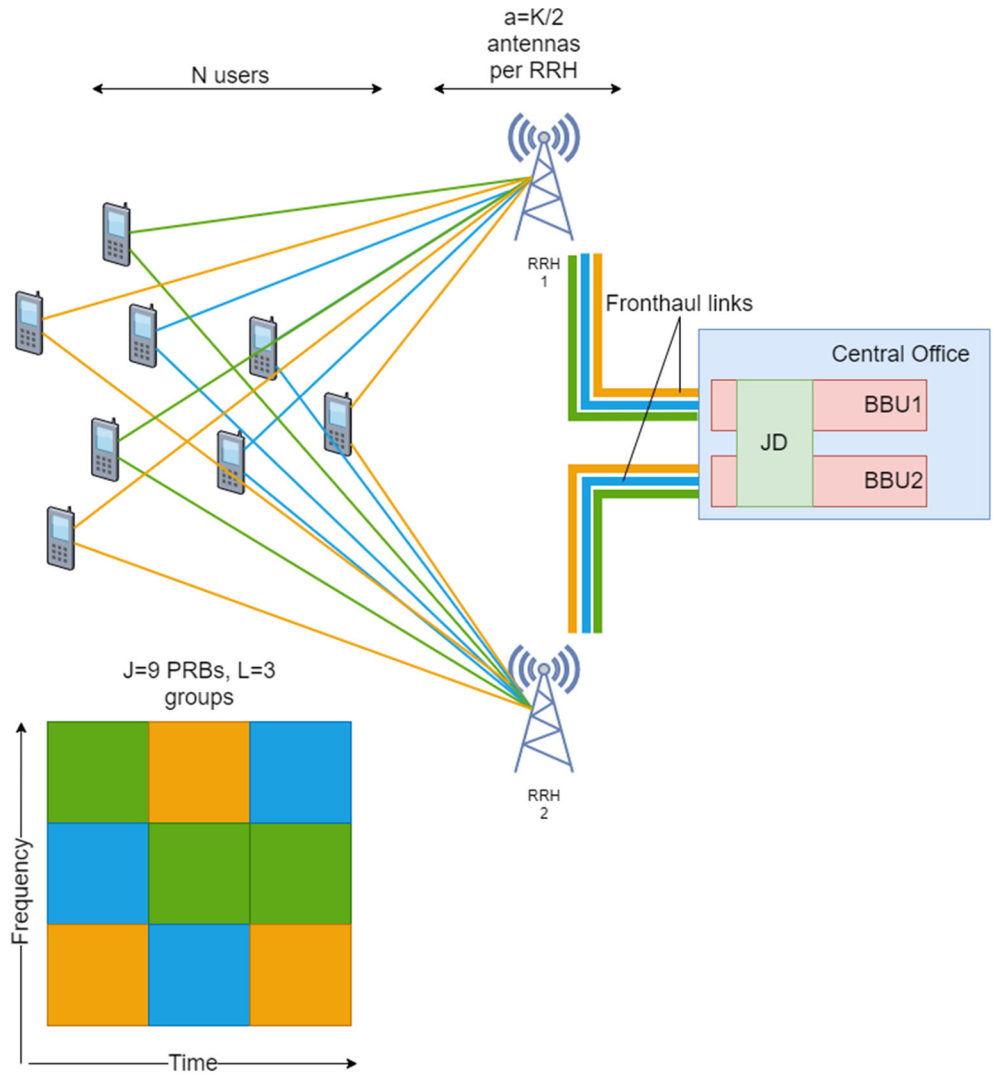

Fig. 3 Uplink NOMA in C-RAN (Source [18])

calculated via IoT users' geographical information and practical statistical models without information exchanges, which enables an open-loop control. The authors in [112] designed users and sub-channel clusters in a cell where number of users compete in a GF manner for several available sub-channels in each of the rings centered around a BS. The formulated long-term cluster throughput problem is solved via DRL based GF-NOMA algorithm for optimal sub-channel and power allocation. The authors in [37] "propose a multi-agent deep Q network (DQN) and double DQN based GF-NOMA algorithm for prototype power pool design, where the BS broadcasts this pool to all IoT users to avoid acquiring Channel state indication (CSI). Each IoT user can randomly select one power level for transmission that reduces complexity at BS and avoids massive information exchange between IoT user and the BS. Power selection from this well-designed prototype power pool guarantees distinct received power levels at the BS for execution of Successive interference cancellation (SIC) processes and reduces collision probabilities by allowing pilot sequence reuse [...]. The authors, in addition, consider uplink transmission in IoT networks with the traffic model of packets following the Poisson distribution. Further, the cell area was divided into different layers and a layer-based transmit power pool prototype was designed via multi-agent reinforcement learning (MARL). In the proposed framework, data transmitting IoT users select a transmit power based on their communication distance (layer) from

the well-designed prototype power pool for GF-NOMA transmission, without any information exchange between IoT user and the BS." IoT user acts as a learning agent and interacts with the environment. After learning from the mistakes, the IoT users in each layer find out the optimal transmit power level that maximizes network throughput. In the downlink, the SIC decoding order is fixed. The user with the stronger channel must first decode the message intended for the weaker user, subtract the corresponding signal, and then decode its own message. The weaker user will only be able to decode its own message. Contrary to that, the decoding order in uplink NOMA can be chosen arbitrarily. Many works related to uplink NOMA [23,35,47,100,107] suggest that the signal from the stronger user should be decoded first, such that the weaker user's signal is interference-free. In [12], the authors investigate the achievable link-layer rate of a two-user NOMA with short-packet communications i.e. the downlink URLLC case. Here the Shannon capacity brings too optimistic bound. Specifically, they formulate the effective capacity of the strong and weak users under heterogeneous delay QoS requirements. The overall reliability, which is the combination of the transmission error probability and the queueing delay violation probability, is investigated. The authors also derived closed-form expressions for the individual effective capacity of the two NOMA users. The achievable data rate of user i is given by equation as well as the queueing model of the NOMA communication for two users. Based on this assumption, effective capacity of both weak and strong users are derived under heterogeneous delay QoS requirements. Whereas in [77] the authors study the effect of delay in the uplink direction, under imperfect CSI and finite-length coding. It was indicated that the delay violation, as important metric in URLLC/mMTC scenarios, will be dependent on the SIC decoding order. In order to determine the delay performance of NOMA systems in the presence of decoding errors and error propagation, one must first analyze the decoding error probabilities due to imperfect CSI and finite block length channel coding. The authors concluded that even under realistic assumptions, NOMA may be suitable for low-latency communications, but only when joint decoding is used and only when there is a large difference between the two users' average SNR values. However, joint decoding may be difficult to implement in practice, especially for low-latency systems. With SIC decoding, NOMA often performs worse than OMA when considering low-latency communications with more realistic system effects. The authors in [22] applied interference cancellation schemes and superposition coding at a NOMA receiver, which can help with multiplexed multiple users on the same subchannel. There are some iterative algorithms with - according to the authors - "guaranteed convergence to deliver a competitive suboptimal solution". The performance evaluations presented indicate that the effectiveness of the proposed algorithm is better than other resource allocation schemes in NOMA or OFDMA system.

4.1.4 Beyond 5G NOMA - Cell Free NOMA

A recent concept studied with respect to NOMA is the CF or cell-less paradigm. This is the beyond 5G approach that changes baseline assumption that cells are independently allocating resource, and UE is attached to a single cell at a time. The cell-free is often combined with Massive MIMO (mMIMO). In [75], analysis of the influence of various linear precoders (maximum ratio transmission (MRT), full-pilot zero-forcing, modified regularized ZF are presented. The authors show that pilot contamination and non-ideal SIC degrade system performance when number of users is low. With the perfect SIC the modified-regularized ZF (mRZF) and Full pilot zero forcing (fpZF) significantly outperform MRT. The main benefit of the mRZF and fpZF is that only local CSI is required at AP, and exchange of CSI is not necessary between APs. Similarly, under massive MIMO scheme the conjugate beamforming is studied in [79] to group users into multiple clusters and decide whether to serve them with OMA or NOMA transmission. In [54] authors show that "OMA counterpart outperforms NOMA in terms of the achievable sum rate in the regime of the low number of users because the latter suffers from intra-cluster pilot contamination and imperfect SIC. However, NOMA can serve twice the (number of) users than that of OMA when two users are grouped into each cluster, while achieving a sum rate gain over OMA when the number of simultaneously served users grows large due to the proposed pilot assignment. A hybrid of primary massive MIMO and underlaid CF massive MIMO-NOMA is investigated in [76]. The authors argued that existing user-clustering methods for non-CF NOMA cannot be applied in the realm of CF-NOMA—and proposed low-complexity, sub-optimal user-clustering method that improves the achievable sum-rate of CF massive MIMO-NOMA network. In the paper, two networks

that co-exist were consider. Hence, Time division duplex (Time division duplex (TDD)) schemes that are perfectly synchronized to prevent PU-SU interference were employed

The authors propose the user-clustering scheme based upon the Jaccard distance coefficient [49] to find the most dissimilar secondary user (SU)s in the secondary network. Power allocation plays an important role in enhancing the performance in multi-cell NOMA [104]. NOMA is promising in the scenario with intensive data traffic and high user densities. Power allocation in multi-cell NOMA has significant impact on resource usage efficiency. The issue of multiple AP transmitting into a region and thus causing co-channel interference is addressed in [73] for the case of indoor Visual light communications (VLC). Here, suitable subcarrier allocation for the overlapping region where multiple VLC-Aps coexist is a suitable solution. The authors proposed a joint NOMA transmission scheme, where the users in multi-cell overlapping regions are jointly served by all the corresponding VLC access points. Two subcarrier allocation techniques, namely, area based subcarrier allocation, and user based subcarrier allocation are also studied. This area based scheme allocates powers to the users present in the overlapping regions based on their effective channel gains, and the different subcarriers are allocated to avoid inter-cell interference. The proposed scheme achieves significantly higher sum rates in comparison with the frequency reuse factor 2 (FR-2) NOMA scheme, which uses independent transmission across multiple cells. In the proposed scheme, the users in the overlapping regions receive data from multiple Light emitting diode (LED), thereby increasing their received SINR.

4.2 NOMA in 6G

4.2.1 Evolution Towards 6G

To understand future directions related to NOMA (and IoT as one of the main use-cases for NOMA) it is essential to understand the foreseen trends in the future mMTC deployments. According to [61] there are 12 trends that will drive evolution of the mMTC market, including among all: (a) autonomous mobility that will increase number of sensors and actuators needed, (b) ubiquitous connectivity in home and smart cities that will facilitate new possibilities at home, (c) Industry 5.0 (follow up trend of the current Industry 4.0 revolution which aims at delivering intelligent factory but at the same time assure sustainable, human-centric and resilient European industry) that will involve much more interactivity with even more diverse and stringent demands regarding wireless connectivity compared to Industry 4.0, (d) MTC that will play a fundamental role in facilitating the user experience and enabling novel and efficient man–machine interfaces to present data coming from machines in a more natural way (e) zero-latency, zero-energy expectations (f) value of data created with MTC is expected to largely increase with the support of AI/ML for its processing. Moreover, it is evident that IoT devices are usually powered by batteries or energy harvesters and are very limited in computing and storage capabilities to reduce costs and extend their lifetime. It is also important to notice [61] that the distributed ledger technologies are expected to reach also the mMTC domain. Additionally, it is highlighted by the authors that the following key trends will influence the shape and design of mMTC in the 6G networks:

- Zero-energy air interfaces are emerging.
- Cell-free networks that imply no explicit connection established between IoT devices of the future (so called "Massive Type Device") and a base station (BS) of a network cell, thus reducing signaling overhead while improving the reliability by enabling multiple BSs to demodulate the incoming transmissions, either separately or jointly.
- Increase of network heterogeneity—more micro-operators, more new RATs, smaller and denser cells.
- Infrastructure-less, dynamically formed networks for mission critical usage.
- The need to support growing number of traffic patterns and service classes.

The holistic network foreseen in 6G to deliver mMTC is presented in Fig. 4 [61].

No matter if it is a zero-energy or self-powered always-on IoT device, the wake-up-signal capture, initially introduced in 3GPP release 15 for NB-IoT and MTC devices and being further enhanced in release 16, necessitate a

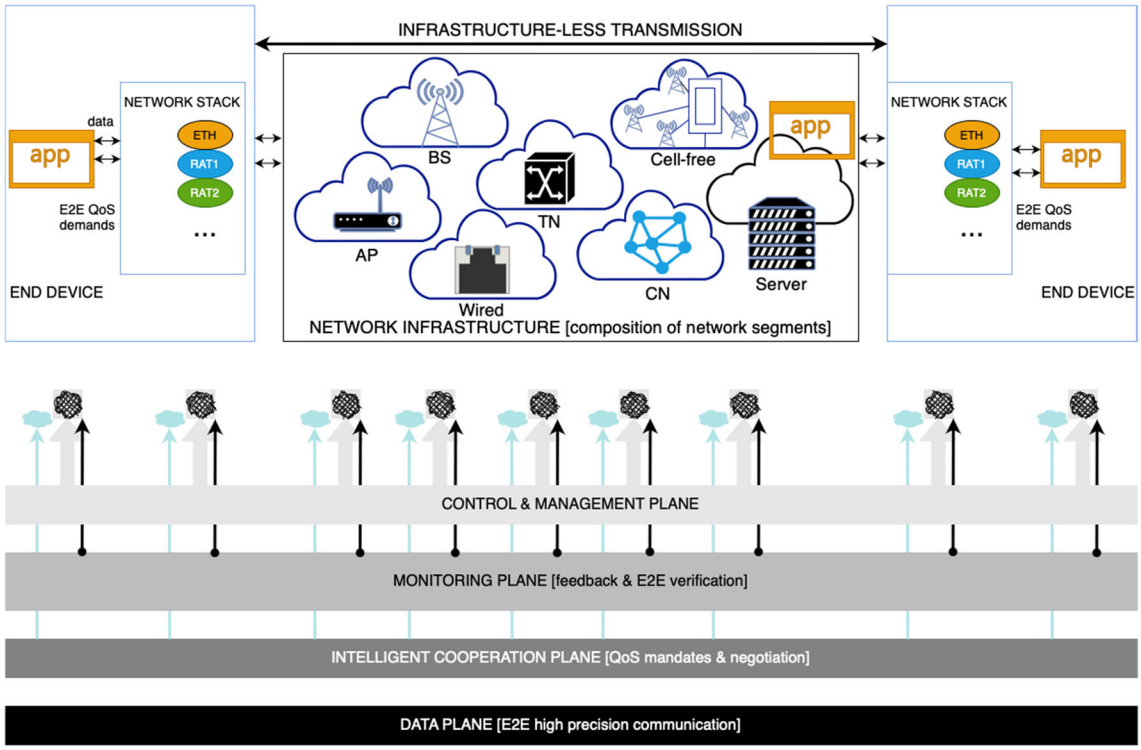

Fig. 4 Holistic view on 6G networks (Source [61])

receiver embedding adaptive blocks that can adapt themselves to the current context. The next generation devices for MTC will benefit from on-device intelligence blocks. However, embedding the required algorithmic functionalities on-chip while limiting the device's power consumption remains a challenge. Regarding NOMA, the authors believe that the success of this approach depends on both user detection and data decoding on the shared resources. The design of efficient non-orthogonal solutions capable of handling massive traffic over grant-free channels, possibly without channel state-information (CSI) required at the physical layer, together with tailored channel code design should be considered as key enablers at the PHY layer. At the MAC layer, specific traffic characteristics of MTC applications require both novel random access schemes and advanced (persistent) scheduling approaches. As presented in [72], the 6G evolution will be mainly about big data, machine learning and other related enhancements in various layers of the mobile network stack. Hence, it is worth noting that NOMA is a technique that is expected to benefit largely from future increase of built-in network intelligence. Examples of works that bring AI/ML into NOMA research are presented in Table 3-1, and the key benefits of their application have been studied in literature.

4.2.2 AI/ML Aspects of NOMA

This sub-section provides examples of research that include AI/ML in NOMA optimization and design. The following topics are discussed from the perspective of machine learning and NOMA:

- User clustering and power allocation for NOMA in mm-Wave.
- CSI prediction, support for spectrum sensing, awareness of user localization and throughput prediction.
- NOMA-based energy-efficient task scheduling for MEC and the total energy consumption with MEC.
- NOMA for ultra-dense deployment and grant-free access.
- Computation offloading and subcarrier allocation problem in multi-cell, Multi-carrier (MC) NOMA.
- Challenges of complex sensing model in cooperative spectrum sensing for non-orthogonal multiple access.

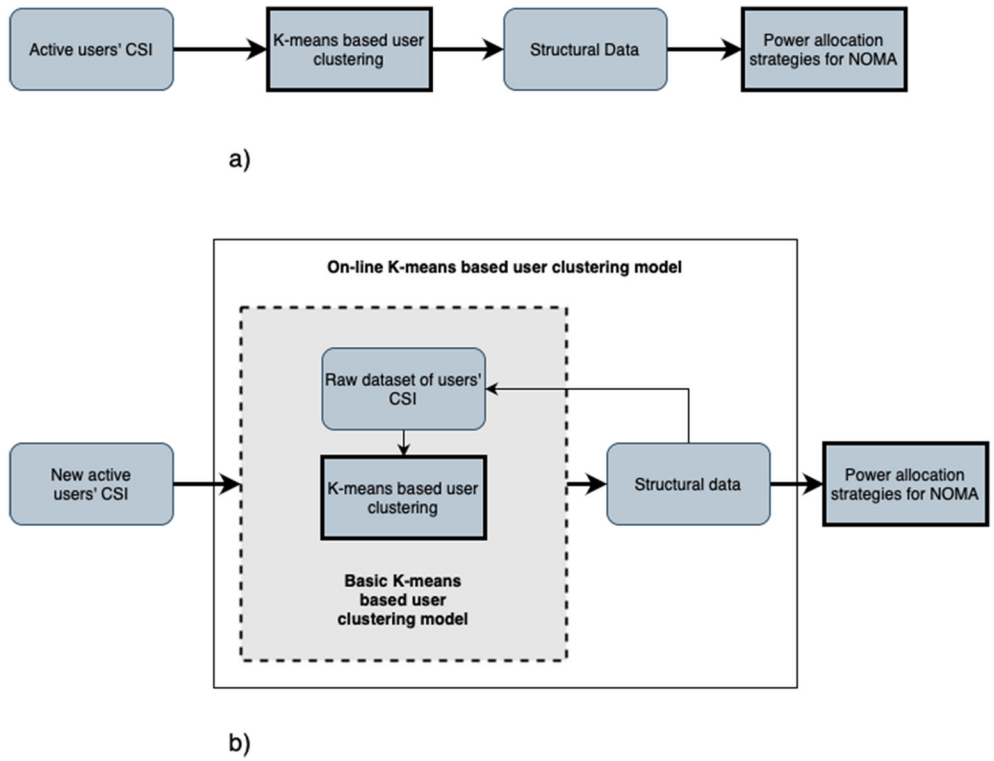

Fig. 5 Approach to machine-learning based NOMA realization (based on [25])

In [25], the analysis of the AI/ML algorithms for application in mmWave based NOMA systems is presented. In such systems, NOMA is applied for users utilizing the same beam. To improve clustering in NOMA, strong channel correlation is a desirable feature to allow improved aggregation of users into clusters. The authors then proposed a K-means enabled ML framework for user clustering for NOMA to maximize sum rate of the transmissions. They especially proposed low-complexity K-means based online user clustering algorithm. In addition, an optimal power allocation strategy is proposed for distributing power among user-clusters. The K-means basic and extended models are presented in Fig. 5. Simulation results revealed that the proposed K-means enabled machine-learning framework for mm-wave-NOMA systems outperforms mmWave-OMA systems. The authors in [24] also focus on application of AI/ML into user clustering for mmWaves, as it has recently been confirmed that machine learning provides a better solution to fast data clustering due to flexibility with more information features of higher dimensionality. However, in [41] it was pointed out that the requirement of a perfect channel state information at the transmitter and high computational complexity at the receiver may be detrimental in NOMA operation. The use of deep learning (DL) techniques is proposed as a potential solution to deal with these challenges. Such DL-based solution can be incorporated in NOMA to detect a completely unknown and sharply changing channel condition. In general, supervised learning is extensively utilized to estimate the CSI, support spectrum sensing, localization, and throughput prediction. Unsupervised learning is mostly used in user clustering and congestion control. While the DL can be extremely useful for complex data processing and to acquire perfect CSI, but to date, DL is mostly proposed for use in resource management. The summary of learning techniques proposed for use with NOMA is presented in Table 1.

In [30], NOMA-based energy-efficient task scheduling among MEC servers for delay-constraint mobile applications is considered in highly-dynamic vehicular edge computing networks. The various movement patterns of vehicles lead to unbalanced offloading requirements and different load pressure for MEC servers. Self-Imitation Learning (SIL)-based Deep Reinforcement Learning (DRL) has emerged, as a promising machine learning tech-

Table 1 Example ML techniques used in combination with NOMA PHY (source [41])

Reference	Type of learning	Deep learning model	Application
11	Supervised	Recurrent neural network	Rapid and optimized resource allocation
14, 13	Supervised	LSTM	Channel estimation
15	Reinforcement	Deep Q-network	subcarrier assignment and power allocation
16	Reinforcement	Attention-based neural network	Joint power allocation and channel assignment
17	Supervised and	Deep belief network	Power optimization assignment
18	Supervised	Deep neural network	Automatic realization of the CSI
21	Unsupervised	Deep neural network	Signal constellation design
25	Supervised	Deep neural network	Reliability improvement of grant-free access
26	Supervised learning	Deep neural network	Optimization for energy-efficient scheduling
27	Reinforcement	Deep neural network [73] based on Q-learning [71]	Power allocation optimization

nique, to break through obstacles in various research fields, especially in time-varying networks. SIL is an off-policy algorithm that aims to reproduce the beneficial experience, i.e., imitating good actions in the replay buffer [69]. Traditionally described in the literature task scheduling schemes for MEC servers focus on users with relatively low speeds, thus the influence of the movement of users can be neglected while users in vehicular networks may suddenly move into the area associated with another MEC server and lose the connection to the current one when the computation results are back, resulting in packet losses and poor quality of user experience. MEC servers are always powered by battery-enabled Road side unit (RSU). Therefore, energy-efficient scheduling policies are necessary to prolong their available service time. However, MEC servers may consume lots of energy for the computations offloaded. In [30], each MEC server can be modeled as an M/G/n/k queueing system, and the complete system model is presented in Fig. 6. The Markov decission process (MDP) based framework is designed to minimize the total energy consumption for tasks processed by MEC servers, while satisfying the task's latency requirement. NOMA-based energy-efficient task scheduling is a promising way to control the total energy consumption of MEC servers and brings various benefits for resource-constraint applications. The use of NOMA for improving random access (RA) procedure in the future ultra-dense networks is considered by authors in [26] as the current RA protocols perform poorly in ultra-dense networks. Machine learning based approach to efficiently accommodate the MTC devices in RA slots is considered for NOMA with Q-learning. Even though Rel.16 and Rel.17 of 3GPP specifications already include some enhancements (2-step Random Access Channel (RACH), reliability improvements, power saving techniques, support for new use-cases including IoT) there is a room for improvements. The benefit of Q-learning is its model-free operation and distributed manner. In [40], NOMA and Q -Learning are utilized in order to maximize energy efficiency in short packet communications. In the Q-learning algorithm, agents are the MTC devices, the environment is the network, while the state-action pair is the combination of the transmit power and the time-slot, with every device having its own Q-table. The simplest way to implement the Q-Learning algorithm is to apply a greedy policy, this way the device always chooses the time-slot with the highest Q -value. The authors in [67] focus on maximizing the computation rate of an MEC system and investigate the computation offloading and subcarrier allocation problem in Multi-carrier (MC) NOMA based MEC systems and address this using Deep Reinforcement Learning for Online Computation Offloading (DRLOCO-MNM) algorithm. The proposed solution helps UEs to decide between local and remote computation modes, and assigns the appropriate subcarrier to the UEs in the case of remote computation mode. Since legacy reinforcement learning becomes less effective when dealing with actual complicated problems of high dimensional-state and action spaces the DRL is composed of two components: offline Deep neural network (DNN) and online Q-leaning to overcome this challenge [53].

Simulation results showed that the proposed algorithm can achieve near-optimal results and achieve significantly higher computation rates compared to the OMA (TDMA) based algorithm. Multiple machine learning-enabled

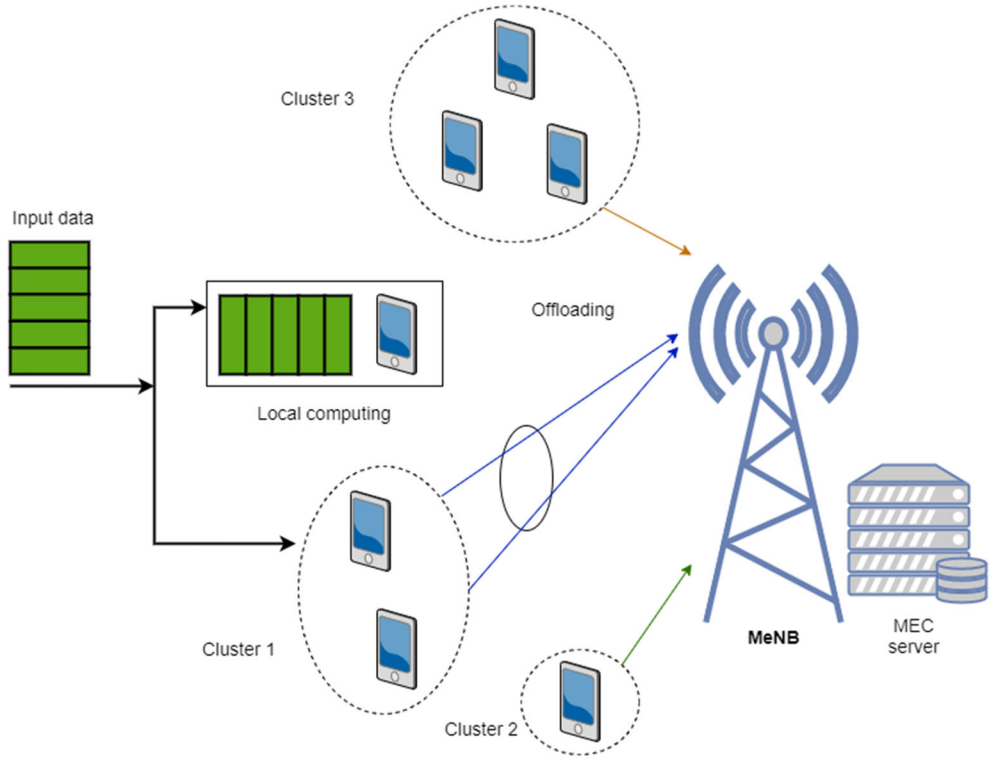

Fig. 6 Schematic of the task offloading for MEC NOMA

solutions are adopted to tackle the challenges of complex sensing model in cooperative spectrum sensing for non-orthogonal multiple access transmission mechanism [81]. Spectrum sensing and dynamic access in unlicensed spectrum are important techniques in the cognitive radio access, where users' terminal activity (relevant to sensing and decision making) is required as the BS does not assign transmission. The Cooperative spectrum sensing (CSS) has been introduced to mitigate multipath fading and shadowing. In this approach one node is nominated as the fusion center that collects radio environment parameters of all SUs in the cluster to identify behavior of PUs and the channel state and provides an opportunity to apply AI (Gaussian mixture models (GMM), Support vector machines (SVM), k-nearest neighbors (KNN)) for signals classification. Good performance improvement was demonstrated in terms of average training time, sample classification delay as well as receiver operating characteristic curve so far. The work presented in [64] also checked possibilities of grouping SU users before training the model with SVM to reduce cooperative overhead and boost detection performance. In the case of CSS for NOMA the SU will be transmitting when either near or far NOMA users are idle. As an alternative to AI/ML, in [8], the authors address the problem of cell-edge users and the efficiency of their transmission using game-theory. The Joint-transmission COMP (JT-COMP) is used in combination with NOMA and coalition game theory. Coalition game is developed to create most-beneficial clusters for the best overall performance. In the game, edge-user throughput is a utility function (pay-off) whereas the cost is the regular users "throughput loss". The results that were obtained show 2-3-fold improvement of edge-user with minimal cost to the other users.

4.2.3 Deep-Learning for NOMA Resource Allocation

The following section introduces deep-learning in various aspects of NOMA's design, operation and optimization, such as:

- User clustering (user pairing, cell-free, cluster identification, power control, grant-free mode)

- Task scheduling (joint with computational resource and power allocation; task offloading, cooperative offloading)
- Resource management (maximize energy efficiency under QoS/interference and power limitation, subchannel and power allocation, outage capacity maximization, caching)
- Heterogeneous requirements (diverse service requirements, beamforming, multiple-access selection, power control)
- NOMA improvements (decoding order, multi-user detection, CSI prediction, multi-access signatures optimization).

The authors in [60] proposed a joint user pairing and association algorithm for multicell NOMA using a DL-based approach. In multi-cell, a paired UEs have a freedom to connect to one of base stations. Therefore, user-association policy is required to assure optimal base station to serve the UE. The network structure called pointer network (PtrNet) is selected due to its non-iterative nature and low complexity. It is shown that for the joint user pairing and association problem for NOMA the PtrNet is only 2% away from optimal solution using exhaustive search. In [92] the authors formulated a joint task scheduling, computational resource allocation, and power allocation problem with an objective to minimize the sum cost (i.e., delay and energy consumptions for all task nodes) realizing energy-delay trade-off. It is challenging to obtain an optimal policy for such a combinatorial optimization problem, which was the reason why the online-learning was considered. The architectural assumption made by the authors was that devices in Fog network were delivering their task to a "helper node" that had some abundant capabilities to process these tasks and as experiencing spectrum under-utilization (with OMA transmissions). This work assumed deep Q-learning based algorithm to tackle the issue of random variations of system states. The NOMA combined with Dual connectivity (DC) feature of 3GPP (release 12) was proposed in [55] to prevent task offloading failures. The main goal was to reduce delays and save energy by being able to utilize two access points and two MEC servers by a single UE—especially the mobile one. By accessing two AP (e.g. macro and small cells) in parallel, the throughput can be improved. The above mentioned reference investigates DC- and NOMA-enabled computing offloading with the aim of minimizing the total energy consumption under the constraints of transmission power and time duration. NOMA is used to upload the task for offloading as well as to download the result, while the DC is used to increase robustness against mobility. The solution proposed is based on intelligent offloading that integrates BranchyNet [87] into deep residual network (deep ResNet) [42]. The optimization objective is to minimize the weighted sum-energy consumption in all four stages (uploading, computing, downloading, local processing). The deep learning module uses deep ResNet to learn a mapping relationship from the channel gain to resource allocation. The authors indicated that vanilla ResNet cannot be applied due to Mixed Integer Nonlinear Programming (MINLP) type of resource allocation problem—as it is problematic to characterize all the features with single output of ResNet. Compared with the single-connectivity approach, dual connectivity can significantly reduce the task drop ratio (when tasks cannot be executed on time) by offloading more computing tasks to remote MEC hosts for execution, thus reducing their computing burden and reducing the weighted sum of the energy consumption of the system, leading to improving task drop performance. The authors showed that intelligent offloading scheme dramatically reduces the time complexity via the inference property of deep learning. Performance features of various edge caching frameworks based on deep learning, deep reinforcement learning, and federated learning algorithms is studied by authors of [66]. It is considered from the perspective of delivering the URLLC services in 5G networks. As it is highlighted in [13], in-network caching is an effective technique for reducing content delivery latency and to improve content availability in IoT systems. The authors highlighted that NOMA among other techniques ("configurable subcarrier spacing and short transmission time interval (sTTI), grant-free access, edge computing, caching, dynamic multiplexing, latency sensitive scheduling schemes, network coding, network slicing, on-device machine learning (ML), and AI can be used to assure low-latency for URLLC. Full section about slicing is devoted to various flavors of NOMA. It was pointed out that the benefits of using DL over ML approach for IoT "include a better performance with large data scale because many IoT applications generate large amounts of data for processing. Also, DL takes much less time to inference information than traditional ML methods." The authors of [10] proposed the usage of NOMA in cell-free scenario combined with deep-double Q learning (DQL) for cluster identification and

deep deterministic policy gradient (DDPG) for uplink beamforming design. The input to both networks is the CSI matrix H (of channel gains). NOMA is used to compensate performance loss due to clustering access points—SIC based signal detection for uplink access is assumed. The upper bound for cell-free architecture outage performance with dynamic clustering is found in [10]. However, the authors in [56] argued that DRL approaches suffer from slow convergence, lack of robustness and unstable performance in changing environments. They proposed a multi-agent DRL framework to achieve long-term performance for cooperative computation offloading, in which a scatter network is adopted to improve its stability (in changing environments) and league learning is introduced for agents to explore the environment collaboratively for fast convergence and robustness. Here NOMA is used in the process of collaborative computation offloading of IoT networks and the authors formulated the joint problem as MDP. Another application of deep-learning is presented in [110], for radio resource management, where authors are maximizing the Energy efficiency (EE) of the system under the constraints of quality of service (QoS), interference limitation, and power limitation. Specifically, user association is solved through the Lagrange dual decomposition method, while semi-supervised learning and deep neural network (DNN) are used for the subchannel and power allocation, respectively, in NOMA networks. It is reflected that DNN advantage is the relatively short training time, so it is very suitable for real-time learning. In[59], the entire two-user NOMA system consisting of base-station, relay and users is redefined as a novel hybrid-cascaded DNN architecture including nine trainable DNN modules. In this way, the whole system can be optimized in a holistic manner. The whole DNN solution learns the mapping between base-station inputs and user-outputs to combat channel fading and noise. Offline-training and online-deploying mode in DL is used, which means there is no need for retraining. It is worth noting that the proposed NOMA setting using DNN significantly reduced the computational complexity especially for long codewords and decoding iterations. Authors in [89] utilize deep neural network and supervised learning to an OFDMA subcarrier assignment and NOMA user grouping problem in downlink video communication systems. It is emphasized that after training the computation time of deep-learning is 5–11 times lower than non-deep learning approaches. As shown by the authors, typically the use of deep learning considers for learning (aggregating the 12 prior-art items):

- inputs: channel (coefficients, for pairs of users in NOMA), SINR, demand from users
- outputs: power control, channel access time, resource allocation requirement

The cross-layer approach to resource allocation is assumed between application and physical layer in order to align the peak-SNR (PSNR) of video of k-users, with the channel quality at the PHY layer. The video distortion is minimized, expressed by the mean square error (MSE). In [106], the authors investigated joint subcarrier assignment and power allocation problem in an uplink multi-user NOMA system to maximize the energy efficiency while ensuring the required quality-of-service (QoS) for all users separately. The two-step deep reinforcement learning (DRL) based algorithm is proposed to solve such non-convex and dynamic optimization problem. The input is channel condition, and the deep-q network (DQN) produces subcarrier assignment policy, which in turn is processed by the deep deterministic policy gradient (DDPG) network to dynamically output transmit power of users. The DQN is not suitable for the power allocation due to its discrete actions which brings quantization error for continuous action tasks (like e.g. power allocation). In [90], the topic of outage capacity maximization in a 5G URLLC scenario is considered for downlink NOMA-video. Subcarrier reallocation (from user with high Quality of Experience (QoE) to user with low QoE) according to outage capacity (percentage of satisfied users whose PSNR is higher than threshold) maximization and not just ergodic capacity (video distortion) is considered where candidate users are assigned subcarriers. The resource management schemes based on deep-learning (DNN) and their non-deep learning counterparts are considered in the cross-layer architecture. Interestingly, the deep-learning versions perform worse than their non-deep learning versions, but their execution time is lower. The non-deep learning proposal is so called “Scheme B”, that is outage-capacity-based proposed cross-layer resource allocation for the NOMA-OFDMA video communication system, and its modification called “Scheme C”—where users with highest video PSNR are providing sub-carriers to lower PSNR users. Hence, this approach is similar to congestion control with fairness assurance, as a result the outage capacity is increased. In [111] the grant-free NOMA in uplink is considered. To reduce collisions in the frequency domain (i.e. when too many users selecting low received power level) and the computational complexity of deep reinforced learning (DRL), subchannel and device clustering are first considered.

The cluster of devices competes for a cluster of subchannels following grant-free NOMA. The clusters define independent grant-free NOMA sub-systems. This is a crucial problem to address in uncoordinated environments. Furthermore, discrete uplink power control is proposed to reduce intra-cluster collisions. Then, the long-term cluster throughput maximization problem is formulated as a partially observable Markov decision process (POMDP). DRL-based grant-free NOMA algorithm is proposed to learn about the network contention status and output subchannel and received power-level selection with less collisions. In [57], the authors presented some research results on a novel multi-dimensional intelligent multiple access (MD-IMA) scheme, which can be considered as a convergence of MIMO, NOMA, and OMA, designed to meet diverse service requirements of heterogeneous equipment while considering their different constraints. First K user terminals (UE) are clustered based on location information into C clusters. Each cluster is allocated beamforming matrix. In the next step K users are allocated to M channels by using MD-IMA. In this scheme each UE can adaptively select multiple-access scheme: OMA, PD-NOMA or SD-NOMA according to the channel information and service requirements. Eventually, power allocation to all users is performed with deep RL-learning. Allocations are done per k-th user, on m-th subcarrier and in the c-th cluster for a given timeslot, as network dynamics and time variation pattern are difficult to be modeled in real-time using the conventional techniques while maintaining accuracy and tractability. Moreover managing the multi-dimensional resource is extremely complicated. That is why the authors considered deep learning approach methods that can extract the dynamics (e.g. channel quality changes) of a network and make decisions based on historical observations automatically. For example, by introducing multiple hidden layers, deep neural network (DNN) is capable of precisely predicting performances of a system, such as traffic volume and power consumption [83]. The authors defined an aggregated KPI called I-QoSE which combines power consumption, delay, and throughput as well as cost, with the different weights. Such weights are adjusted based on predicting network situation and QoS requirements based on historical data. The MD-IMA objective is to allocate resources in a way that maximizes the I-QoSE under constraints of transmit power and each user's QoS requirements "the Long-short term memory (LTSM) performs large timescale prediction of traffic volume (in terms of sum rate requirement), and then adjusts the weight factors of the I-QoSE metric to follow the changes in the service requirements dynamically. For example, the weight of the sum-rate should be increased if the network traffic is predicted to rise". The scheme is evaluated for time-division-duplex (TDD) downlink transmission in a single cell downlink scenario. Performance of the MD-IMA scheme has been evaluated with the real-world cellular traffic datasets (i.e. Telecom Italia data sets [15]). Two problems are defined—network situation prediction for all the grids which constitute the cell and real-time resource allocation in multiple-access domains. The DRL techniques are also applied in [82] to deal with NOMA decoding order and user all-access strategy in case of imperfect SIC. Dynamic access and dynamic decoding order in NOMA system is studied and the authors proposed an intelligent hierarchical DRL-based user access and decoding order selection (HDRUD) algorithm. Topics of user-detection for uplink grant-free NOMA are also studied in [34]. The off-line-trained LTSM-based network is used for multi-user detection and channel estimation (DeepMuD). Multi-user detection is based on pilot signals and does not require perfect CSI. The DeepMuD has been trained offline and implemented as an online detector in grant-free IoT NOMA networks. The authors showed that error performance of DeepMuD solution has been improved significantly, with 10 dB less power consumption than conventional detectors. Complexity of the proposed solution is independent of a modulation order or a number of IoT devices (which is the case of the legacy SIC algorithm). In [102], the authors considered a joint optimization of NOMA and identify the key NOMA components as multi-access signatures (MAS) and Multi-user detection (MUD). Examples of signatures considered are: power-level, and bit-to-symbol sequence mapping. On the other hand, MUD techniques are mostly based graph model-based methods such as interference cancellation (ICT) and message passing. The MAS determines the performance upper bound, and MUD decides how close that bound can be approached (Figs. 7, 8).

According to [102], it is not optimal to design NOMA transceivers in isolation (which is the current trend in NOMA research), they promote the concept of unified framework for simultaneous design, and end-to-end optimization of NOMA transceivers. The authors assumed the use of deep learning (namely DNN) to approximate optimal NOMA transceivers. Multi-tasking DNN treating non-orthogonal signal transmissions as multiple distinctive but interrelated tasks is utilized. A novel multi-task balancing loss function was then proposed to ensure fairness among

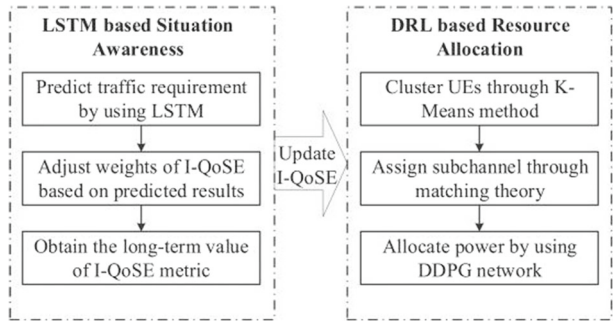

Fig. 7 MD-IMA deep learning framework (source [57])

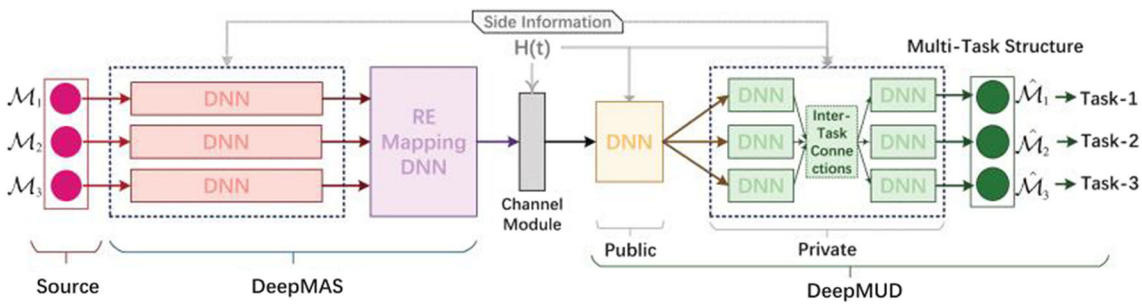

Fig. 8 Holistic approach to NOMA design and optimization with DL (source [102])

various tasks and to avoid The "RE mapping DNN" is there to allow generalizing both sparse and dense NOMA schemes. The results of comparing DeepNOMA with other examples of non-orthogonal techniques i.e. Multi User Shared Access (MUSA), WBE and SCMA NOMA—shows that the former outperforms the latter schemes. It is highlighted that DeepNOMA was also a unified framework for linear or non-linear, power or code domain, and grant-based or grant-free NOMA in the other paper of the same authors [101]. An interesting survey on PD-NOMA, that overviews of multiple aspects of NOMA, including the application of ML/AI, cell-free and mMIMO has been presented in [64]. Interesting aspect of the mentioned study is that the authors provided lessons learned for multiple NOMA design aspects and related techniques. This can be considered as vaulable complementary material to the current paper.

4.2.4 Overview of Recent R&D Projects Dealing with NOMA

In the last few years, there have been several R&D project initiatives considering NOMA dedicated to 5G like EC-funded R&D projects: 5GCity or One5G that were also dealing with these multi-access techniques. In addition to that there are projects related to NOMA in other countries like: Japan, Korea, US. In the following, some of the key research efforts on NOMA that can be attributed to those international research teams are highlighted. There are also projects investigating the use of NOMA that are in progress or just ended, the [68] is one of them. Some of the researchers working in this area were trying to prove that NOMA has a better potential than OMA. In some of their investigations, they came to a conclusion that that OMA is better than NOMA because it is easier to develop with corresponding lower computational complexity cost. Despite this, NOMA has higher potential and in the future, it should deliver more powerful wireless systems. The researchers in [39] have similar conclusions that NOMA in the future might be better that OFDMA that has a problem of inability for frequency reuse within a single cell. In NOMA there is a SIC technique, but the authors in [39] proposed their own technique—T-SIC that can benefit from frequency reuse as well as the near-far effect. The proposed T-SIC technique with iterative signal processing provides significant bit error rate (BER) performance improvement. NOMA with the proposed T-SIC technique significantly outperforms OFDMA due to frequency reuse. In [84], the authors investigated both downlink and uplink NOMA systems, and proposed a scheme based on signal spreading. Massive simulations and

tests were performed. Theoretically, according to their findings, their proposed scheme can reach the capacity of the multi-user MIMO channel, and outperforms the state-of-the-art alternatives in terms of maximum achievable rate over discrete constellations. In the paper, the authors have used a highly optimized probabilistic algorithm called Rubik-PTST. It was proved through simulation that their scheme, called MC-NOMA, outperforms well-established NOMA schemes such as SCMA or PDMA. The authors in [9] made both mathematical analysis and simulation with IoT devices using Power-Domain NOMA (PD-NOMA). Especially, they focused on the sum-rate upper bound and the reliability in terms of target received power values, and power difference threshold for Rayleigh and Nakagami-m fading channels. The simulations were performed to satisfy the communication requirements of IoT-like services, which means a compromise between reliability and sum-rate upper bound. The authors in [6] were using NOMA with FPGA to secure a user-fairness and connectivity for two users in downlink communication. FPGA was used to both design the system and to evaluate its performance. The results obtained with the FPGA (i.e. BER over SNR) implementation of NOMA were compared with Monte-Carlo simulations using MATLAB. In the proposed version of FPGA the utilization of the device was about 8–15%, except for the number of the bonded IOBs, which was 98 percent. The authors indicated that a new upgraded hardware might be a solution to this problem, but it was not investigated. In [79], the authors have analyzed MIMO-NOMA systems based on Dynamic user pairing scheme (DUCGD). The DUCGD is implemented by calculating the channel gain of every prospective user and sorting it in descending order to pair ones with the highest difference in channel gains to function together. The numerical analysis showed that the increase of the number of users leads to an increase in the systems' gain and the total achieved sum rates. In [79], this improvement observed based on the combined use of more antennas and DUCGD. The team in [76] used NOMA with AF and DF protocols. The numerical results confirmed that in the incremental cooperative NOMA (ICN) system, remote user achieved better shutdown behavior than in the conventional cooperative NOMA (CCN) system. The nearby ICN system user was able to achieve better performance than in the case of the CCN system user in the low SNR region, and the remote user downtime performance for the DF-based ICN system was better than the AF-based ICN system when the system was operating in the NOMA cooperative transmission mode. Moreover, the capacity of the ICN system was higher than that of the CCN system in the low SNR region. Single cell NOMA downlink network was considered in [104]. The authors proposed a virtual full-duplex cooperative NOMA structure where the BS sends superimposed signals for two UEs in each transmission phase while the selected RS sends signals to two UE simultaneously. This framework outperforms the conventional half-duplex NOMA cooperative platform, which provides the best performance in terms of shutdown probability as well as diversity-multiplexing tradeoff (DMT). To conclude the research highlighted in the this section, the AI/ML/DL has the capability of bringing significant changes to both the design the future 6G networks and the way NOMA techniques can be further enhanced. However, in the description so far one key ingredient behind all the opportunities has been not mentioned at all, which is the energy efficiency of NOMA itself or the energy consumption improvements that can be achieved with the help of NOMA. In the next section, we present an overview of such techniques in combination with NOMA (Table 2).

5 Sustainablity Aspects of NOMA

5.1 Energy Efficiency in NOMA

Irrespectively of the cellular (or cell-free) RAN network layout and operation, and possible use of deep learning to boost design of NOMA transceivers, the issue of energy efficiency of NOMA is an important one. With a tendency of introducing software defined mobile networks, the challenge starts to grow, due to a trend in shifting majority of processing (i.e. baseband as well as part of radio units) onto the servers in the cloud or edge. Networks implementing NOMA needs to consider analysis of computing resources together with radio resources. According to the authors in [16] there are multiple aspects of NOMA that contribute to energy efficiency of NOMA techniques. The picture provided by the authors nicely presents this view and it is shown in Fig. 9.

Table 2 Energy efficiency papers overview

Paper	Energy efficiency	Algorithm/Aproach	Complexity result
[5]	EE(NOMA)>EE(OMA) EE(proposed)>EE(NOMA)	User association Subcarier assignment	$\theta(NM(M+T_t))\theta(MK(MK+NT_t^{'}))$
[32]	Proposed solution needs only 20–30% less energy consumed compared to random clustering and power control	User clastering and computation resource allocaction transmit power control	Low complexity
[80]	Only the power level are considered P_h, P_l for power control. No other mentions.	Grant-free access methods based on Q-learning is proposed	N/A
[11]	Basic OMA is always less energy efficient than NOMA.But optimized-OMA can sometimes be more efficient than NOMA.	OMA scheme with optimal time and power allocation (O-OMA) H-NOMA scheme with optimal user selection.	N/A
[71,91]	Low complexity solution is robust against the increase in number of IoT devices (NOMA clients).	Computation Resource Allocation and SIC Ordering Algorithm.	Optimal solution Low-complexity solution
[14]	In DL: Algorithms applied to OMA and NOMA bring similar average cluster energy In UL: OMA case consumes more energy for the same algorithms as in NOMA average transmit energy per user in the UL under OMA scheme is higher than the NOMA schemes, due to orthogonal transmission.	SWIPT enabled NOMA cell with join sub carrier assignment, time-switching and power allocation (J-SA-TS-PA)	J-TS-PA-polynomial complexity $\theta(2\|\ell_m\|^2 T_{m,k}(\in))$ Subcarier allocation- polynomial complexity $\theta(KM^2)$ Joint optimization has higher complexity
[103]	Focus on latency and task completion probability	Closed-form mathematical expressions of the successful computation probability in MEC offloading NOMA system	N/A
[70]	Energy efficiency of task offloading to MEC with NOMA is always better than reference FDMA system with bandwidth equally divided between K-users. Also as number of users growths the NOMA is always more efficient.	With the obtained optimal power control solutions, the task offloading partitions and time allocation are obtained by the successive convex approximation algorithm	N/A
[44]	Exhaustive search max-min computational efficiency is lower than for NOMA schemes proposed. Computing efficiency with proposed NOMA scheme is better than OMA and local computing on the UE.	Heuristic search algorithm to obtain optimal policy, including uplink transmit power allocation policy and local computing resource allocation policy	$\Theta[(\frac{K^2+13K}{2})(HMS+NI)]$
[96]	There is no detailed energy efficiency analysis provided.	Design efficient iterative algorithm by jointly designing the secrecy rate, local computing bits, and power allocation	N/A
[51]	Energy consumption of NOMA offloading decreases by increasing the available number of frequency RBs; second, the energy consumption improves by increasing number of users sharing PRBs	Minimizes the energy consumption of MEC users via optimizing the user clustering, computing and communication resource allocation, and transmit power	N/A
[20]	Proposed NOMA scheme always achieves lower energy consumption than FDMA (OMA) scheme.	Non-convex problem of energy consumption minimization is transformed into a task and power allocation problem, and a subchannel allocation problem, plus the delay constraints are considered.	Proposed algorithms complexity is $O(n^3)$ for both algorithms with guaranteed convergence. Where n depicts number of UEs in the center of the cell

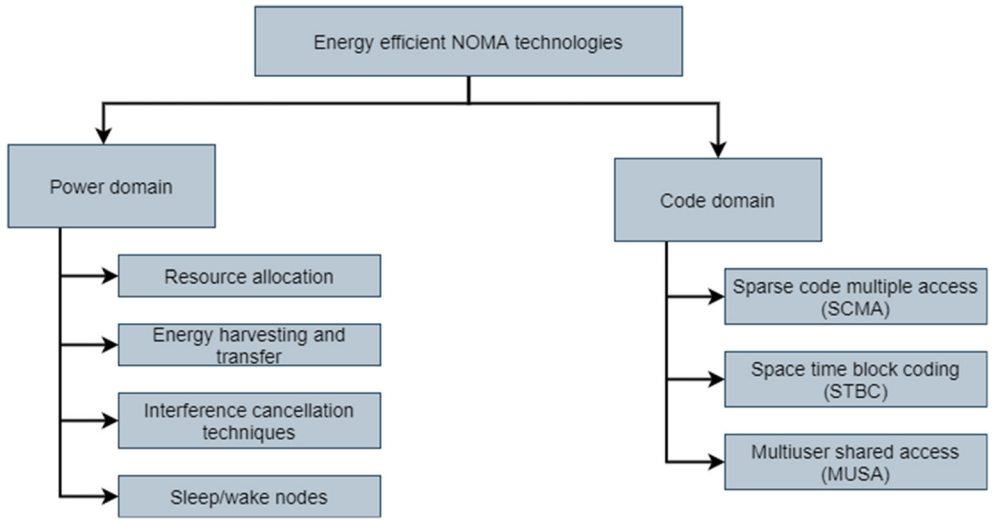

Fig. 9 NOMA technologies for energy efficiency

Based on the Fig. 9 classification, the authors the in [16] considered each of the listed technologies for NOMA. From the architectural perspective it is important to remind after the AIOTI alliance, that "[...] Edge computing functions can be hosted on a micro modular edge device, a gateway, micro server or other processing units where the edge server operates as a decentralized processing unit or an extension of the centralized cloud in the context of an overall application or workflow that is managed in the edge or cloud and executed at the edge" [7]. The authors in [5] addressed energy efficiency maximization problem for multi-cell, multi-carrier NOMA in downlink. The original optimization (Mixed integer non-linear NP –hard) problem is decomposed into two sub-problems: (a) user and BS association and sub-channel assignment using binary whale optimization (BWOA) [45] and (b) successive pseudo-convex approximation (SPCA) [99] is used to allocate power. The resulting energy efficiency (EE) for NOMA is higher than what is obtainable in the case of OMA. In [32], the authors addressed the energy-efficient user clustering and resource management for NOMA based MEC systems. In such an approach, optimal amount of physical resource block (RB) is allocated to exchange of data that needs computation for optimal clusters of UEs. The problem is formulated as mixed binary-integer and continuous variables programming optimization problem. In this way, the energy efficiency problem degrades to transmit power control optimization. The solution is derived using low complexity heuristic—Firework Algorithm [32]. In [80], the uplink NOMA is considered in federated learning model aggregation. The objective was to find an optimal user scheduling and power allocation scheme to maximize the weighted sum rate by transforming original problem into the maximum-weight independent set problem that can be solved with graph theory. The system model assumes K out of M user edge devices that can upload the learned local parameters into parameter server, and NOMA was used to enable multiple users to upload simultaneously. Novel OMA and hybrid NOMA (H-NOMA) schemes for MEC Offloading are introduced in [11]. These were both OMA scheme with optimal time and power allocation (O-OMA), and a H-NOMA scheme with optimal user selection (H-NOMA-S) scheme. To guarantee the fairness of task execution latency across NB-IoT devices in the uplink, computation resources of MEC units have to be fairly allocated to tasks from the IoT devices according to the task size [71]. For these reasons, the authors investigated the joint optimization of SIC ordering (ascending/descending order of channel gains, ascending order of task size) and computation resource allocation of MEC. Specifically, a combinatorial optimization problem was formulated with the objective to minimize the maximum task execution latency required per task bit across NB-IoT devices under the limitation of computation resource. Hence, three different algorithms were proposed to solve the optimization. It is also shown that for an increased number of NB-IoT devices, the use of NOMA for task offloading reduces execution time as compared to FDMA and TDMA by 60-70%. Very recent study about the minimization of latency of task offloading with the use

of Deep Deterministic Policy Gradient (DDPG) in IoT mobile systems with NOMA is presented in [21], where the latency reduction by 20% is achieved compared to reference algorithm. The authors in [14] considered network sum-rate maximization in Simultaneous wireless power and energy transfer (SWIPT) enabled energy-harvesting (EH) clustered NOMA networks in the downlink and uplink directions. Specifically, the aim was to utilize the harvested energy at the base-station to maximize the network sum-rate in both link directions via joint subcarrier-assignment, time-switching, and power allocation. The use of MEC in future applications for 5G networks will be necessary to ensure low latency and high computational requirements. The cooperation between MEC and local computing resources of users allows to shift the focus in terms of computing, but the efficiency and availability of effective communication techniques at the edge of the network are a significant limitation [58,86]. The advantage of NOMA is, among others improvement of spectral efficiency, improvement of communication possibilities and reduction of delays. Moreover, the main advantage of NOMA in the grid edge computation scenario is the improvement of energy efficiency, latency, and increased probability of completing the computations on time [103]. The use of MEC and Fog nodes is the base scenario here. In addition, the use of supplementary energy sources such as SWIPT mechanism (simultaneous data and energy transfer) and, for example, alternative energy sources in a form of renewable energy, constitute an interesting alternative considered by researchers. The authors in [103] considered a new hybrid computation balancing scheme (more complex task offloading) that can function as a partial computation offload, full local computation, or a complete computation transfer to MEC. They also considered computing capabilities of the MEC server. The paper includes derivation of the probability of completing the calculations within a set delay budget and presents the optimal calculation parameters, i.e. the time when the calculations must be completed, power allocation and coefficients controlling the distribution of the calculation availability for two stationary users. It has been shown that the optimal availability rates for offload techniques are determined by taking into account the difference in computing capabilities between the end user (Edge) and the MEC server, and that user locations have a large impact on the probability of successful completion of the computations. Three schemes for selecting users from two groups are presented in the paper as follows users located near the base station and users at the cell edge. The selection of users and transmission parameters is made by the MEC server. The control of the SIC decoding order also plays an important role, as at random times the user at cell edge may have better parameters than the user near base station due to small scale fading. As indicated by the authors in [91], the order of signal decoding in the SIC mechanism becomes a bottleneck limiting the improvement of uplink performance, which is the dominant type of traffic in NB-IoT communication. In addition, to guarantee a fair level of computation latency on NB-IoT devices, MEC computing resources must be fairly allocated to tasks from IoT devices according to the size of the task. For these reasons, the authors investigated a joint optimization of SIC decoding order and computing resource allocation. In particular, the problem of combinatorial optimization was formulated to minimize the maximum delay of the computation per one task bit on NB-IoT devices with limited computing resources. The proposed heuristic algorithm for SIC decoding, the complexity of which is O (N2) where N is the total number of NB-IOT devices in the network. As part of the literature review, the authors cited the work in [51], in which the NOMA-based optimization framework for 5G networks was proposed. The proposed optimization's aims was to minimize energy consumption by MEC users by optimizing methods for user grouping, allocation of computing and communication resources and the transmission power. Similarly, [70] tested MEC in conjunction with NOMA techniques for both submitting tasks and downloading results. Energy consumption is minimized by optimizing the transmitter power, allocating the transmission time, and configuring the calculation process offload. In particular, with the obtained optimal power control solutions, the MEC offload and time allocation are achieved by the convex approximation algorithm. The numerical results show that the proposed NOMA-based MEC offload scheme can significantly reduce power consumption of the system compared to traditional OMA schemes. On the other hand, as shown in [29], the assumption that a close user (with a high gain value) ends offloading calculations exactly during transmission by a distant user in the OMA mode, may lead to a low energy efficiency of the NOMA-MEC connection. The authors show that for energy efficiency it is better partially offloading the near user tasks during the delegation of tasks by the remote NOMA user and then handing over the remaining near user tasks at a dedicated transmission time. That way, it is possible to offload more data than with OMA. It has been shown that the increased number of NB-IoT devices using NOMA for offloading reduces execution times compared to FDMA and TDMA

by 60-70%. To ensure a compromise between delays and MEC energy consumption, a dynamic offloading scheme for computing tasks has been developed, assuming that mobile devices can accumulate energy from e.g. renewable energy sources. In [94] and [43], a MEC scenario was considered in which simultaneous wireless information and power transfer (SWIPT) was used in conjunction with MEC to improve user experience. In [32] the authors deal with energy efficient user grouping and resource management for NOMA based MEC systems. In such an approach, an optimal amount of radio resource block (RB) is allocated to the data exchange that requires computation for the optimal UE clusters. The problem was formulated as a combined programming optimization problem for binary and continuous variables. Thus, the problem of energy efficiency was transformed into an optimal transmission power control. The solution was obtained using a low complexity heuristic—i.e. the Fireworks algorithm. Combined optimization of power in the transmitter and local allocation of computing resources was considered in [44], where the self-adaptive harmony search (SAHS) algorithm [93] with low computational complexity was used to solve the optimization problem (non-convex). With the SAHS algorithm, optimal policies can be obtained in a similar way to the improvisation process in a group of musicians who synchronize in search of optimal harmony. The performed simulations compared offloading with OMA, NOMA, and local calculation on the mobile device. The results obtained confirmed that the highest energy efficiency when using partial offloading with the use of NOMA compared to the other schemes. The work in [96] shows how to consider the requirements for the minimum level of transmission protection against eavesdropping by an unauthorized user in the NOMA-MEC optimization problems. As shown in the analysis, to increase the level of security, the key is to properly control the power allocated to NOMA users for the uplink. The choice of the power level should be based on the allowed maximum power threshold, the total energy budget, and the level of interference to/from other users. It is based on such an analysis that the power levels of other users can be adjusted. It was also shown that it is impossible to meet the requirements for a high level of security and low delays in calculating tasks (offloading tasks)—increasing the level of security comes at the expense of the increased level of delays due to the need to introduce redundant information in transmission. The authors of [51] formulated the problem of minimizing the total energy of NOMA mobile users with a limitation on the total transmission time and the total calculation time. To reduce the computational complexity, the authors proposed a heuristic method for user grouping and radio resource block (RB) allocation. Thereafter the optimization processes, the problem of convex optimization was formulated to control the power level of user terminals separately for each NOMA cluster. It has been shown that as the number of radio resources increases and the number of users in the cluster increases, the total energy consumption on the part of mobile users decreases. In turn, the spectral efficiency increases with the increase in the number of users and the size of tasks to be sent to the MEC server. The model of energy consumption by mobile users (CUE) and NOMA (EUE) was presented in [20]. It was considered that (a) the NOMA access point (EAP) also consumes energy for the execution of computing tasks from NOMA users, and that (b) the computations of tasks sent to the MEC have their maximum execution times. The optimization problem is the partitioning of tasks to be performed, the allocation of sub-channels and the allocation of power of NOMA user transmitters (EUE) to minimize power consumption of the system. Optimization is implemented by game theory algorithms. The results in [20] are comparable to the results from [51]. In [79], the NOMA uplink is considered for aggregation purposes in the federated learning model. The goal is to find the optimal user data ranking schedule and a power allocation scheme to maximize the weighted sum of data—a problem that can be solved by graph theory. The system model assumes that K of M endpoints can send learned local parameters of the target model to a central server storing the model parameters using NOMA to enable a simultaneous uplink data transfer by multiple users. The work presented in [95] focuses on research on IoT routing algorithms in conjunction with NOMA. In the paper, the authors detailed analytical considerations for a situation in which the Fog node collects data from IoT users (in the uplink) and then sends this data to the appropriate users (recipients) in the downlink. Additionally, the authors compared two versions of NOMA with each other: NOMA-NC implemented in the FoG relay node and NOMA-NOMA being the reference to the base version. NOMA-NC is a typical NOMA uplink solution, while downlink messages to target users are encoded and multicast. The use of NOMA-NC allows for a 25% reduction in the complexity of detection in the receiver compared to a typical NOMA receiver. The summary of the key contributing papers of this section are listed in Table 4 1.

6 Conclusions

This paper presented an overview of some NOMA techniques and the recent developments related to this multiple-access scheme. The primary aim of the authors is document various most recent directions in NOMA (including virtual RAN, cell-free, ML-based solutions—and especially application of deep learning) that are very much oriented towards 6G wireless networks. This is with an emphasis on enhancements in NOMA design and optimizations towards B5G and 6G. The general observed trend is the rising interest of combining the NOMA and MEC for computation task offloading, while assuring high energy efficiency. As the key constituent of 6G, the security topics are presented in combination with NOMA techniques, where the physical-layer security can be addressed with multiple NOMA elements (cooperation, relaying, resource allocation). Still most important is that NOMA is considered as strong candidate for beyond-5G standards in 3GPP, though decisions have not been made with respect to this, up to the time of documenting this paper. Hence, NOMA (accompanied with the RSMA), seems to be in a phase of intense research, especially considering topics important in connection with 6G. There are multiple existing research challenges in NOMA, but, in the authors' opinion the most important are:

- NOMA performance and complexity: essentially, NOMA introduces a tradeoff between performance (sum rate) and complexity. Analyzing this tradeoff is a worthwhile research topic. [...] energy efficiency of cell free massive MIMO-NOMA could be quantified [76]. The effect of SIC processing delays, SIC decoding error propagation, unpredictable interference caused by grant-free access, and imperfect CSIs should be analyzed in further details [50]
- Realistic receiver design: realistic receiver needs to take into account practical issues such as active user detection for grant-free transmission, non-ideal channel estimation, time and frequency offset handling and receiver complexity The key issue for grant-free is that how to perform UE identifications [85]. Although spatial domain NOMA is quite effective in improving the spectral efficiency, the use of conventional pilots to acquire channel information causes severe pilot contamination [61]. Possible solutions include blind (pilot-free) data-driven methods [85], channel predictions using non-RF data [8] and the enhancement of pilot design. The use of multi-pilots along with the application of strategies similar to modern random access to decode them is an example of the latter [32].
- Grant-free schemes: while advanced NOMA has been well researched in configured grant-free schemes, in other grant-free approaches,the global power control, resource allocation and configuration cannot be accomplished efficiently, calling for advancements towards uncoordinated access policies. Open issues for grant-free access are: UL transmission detection, HARQ related procedures, RRC and L1 signaling, link adaptation, switching between orthogonal and NOMA [88]. This poses the further challenge of multi-user interference (MUI), for which the one-dimensional randomness of the power domain yielded thanks to the near-far effect may not be enough. Instead, higher-dimensional randomness including also e.g. code- and spatial-domains should be introduced [61]
- User clustering and prediction more complicated clustering algorithms that are robust to noises and outliers should be studied, incorporating the optimization of the number of user clusters into mmwave NOMA systems is capable of further improving the sum rate of the mmWave NOMA systems. Future work can also consider more sophisticated on-line and reinforcement learning procedures that update the partition according to the dynamic mmWave NOMA scenarios [25]. The analysis of customer behavior is one of the hardest challenges during the prediction of customer requests [36]
- Role of AI/ML in NOMA: the efficient design of the core neural networks is still a headache. Most of the current deep neural networks employed in NOMA provides a high computational cost. Therefore, reducing the computational complexity is a significant research issue for the future. Moreover, using a specific neural network, different researchers used different architecture. The number of layers and the quality and quantity of the training dataset is different for a different architecture, however, the optimized system is yet to be discovered [41]
- MEC offloading and orchestration in TN/NTN huge challenges on MEC systems have been brought by the highly unpredictable task flows arriving at MEC servers, requiring advanced optimization and estimation tech-

nologies to balance a trade-off between system performances and energy consumption [30] Recently NOMA and SIC has been indicated by the European Space Agency as one among multiple items to consider for research in the 5G/6G strategic appraoch. New research is required to address the challenges of orchestrating resources in a comprehensive ecosystem where IoT, edge/fog and cloud converge to form a computing continuum [7]. Finally, it is interesting to consider a multi-objective optimization problem in DC-assisted MEC offloading, i.e., minimization of power consumption with an interruption probability constraint and minimization of interruption probability with a power consumption constraint [55]

According to the authors of this paper the wide scale industrial trials also need to be launched by equipment providers in cooperation with mobile network operators.

References

1. 3rd Generation Partnership Project (3GPP), 2013, TR 36.888 Study on provision of low-cost Machine-Type Communications (MTC) User Equipments (UEs) based on LTE, Release 12 v(12.0.0)
2. 3rd Generation Partnership Project (3GPP), 2020, TR 21.916 "Summary of Rel-16 Work Items", Release 16 (v1.0.0)
3. 3rd Generation Partnership Project; Technical Specification Group Radio Access Network; Study on New Radio Access Technology Physical Layer Aspects (Release 14), 3GPP TR 38.802 (2017). https://www.3gpp.org/ftp//Specs/archive/38_series/38.802/38802-e20.zip
4. 3rd Generation Partnership Project; Technical Specification Group Radio Access Network; Study on Non-Orthogonal Multiple Access (NOMA) for NR (Release 16), 3GPP TR 38.812 (2018). https://www.3gpp.org/ftp//Specs/archive/38_series/38.812/38812-g00.zip
5. Adam, A.B.M., Wan, X., Wang, Z.: Energy efficiency maximization for multi-cell multi-carrier NOMA networks. Sensors **20**(22), 6642 (2020). https://doi.org/10.3390/s20226642
6. Ahmed, M.A., Mmahmmod, K.F., Azeez, M.M.: On the performance of non-orthogonal multiple access (NOMA) using FPGA. Int. J. Electr. Comput. Eng. **10**, 2151 (2020). https://doi.org/10.11591/ijece.v10i2.pp2151-2163
7. AIOTI Whitepaper: IoT and Edge Computing Convergence (2020). https://aioti.eu/wp-content/uploads/2020/10/IoT-and-Edge-Computing-Published.pdf
8. Akhtar, T. et al.: Efficient radio resource management with coalition games using NOMA in small cell networks. GLOBECOM 2020 - 2020 IEEE Global Communications Conference, pp. 1-6 (2020). https://doi.org/10.1109/GLOBECOM42002.2020.9348260
9. Aldababsa, M., Toka, M., Gökceli, S., Karabulut-Kurt, G., Kucur, O.: A tutorial on nonorthogonal multiple access for 5G and beyond. Wirel. Commun. Mobile Comput. **2018**, 9713450:1-9713450:24 (2018)
10. Al-Eryani, Y., Akrout, M., Hossain, E.: Multiple access in cell-free networks: outage performance, dynamic clustering, and deep reinforcement learning-based design. IEEE J. Sel. Areas Commun. **39**(4), 1028–1042 (2021). https://doi.org/10.1109/JSAC.2020.3018825
11. Altin, İ, Akar, M.: Novel OMA and hybrid NOMA schemes for MEC offloading. In: 2020 IEEE International Black Sea Conference on Communications and Networking (BlackSeaCom), 2020, pp. 1–5 (2020). https://doi.org/10.1109/BlackSeaCom48709.2020.9235017
12. Amjad, M., Musavian, L., Aïssa, S.: Link-layer rate of NOMA with finite blocklength for low-latency communications. In: IEEE 31st Annual International Symposium on Personal. Indoor and Mobile Radio Communications, London, UK, 2020, pp. 1–6 (2020). https://doi.org/10.1109/PIMRC48278.2020.9217106
13. Asmat, H., Din, I.U., Ullah, F., Talha, M., Khan, M., Guizani, M.: ELC: Edge linked caching for content updating in information-centric Internet of Things. Comput. Commun. **156**, 174–182 (2020)
14. Baidas, M.W., Alsusa, E., Shi, Y.: Resource allocation for SWIPT-enabled energy-harvesting downlink/uplink clustered NOMA networks. Comput. Netw. **182**, 107471 (2020)
15. Barlacchi, G., et al.: A multi-source dataset of urban life in the city of Milan and the Province of Trentino. Sci. Data **2**, 150055 (2015). https://doi.org/10.1038/sdata.2015.55
16. Basnayake, V., Jayakody, D., Sharma, V., Sharma, N., Muthuchidambaranathan, P., Mabed, H.: A new green prospective of non-orthogonal multiple access (NOMA) for 5G. Information **11**, 89 (2020)

17. Boviz, D.: Communications multi-utilisateurs dans les réseaux d'accès radio centralisés : architecture, coordination et optimisation. Autre. Université Paris-Saclay, (2017). Français. ffNNT : 2017SACLC035ff. fftel-01591285f
18. Boviz, D.: Communications multi-utilisateurs dans les réseaux d'accès radio centralisés: archtecture, coordination et optimisation. Autre. Université Paris-Saclay, (2017). Français. NNT: 2017SACLC035
19. Budhiraja, I., Tyagi, S., Tanwar, S., Kumar, N., Guizani, M.: Cross layer NOMA interference mitigation for femtocell users in 5G environment. IEEE Trans. Veh. Technol. **68**(5), 4721–4733 (2019). https://doi.org/10.1109/TVT.2019.2900922
20. Cao, X., Liu, C., Peng, M.: Energy-efficient mobile edge computing in NOMA-based wireless networks: a game theory approach. In: ICC 2020 - 2020 IEEE International Conference on Communications (ICC), pp. 1–6 (2020). https://doi.org/10.1109/ICC40277.2020.91490562
21. Cao, S., Chen, S., Chen, H., Zhang, H., Zhan, Z., Zhang, W.: HCOME: research on hybrid computation offloading strategy for MEC based on DDPG. Electronics **12**(3), 562 (2023). https://doi.org/10.3390/electronics12030562
22. Chang, Z., et al.: Energy-efficient and secure resource allocation for multiple-antenna NOMA with wireless power transfer. IEEE Trans. Green Commun. Netw. **2**(4), 1059–1071 (2018). https://doi.org/10.1109/TGCN.2018.2851603
23. Chen, X., Benjebbour, A., Li, A., Harada, A.: Multi-user proportional fair scheduling for uplink non-orthogonal multiple access (NOMA). In: Proceedings of the IEEE 79th Vehicular Technology Conference (VTC Spring), pp. 1–5, May 2014
24. Cui, J., Ding, Z., Fan, P.: The application of machine learning in mmWave-NOMA systems. In: 2018 IEEE 87th Vehicular Technology Conference (VTC Spring), pp. 1–6 (2018). https://doi.org/10.1109/VTCSpring.2018.8417523
25. Cui, J., Ding, Z., Fan, P., Al-Dhahir, N.: Unsupervised machine learning-based user clustering in millimeter-wave-NOMA systems. IEEE Trans. Wirel. Commun. **17**(11), 7425–7440 (2018). https://doi.org/10.1109/TWC.2018.2867180
26. da Silva, M.V., Souza, R.D., Alves, H., Abrão, T.: A NOMA-based Q-learning random access method for machine type communications. IEEE Wirel. Commun. Lett. **9**(10), 1720–1724 (2020). https://doi.org/10.1109/LWC.2020.3002691
27. Dai, L., Wang, B., Ding, Z., Wang, Z., Chen, S., Hanzo, L.: A survey of non-orthogonal multiple access for 5G. IEEE Commun. Surv. Tutor. **20**(3), 2294–2323 (2018). https://doi.org/10.1109/COMST.2018.2835558. (**thirdquarter**)
28. Ding, Z., Lei, X., Karagiannidis, G.K., Schober, R., Yuan, J., Bhargava, V.K.: A survey on non-orthogonal multiple access for 5G networks: research challenges and future trends. IEEE J. Sel. Areas Commun. **35**(10), 2181–2195 (2017). https://doi.org/10.1109/JSAC.2017.2725519
29. Ding, Z., Fan, P., Poor, H.V.: Impact of non-orthogonal multiple access on the offloading of mobile edge computing. IEEE Trans. Commun. **67**(1), 375–390 (2019). https://doi.org/10.1109/TCOMM.2018.2870894
30. Dong, P., Ning, Z., Ma, R., Wang, X., Hu, X., Hu, B.: NOMA-based energy-efficient task scheduling in vehicular edge computing networks: a self-imitation learning-based approach. China Commun. **17**(11), 1–11 (2020). https://doi.org/10.23919/JCC.2020.11.001
31. Driouech, S., Sabir, E., Ghogho, M., Amhoud, E.-M.: D2D mobile relaying meets NOMA–Part I: a biform game analysis. Sensors **21**, 702 (2021). https://doi.org/10.3390/s21030702
32. Du, J., Xue, N., Zhai, D., Cao, H., Feng, J., Lu, G.: Energy-efficient user clustering and resource management for NOMA based MEC systems. 1–6 (2020). https://doi.org/10.1109/GCWkshps50303.2020.9367499
33. Ebrahim, A., Celik, A., Alsusa, E., Eltawil, A.M.: NOMA, OMA mode selection and resource allocation for beyond 5G networks. In: IEEE 31st Annual International Symposium on Personal. Indoor and Mobile Radio Communications, 2020, pp. 1–6 (2020). https://doi.org/10.1109/PIMRC48278.2020.9217161
34. Emir, A., Kara, F., Kaya, H., Yanikomeroglu, H.: DeepMuD: multi-user detection for uplink grant-free NOMA IoT networks via deep learning. IEEE Wirel. Commun. Lett. (2021). https://doi.org/10.1109/LWC.2021.3060772
35. Endo, Y., Kishiyama, Y., Higuchi, K.: Uplink non-orthogonal access with MMSE-SIC in the presence of inter-cell interference. In: Proceedings of the International Symposium on Wireless Communication Systems (ISWCS), pp. 261–265, Aug. 2012
36. Fantacci, R., Picano, B.: When network slicing meets prospect theory: a service provider revenue maximization framework. IEEE Trans. Veh. Technol. **69**(3), 3179–3189 (2020)
37. Fayaz, M., Yi, W., Liu, Y., Nallanathan, A.: Transmit power pool design for grant-free NOMA-IoT networks via deep reinforcement learning (2020)
38. Gan, M., Jiao, J., Li, L., Wu, S., Zhang, Q.: Performance Analysis of Uplink Uncoordinated Code-Domain NOMA for SINs. In: 2018 10th International Conference on Wireless Communications and Signal Processing (WCSP), pp. 1–6 (2018). https://doi.org/10.1109/WCSP.2018.8555942
39. Haci, H., Zhu, H., Wang, J.: Performance of non-orthogonal multiple access with a novel asynchronous interference cancellation technique. IEEE Trans. Commun. **65**(3), 1319–1335 (2017). https://doi.org/10.1109/TCOMM.2016.2640307
40. Han, S., et al.: Energy-efficient short packet communications for uplink NOMA-based massive MTC networks. IEEE Trans. Veh. Technol. **68**(12), 12066–12078 (2019)
41. Hasan, M.K., Shahjalal, M., Islam, M.M., Alam, M.M., Ahmed, M.F., Jang, Y.M.: The role of deep learning in NOMA for 5G and beyond communications. In: International Conference on Artificial Intelligence in Information and Communication (ICAIIC), 2020, pp. 303–307 (2020). https://doi.org/10.1109/ICAIIC48513.2020.9065219
42. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. (CVPR), pp. 770–778 (2016)
43. Hu, X., Wong, K.-K., Yang, K.: Wireless powered cooperation-assisted mobile edge computing. IEEE Trans. Wirel. Commun. **17**(4), 2375–2388 (2018)

44. Huang, X., Zeng, S., Li, D., Zhang, P., Yan, S., Wang, X.: Fair computation efficiency scheduling in NOMA-aided mobile edge computing. IEEE Wirel. Commun. Lett. **9**(11), 1812–1816 (2020). https://doi.org/10.1109/LWC.2020.3001994
45. Hussien, A.G., Oliva, D., Houssein, E.H., Juan, A.A., Yu, X.: Binary whale optimization algorithm for dimensionality reduction. Mathematics **8**(10), 1821 (2020). https://doi.org/10.3390/math8101821
46. Islam, S.M.R., Ming, Z., Octavia, D., Kyung, K.: Non-orthogonal multiple access (NOMA): how it meets 5G and beyond (2019)
47. Islam, S.M.R., Avazov, N., Dobre, O.A., Kwak, K.-S.: Power-domain non-orthogonal multiple access (NOMA) in 5G systems: potentials and challenges. IEEE Commun. Surv. Tutor. **19**(2), 721–742 (2017). (**2nd Quart.**)
48. Islam, S.M.R., Avazov, N., Dobre, O.A., Kwak, K.: Power-domain non-orthogonal multiple access (NOMA) in 5G systems: potentials and challenges. IEEE Commun. Surv. Tutor. **19**(2), 721–742 (2017). https://doi.org/10.1109/COMST.2016.2621116. (**Secondquarter**)
49. Jaccard distance definition. https://www.statisticshowto.com/jaccard-index/
50. Kaneko, M., Randrianantenaina, I., Dahrouj, H., Elsawy, H., Alouini, M.-S.: On the opportunities and challenges of NOMA-based fog radio access networks: an overview. IEEE Access **8**, 205467–205476 (2020). https://doi.org/10.1109/ACCESS.2020.3037183
51. Kiani, A., Ansari, N.: Edge computing aware NOMA for 5G networks. IEEE Internet Things J. **5**(2), 1299–1306 (2018). https://doi.org/10.1109/JIOT.2018.2796542
52. Larsen, L.M.P., Checko, A., Christiansen, H.L.: A survey of the functional splits proposed for 5G mobile crosshaul networks. IEEE Commun. Surv. Tutor. **21**(1), 146–172 (2019). https://doi.org/10.1109/COMST.2018.2868805. (**Firstquarter**)
53. Li, H., Wei, T., Ren, A., Zhu, Q., Wang, Y.: Deep reinforcement learning: Framework applications and embedded implementations: Invited paper. Proc. IEEE/ACM Int. Conf. Comput.-Aided Design (ICCAD), pp. 847–854, Nov. 2017
54. Li, Y., Aruma Baduge, G.A.: NOMA-aided cell-free massive MIMO systems. IEEE Wirel. Commun. Lett. **7**(6), 950–953 (2018). https://doi.org/10.1109/LWC.2018.2841375
55. Li, C., Wang, H., Song, R.: Intelligent offloading for NOMA-assisted MEC via dual connectivity. IEEE Internet Things J. **8**(4), 2802–2813 (2021). https://doi.org/10.1109/JIOT.2020.3020542
56. Li, Z., Xu, M., Nie, J., Kang, J., Chen, W., Xie, S.: NOMA-enabled cooperative computation offloading for blockchain-empowered internet of things: a learning approach. IEEE Internet Things J. **8**(4), 2364–2378 (2021). https://doi.org/10.1109/JIOT.2020.3016644
57. Liu, Y., Wang, X., Mei, J., Boudreau, G., Abou-Zeid, H., Sediq, A.B.: Situation-aware resource allocation for multi-dimensional intelligent multiple access: a proactive deep learning framework. IEEE J. Sel. Areas Commun. **39**(1), 116–130 (2021). https://doi.org/10.1109/JSAC.2020.3036969
58. Liu, X., Jiang, S., Yi, W.: A novel deep reinforcement learning approach for task offloading in MEC systems. Appl. Sci. **12**(21), 11260 (2022). https://doi.org/10.3390/app122111260
59. Lu, Y., Cheng, P., Chen, Z., Mow, W.H., Li, Y., Vucetic, B.: Deep multi-task learning for cooperative NOMA: system design and principles. IEEE J. Sel. Areas Commun. **39**(1), 61–78 (2021). https://doi.org/10.1109/JSAC.2020.3036943
60. Ma, M., Wong, V.W.S.: Joint user pairing and association for multicell NOMA: a pointer network-based approach. In: 2020 IEEE International Conference on Communications Workshops (ICC Workshops), Dublin, Ireland, pp. 1–6 (2020). https://doi.org/10.1109/ICCWorkshops49005.2020.9145383
61. Mahmood, N.H., et al.: White paper on critical and massive machine type communication towards 6G. arXiv:2004.14146 (2020)
62. Mahmood, N.H., et al.: White paper on critical and massive machine type communication towards 6G. arXiv:2004.14146v2 (2020)
63. Mankar, P.D., Dhillon, H.S.: Downlink analysis of NOMA-enabled cellular networks with 3GPP-inspired user ranking. IEEE Trans. Wirel. Commun. **19**(6), 3796–3811 (2020). https://doi.org/10.1109/TWC.2020.2978481
64. Maraqa, O., Rajasekaran, A.S., Al-Ahmadi, S., Yanikomeroglu, H., Sait, S.M.: A survey of rate-optimal power domain NOMA with enabling technologies of future wireless networks. IEEE Commun. Surv. Tutor. **22**(4), 2192–2235 (2020). https://doi.org/10.1109/COMST.2020.3013514. (**Fourthquarter**)
65. Marcano, A.S., Christiansen, H.L.: A novel method for improving the capacity in 5G mobile networks combining NOMA and OMA. In: 2017 IEEE 85th Vehicular Technology Conference (VTC Spring), pp. 1–5 (2017). https://doi.org/10.1109/VTCSpring.2017.8108677
66. Mutalemwa, L.C., Shin, S.: A classification of the enabling techniques for low latency and reliable communications in 5G and beyond: AI-enabled edge caching. IEEE Access **8**, 205502–205533 (2020). https://doi.org/10.1109/ACCESS.2020.3037357
67. Nduwayezu, M., Pham, Q., Hwang, W.: Online computation offloading in NOMA-based multi-access edge computing: a deep reinforcement learning approach. IEEE Access **8**, 99098–99109 (2020). https://doi.org/10.1109/ACCESS.2020.2997925
68. Next Generation Internet of Things, "D3.3: A Roadmap for IoT in Europe. Research, innovation and implementation 2021-2027", NGIoT. https://www.ngiot.eu/ngiot-report-a-roadmap-for-iot-in-europe/ (2022)
69. Oh, J., Guo, Y., Singh, S., Lee, H.: Self-Imitation Learning. In: Proceedings of the 35th International Conference on Machine Learning, PMLR, vol. 80, pp. 3878–3887 (2018)
70. Pan, Y., Chen, M., Yang, Z., Huang, N., Shikh-Bahaei, M.: Energy-efficient NOMA-based mobile edge computing offloading. IEEE Commun. Lett. **23**(2), 310–313 (2019). https://doi.org/10.1109/LCOMM.2018.2882846
71. Qian, L.P., Feng, A., Huang, Y., Wu, Y., Ji, B., Shi, Z.: Optimal SIC ordering and computation resource allocation in MEC-aware NOMA NB-IoT networks. IEEE Internet Things J. **6**(2), 2806–2816 (2019). https://doi.org/10.1109/JIOT.2018.2875046
72. Rahman, A.: Network intelligentizing for future 6G wireless networks, future communication summit, Lisbon, November 2019. https://futurecomresearch.eu/previous/2019/slides/Md_Arifur_Rahman.pdf

73. Rajput, V.S., Ashok, D.R., Chockalingam, A.: Joint NOMA transmission in indoor multi-cell VLC networks. In: IEEE 30th Annual International Symposium on Personal. Indoor and Mobile Radio Communications (PIMRC), vol. 2019, pp. 1–6 (2019). https://doi.org/10.1109/PIMRC.2019.8904250
74. Randrianantenaina, I., Kaneko, M., Dahrouj, H., ElSawy, H., Alouini, M.-S.: Interference management in NOMA-based fog-radio access networks via scheduling and power allocation. IEEE Trans. Commun. **68**(8), 5056–5071 (2020). https://doi.org/10.1109/TCOMM.2020.2988564
75. Rezaei, F., Tellambura, C., Tadaion, A.A., Heidarpour, A.R.: Rate analysis of cell-free massive MIMO-NOMA with three linear precoders. IEEE Trans. Commun. **68**(6), 3480–3494 (2020). https://doi.org/10.1109/TCOMM.2020.2978189
76. Rezaei, F., Heidarpour, A.R., Tellambura, C., Tadaion, A.: Underlaid spectrum sharing for cell-free massive MIMO-NOMA. IEEE Commun. Lett. **24**(4), 907–911 (2020). https://doi.org/10.1109/LCOMM.2020.2966195
77. Schiessl, S., Skoglund, M., Gross, J.: NOMA in the uplink: delay analysis with imperfect CSI and finite-length coding. IEEE Trans. Wirel. Commun. **19**(6), 3879–3893 (2020). https://doi.org/10.1109/TWC.2020.2979114
78. Schulz, P., et al.: Latency critical IoT applications in 5G: perspective on the design of radio interface and network architecture. IEEE Commun. Mag. **55**(2), 70–78 (2017)
79. Senel, K., Cheng, H.V., Björnson, E., Larsson, E.G.: What role can NOMA play in massive MIMO? IEEE J. Sel. Top. Signal Process. **13**(3), 597–611 (2019). https://doi.org/10.1109/JSTSP.2019.2899252
80. Shi, Z., Gao, W., Liu, J., Kato, N., Zhang, Y.: Distributed Q-learning-assisted grant-free NORA for massive machine-type communications 1–5. (2020). https://doi.org/10.1109/GLOBECOM42002.2020.9322273
81. Shi, Z., Gao, W., Zhang, S., Liu, J., Kato, N.: Machine learning-enabled cooperative spectrum sensing for non-orthogonal multiple access. IEEE Trans. Wirel. Commun. **19**(9), 5692–5702 (2020). https://doi.org/10.1109/TWC.2020.2995594
82. Shi, Z., Xie, X., Lu, H., Yang, H., Cai, J.: Deep reinforcement learning based dynamic user access and decode order selection for uplink NOMA system with imperfect SIC. IEEE Wirel. Commun. Lett. (2020). https://doi.org/10.1109/LWC.2020.3040402
83. Shone, N., Nguyen Ngoc, T., Dinh Phai, V., Shi, Q.: A deep learning approach to network intrusion detection. IEEE Trans. Emerg. Top. Comput. Intell. **2**(1), 41–50 (2018)
84. Stoica, R., De Abreu, G.T.F., Hara, T., Ishibashi, K.: Massively concurrent non-orthogonal multiple access for 5G networks and beyond. IEEE Access **7**, 82080–82100 (2019). https://doi.org/10.1109/ACCESS.2019.2923646
85. Study on new radio access technology Physical layer aspects, 3GPP 38.802 (2017)
86. Sun, Yu., He, Q.: Computational offloading for MEC networks with energy harvesting: a hierarchical multi-agent reinforcement learning approach. Electronics **12**(6), 1304 (2023). https://doi.org/10.3390/electronics12061304
87. Teerapittayanon, S., McDanel, B., Kung, H.-T.: BranchyNet: Fast inference via early exiting from deep neural networks. In: Proceedings of the IEEE 23rd Proceeding International Conference on Pattern Recognition (ICPR), pp. 2464–2469 (2016)
88. Tian, L., Yan, C., Li, W., Yuan, Z., Cao, W., Yuan, Y.: On uplink non-orthogonal multiple access for 5g: opportunities and challenges. China Commun. **14**(12), 142–152 (2017). https://doi.org/10.1109/CC.2017.8246331
89. Tseng, S., Chen, Y., Tsai, C., Tsai, W.: Deep-learning-aided cross-layer resource allocation of OFDMA/NOMA video communication systems. IEEE Access **7**, 157730–157740 (2019). https://doi.org/10.1109/ACCESS.2019.2950127
90. Tseng, S.-M., Tsai, C.-S., Yu, C.-Y.: Outage-capacity-based cross layer resource management for downlink NOMA-OFDMA video communications: non-deep learning and deep learning approaches. IEEE Access **8**, 140097–140107 (2020). https://doi.org/10.1109/ACCESS.2020.3004865
91. Vaezi, M., et al.: Cellular, wide-area, and non-terrestrial IoT: a survey on 5G advances and the road toward 6G. IEEE Commun. Surv. Tutor. **24**(2), 1117–1174 (2022). https://doi.org/10.1109/COMST.2022.3151028. (**Secondquarter**)
92. Wang, K., Zhou, Y., Yang, Y., Yuan, X., Luo, X.: Task offloading in NOMA-based fog computing networks: a deep Q-learning approach. In: IEEE Global Communications Conference (GLOBECOM). Waikoloa, HI, USA **2019**, pp. 1–6 (2019). https://doi.org/10.1109/GLOBECOM38437.2019.9013841
93. Wang, Huang, C.-M., Yin-Fu: Self-adaptive harmony search algorithm for optimization. Expert Syst. Appl. **37**, 2826–2837 (2010). https://doi.org/10.1016/j.eswa.2009.09.008
94. Wang, F., Xu, J., Wang, X., Cui, S.: Joint offloading and computing optimization in wireless powered mobile-edge computing systems. IEEE Trans. Wirel. Commun. **17**(3), 1784–1797 (2018)
95. Wei, F., Zhou, T., Xu, T., Hu, H., Tao, X.: A Joint Mechanism for Fog-Relay Networks Based on NOMA and Network Coding. 2019 IEEE Globecom Workshops (GC Wkshps), pp. 1–6 (2019). https://doi.org/10.1109/GCWkshps45667.2019.9024638
96. Wu, W., Wang, X., Zhou, F., Wong, K., Li, C., Wang, B.: Resource Allocation for Enhancing Offloading Security in NOMA-Enabled MEC Networks. IEEE Syst. J. https://doi.org/10.1109/JSYST.2020.3009723
97. Wysocki, T., Flizikowski, A., Marciniak, T.: Selected aspects of non-orthogonal multiple access for future wireless communications - for IoT. Sci. J. Telecommun. Electron. **24** (2020)
98. Xiao, C., Zeng, J., Liu, B., Su, X., Wang, J.: Cross-layer power control for uplink NOMA in IoT applications with statistical delay constraints. In: IEEE Global Communications Conference (GLOBECOM), Abu Dhabi, United Arab Emirates, 2018, pp. 1–7 (2018). https://doi.org/10.1109/GLOCOM.2018.8647452
99. Yang, Y., Marius, P.: A unified successive pseudo-convex approximation framework. IEEE Trans. Signal Process. (2017). https://doi.org/10.1109/TSP.2017.2684748
100. Yang, Z., Ding, Z., Fan, P., Al-Dhahir, N.: A general power allocation scheme to guarantee quality of service in downlink and uplink NOMA systems. IEEE Trans. Wirel. Commun. **15**(11), 7244–7257 (2016)

101. Ye, N., Li, X., Yu, H., Wang, A., Liu, W., Hou, X.: Deep learning aided grant-free NOMA toward reliable low-latency access in tactile Internet of Things. IEEE Trans. Ind. Inform. **15**(5), 2995–3005 (2019)
102. Ye, N., Li, X., Yu, H., Zhao, L., Liu, W., Hou, X.: DeepNOMA: a unified framework for NOMA using deep multi-task learning. IEEE Trans. Wirel. Commun. **19**(4), 2208–2225 (2020). https://doi.org/10.1109/TWC.2019.2963185
103. Ye, Y., Hu, R.Q., Lu, G., Shi, L.: Enhance latency-constrained computation in MEC networks using uplink NOMA. IEEE Trans. Commun. **68**(4), 2409–2425 (2020). https://doi.org/10.1109/TCOMM.2020.2969666
104. You, L., Yuan, D.: A Note on Decoding Order in Optimizing Multi-Cell NOMA. arXiv:1909.08651 (2019)
105. You, X., Wang, C.X., Huang, J., et al.: Towards 6G wireless communication networks: vision, enabling technologies, and new paradigm shifts. Sci. China Inf. Sci. **64**, 110301 (2021). https://doi.org/10.1007/s11432-020-2955-6
106. Zhang, Y., Wang, X., Xu, Y.: Energy-efficient resource allocation in uplink NOMA systems with deep reinforcement learning. In: 2019 11th International Conference on Wireless Communications and Signal Processing (WCSP), Xi'an, China, pp. 1–6 (2019). https://doi.org/10.1109/WCSP.2019.8927898
107. Zhang, N., Wang, J., Kang, G., Liu, Y.: Uplink nonorthogonal multiple access in 5G systems. IEEE Commun. Lett. **20**(3), 458–461 (2016)
108. Zhang, H., Qiu, Y., Chu, X., Long, K., Leung, V.C.M.: Fog radio access networks: mobility management, interference mitigation, and resource optimization. IEEE Wirel. Commun. **24**(6), 120–127 (2017). https://doi.org/10.1109/MWC.2017.1700007
109. Zhang, H., Qiu, Y., Long, K., Karagiannidis, G.K., Wang, X., Nallanathan, A.: Resource allocation in NOMA-based fog radio access networks. IEEE Wirel. Commun. **25**(3), 110–115 (2018). https://doi.org/10.1109/MWC.2018.1700326
110. Zhang, H., Zhang, H., Long, K., Karagiannidis, G.K.: Deep learning based radio resource management in NOMA networks: user association, subchannel and power allocation. IEEE Trans. Netw. Sci. Eng. **7**(4), 2406–2415 (2020). https://doi.org/10.1109/TNSE.2020.3004333
111. Zhang, J., Tao, X., Wu, H., Zhang, N., Zhang, X.: Deep reinforcement learning for throughput improvement of the uplink grant-free NOMA system. IEEE Internet Things J. **7**(7), 6369–6379 (2020). https://doi.org/10.1109/JIOT.2020.2972274
112. Zhang, J., Tao, X., Wu, H., Zhang, N., Zhang, X.: Deep reinforcement learning for throughput improvement of the uplink grant-free NOMA system. IEEE Internet Things J. **7**(7), 6369–6379 (2020)