A Survey on AI-Driven Energy Optimization in Terrestrial Next Generation Radio Access Networks

IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 24 September 2024, accepted 9 October 2024, date of publication 17 October 2024, date of current version 4 November 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3482561

SURVEY

A Survey on AI-Driven Energy Optimization in Terrestrial Next Generation Radio Access Networks

KISHAN STHANKIYA[1], (Graduate Student Member, IEEE), NAGHAM SAEED[2], (Senior Member, IEEE), GREG MCSORLEY[3], MONA JABER[1], (Senior Member, IEEE), AND RICHARD G. CLEGG[1]

[1]School of Electronic Engineering and Computer Science, Queen Mary University of London, E1 4NS London, U.K.
[2]School of Computing and Engineering, University of West London, W5 5RF London, U.K.
[3]Applied Research BT, IP5 3RE Suffolk, U.K.

Corresponding author: Nagham Saeed (nagham.saeed@uwl.ac.uk)

This work was supported in part by the University of West London Knowledge Exchange Seed Fund through the project "Working Toward Energy Efficient Wireless Network: A collaboration with BT Group to study and evaluate the AI and ML usage" under Project SF16-15036, and in part by the U.K. Engineering and Physical Sciences Research Council (EPSRC) under Grant EP/V519935/1.

ABSTRACT This survey uncovers the tension between AI techniques designed for energy saving in mobile networks and the energy demands those same techniques create. We compare modeling approaches that estimate power usage cost of current commercial terrestrial next-generation radio access network deployments. We then categorize emerging methods for reducing power usage by domain: time, frequency, power, and spatial. Next, we conduct a timely review of studies that attempt to estimate the power usage of the AI techniques themselves. We identify several gaps in the literature. Notably, real-world data for the power consumption is difficult to source due to commercial sensitivity. Comparing methods to reduce energy consumption is beyond challenging because of the diversity of system models and metrics. Crucially, the energy cost of AI techniques is often overlooked, though some studies provide estimates of algorithmic complexity or run-time. We find that extracting even rough estimates of the operational energy cost of AI models and data processing pipelines is complex. Overall, we find the current literature hinders a meaningful comparison between the energy savings from AI techniques and their associated energy costs. Finally, we discuss future research opportunities to uncover the utility of AI for energy saving.

INDEX TERMS Next generation mobile communication, energy efficiency, machine learning, power consumption, radio access networks.

I. INTRODUCTION

Energy and carbon reductions for mobile networks have never been more important given the goal to meet net-zero by 2050 and user data traffic is estimated to rise five-fold in moving to fifth generation (5G). The radio access network (RAN) remains a significant energy consumer (estimated 87% of network operations and up to 40% of operational expenditure (OPEX)) [1]. This has led to a push for artificial intelligence (AI) driven solutions for energy reduction in RAN deployments [2]. However, AI itself can have a large energy cost. Estimates for the energy cost of training a large-language model (LLM) such as OpenAI's GPT-3 stand at 1,287 MWh, whereas estimates for operational energy demand stand at 564 MWh [3]. Meta [4] estimates the energy footprint of AI inference of an in-house recommendation model (RM) to account for 40% of the whole model energy consumption. Similarly, Google [5] estimates AI inference alone accounted for 9% of their total energy use between 2019 and 2021.

This survey paper focuses on the RAN and looks at how AI/machine learning (ML) can be used to reduce power

The associate editor coordinating the review of this manuscript and approving it for publication was Adamu Murtala Zungeru.

consumption but also to consider the power consumption of the required AI inference. In particular, we investigate if the power cost of algorithms to reduce energy consumption can ever approach or exceed the energy saved. A high-level overview of the topics covered can be found in Fig. 1. We begin with a survey of RAN power consumption models asking whether the research community has a good and well-evidenced model of the power used by a RAN and this will be the basis for an accurate estimate of power saved. Following this we look at the different optimization models used to reduce power consumption considering the physical techniques used (what RAN parameters are being changed to get the power savings) and what AI techniques are being used to achieve this. We limit our survey to techniques that are already deployed or standardized and ready to deploy RAN technologies and report results with improvements in energy saving or energy efficiency. Finally, we investigate the question of how much energy might be consumed by AI models deployed for energy reduction. Because timeliness is vital in a rapidly moving field like this one we have chosen papers published in 2020 or afterwards with a few exceptions where older papers are a vital part of later understanding.

To answer the questions above, this survey is structured as follows. The remainder of this section reviews related survey papers highlighting the key differences of this work. This is followed by an outline of the scope of this survey. Section II introduces the 5G RAN architecture as a grounding for discussing power models in Section III. In Section IV, we survey the literature on energy-saving techniques, highlighting the key contributions in the time, frequency, power and spatial domains. In Section V, we review the areas that impact the energy cost of AI inference in the next-generation radio access network (NG-RAN) and, where required, draw in the broader research literature. Finally, in Section VI, we present concluding remarks with suggestions for future research directions.

A note on terminology: the terms ML and AI are often used somewhat interchangeably, to avoid the somewhat clumsy ML/AI we will use AI throughout in this survey unless there is a good reason to prefer the term ML in context (for example where the authors of a paper use this term). Many (but not all) techniques discussed have both a training phase (done once only or at infrequent intervals) which produces the parameters used by the model and an inference phase that produces the answer given a set of parameters. The training phase is typically more computationally intensive but, in a production network, the inference phase needs to be used every time an answer is required hence cannot be avoided as an operational cost.

A. RELATED SURVEY PAPERS

Reviews that focus on AI for power-saving in the RAN are well studied and the major competing surveys in this space since 2020 are [6], [7], [8], [9], [10], and [11]. To the best of the authors knowledge, the novelty in this work is an emphasis on also considering the energy cost of AI.

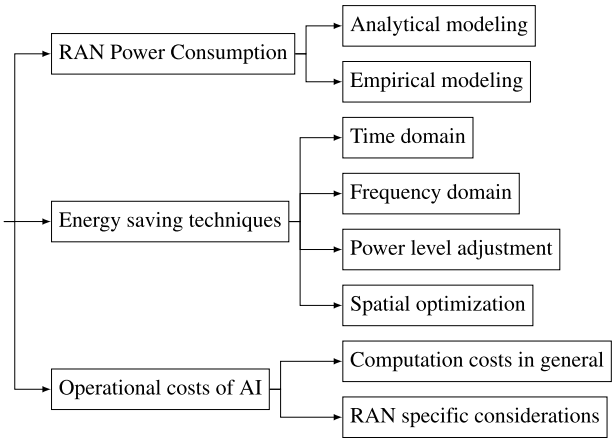


FIGURE 1. High-level taxonomy of topics covered in this survey.

A summary of the other papers compared with this one is given in Table 1. Some surveys prefer to give their attention to future enablers for 6G technology [6], [10], [11] that are not covered by this paper. By contrast, our focus is on technologies deployed today or standardized and ready for deployment, by studies where quantitative energy savings are reported which could be immediately beneficial.

While three of the other studies include power consumption models [7], [9], [11] the surveys [7], [9] do not break down models into analytical or empirical and the other [11] uses only older third generation (3G) models of power. This survey, by contrast, offers a timely breakdown of the analytical and empirical power consumption models using current generation technology. Most surveys do not cover the downside of optimization, the energy cost of AI. While [9] highlights computational effort as the number of operations per second, this still misses a huge number of factors that contribute to algorithmic power consumption. By contrast, this survey details the factors involved in the power consumption of an AI algorithm. The only survey the authors found that covers this field reasonably is [12] but this survey is now five years old whereas we focus on AI techniques from 2020 onward. This is the key differentiator between this survey and others in the field.

Other surveys have included a number of works that look at techniques to manage energy consumption in the RAN but we believe this to be the most up-to-date and complete. Both [7] and [10] give extensive explanations on how sleep modes and different levels of shutdowns work for power saving at a base station, whereas [11] focus on ways to maximize sleep duration. These surveys are from 2022 and 2023 respectively so our survey complements and updates them.

Interference management for energy efficiency is covered in [6] and [11] but the former focuses only on remote radio head clustering in cloud RAN and the latter on only techniques that modify transmit power. The survey [10] highlights the novelty of rate splitting (RS) for efficiency which we also cover. In this survey, we look at how

TABLE 1. A comparison of our work with other survey papers since 2020.

Topic	Refs						
	[6]	[7]	[8]	[9]	[10]	[11]	Our work
Current RAN challenges	×	✓	✓	✓	×	×	✓
Empirical RAN power models	×	✓	×	✓	×	✓	✓
AI power factors	×	×	×	✓	×	✓	✓
Sleep modes	✓	✓	✓	✓	✓	✓	✓
Rate splitting	×	×	×	×	✓	×	✓
Interference management	✓	×	×	×	✓	✓	✓

scheduling techniques can help to reduce delay, power consumption and maximize profit for an operator. This is a promising area of research, but discussions in the literature have been sparse in recent works. For instance, [7] do not consider it and [9] limits their discussion to one study. In contrast, [10] covers multiple operator sharing and baseband workload scheduling.

B. SCOPE AND CONTRIBUTIONS

This survey focuses on the impact of AI-based algorithms on reducing power usage and the energy cost of that AI and focuses on developments since 2020 (although older papers are included, particularly in considering power estimation, where they remain the state-of-the-art). The survey is of viable techniques used in current 5G installations where energy savings are explicitly reported. We categorically do not cover supply-side power management technologies such as improved power generation, renewable energy, battery, or smart grid technologies. We recognize the potential utility of post 5G technologies, such as nonterrestrial networks (NTNs) [13], optical wireless communication (OWC) [14], [15] and terahertz (THz) [16], [17] communications but these are not our focus here. Our main contributions include:

1) Identifying the analytical and empirical power consumption models in the RAN. We compare how power consumption models are delineated based on the scope and architecture.
2) A timely review of leading research on RAN energy efficiency (EE), classifying the studies by their leading degree of freedom (e.g. time, frequency, power, and spatial domains).
3) Discussion of the factors that impact the operational energy cost of using AI techniques as this may mitigate any savings made.

Following the survey, we highlight the gaps in the existing research, providing insights into directions for future research on AI for improving RAN EE.

II. NG-RAN ARCHITECTURE

When planning a RAN deployment, the design is typically over-provisioned in order to be future-proof (because deployment is costly), and to cope with peak load. Breaking down hardware functions into decoupled logical units creates opportunities for more granular scaling, gains in power consumption and efficiency. Hosting network functions in different physical locations and hardware can allow efficient responses to changes in demand patterns. In this section, our focus is to describe the logical units that constitute the NG-RAN, forming a foundation for later discussions on similarities and differences between power models.

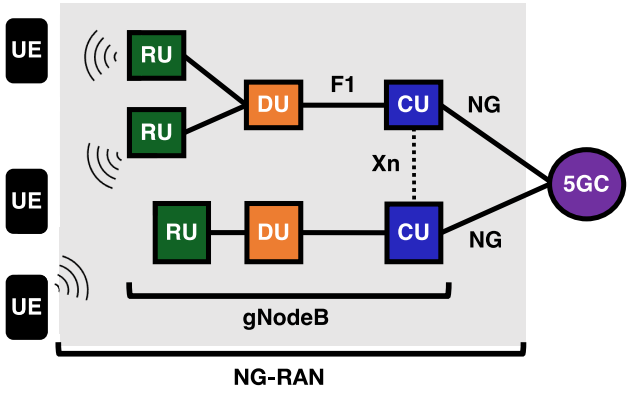


FIGURE 2. Overview of 5G System. Composed of UE, NG-RAN (shaded) and 5G core network.

Depicted in Fig. 2, the NG-RAN is a collection of several of base station known as next generation NodeB (gNB) [18]. Each gNB contains one centralised unit (CU) and one or more distributed units (DUs).[1] Complementary to the third generation partnership project (3GPP) specifications, the Open Radio Access Network (O-RAN) Alliance defines standards to promote architectures that use open interfaces while fostering hardware disaggregation, flexibility and network intelligence [8]. This makes O-RAN a key technology to allow AI to be used in RAN. In addition to the aforementioned logical units, the O-RAN describes radio unit (RU),[2] where each DU connects to one or more RUs. The connections between physical or logical nodes in the 3GPP specifications for NG-RAN are described in [19]. The edges between logical units describe the crosshaul transport network (xHaul).

The radio signals between user equipment (UE) and NG-RAN are transmitted and received by the RU. The RU are always located at network operator cell sites, which are spatially distributed to ensure geographical coverage. The RU

[1]Formally, a CU and DU are referred to as gNB-CU and gNB-DU, respectively, but we omit the prefixes for simplicity.

[2]The 3GPP specifications do not formally include RU as part of the gNB. However, we do here because of the significant impact on power consumption.

converts between the analogue radio signals used by antennae and the digital signals used by the DU. The DU connects to one or more RUs and the CU typically handles the higher level protocol stack.

As previously mentioned, the xHaul describes the transport network supporting the sending and receiving of signals between RU, DU and CU nodes. It is made up of the fronthaul (RU-DU), midhaul (DU-CU) and backhaul (CU-Core).

III. RAN POWER CONSUMPTION MODELING

In order to properly evaluate the influence of AI on the EE of the NG-RAN, it is crucial to understand the assumptions of models that provide estimates of power consumption. These models must consider the enabling technologies used in the NG-RAN while remaining flexible to evolving RAN architectures. Two main types of studies emerge from the literature, *analytical models* and *empirical models*. Analytical models here attempt to derive equations from physical principles that could estimate power consumption given correct input data and physical parameters. By contrast, the empirical models use measured data to attempt to ground these estimates in the real world. In this section, we present prominent models and approaches used for NG-RAN power consumption. It should be noted that publications creating new power models are far less frequently published and, hence, the references in this section are older since in this area papers that represent the current best state of understanding can be more than ten years old.

A functional split describes the division of the baseband processing chain and which logical nodes are responsible for carrying out that function. When considering the power consumption of the logical nodes in the NG-RAN (e.g. RU, DU, or CU), it is essential to consider how power consumption will be affected by the chosen functional split since the network functions hosted at each node will impact the computational load and therefore the energy consumed. Some authors have looked in detail at the effects of functional split on energy efficiency [20], [21], [22], [23]. For example, *service differentiation*, a technique using backup virtual network functions (VNFs) to improve resilience and central processing units (CPUs) over-provisioning to decrease the queuing delay of the VNFs, improves EE [20]. Live migration of virtualized resources reduces the number of ''switched-on'' servers reducing the average energy consumption by 8% [21]. Similarly, placement of CUs and DUs in a metro access network is solved using a heuristic [22] saving almost 8% of total power when compared to a static mixed-integer linear programming (MILP) approach. In contrast, [23] use deep reinforcement learning (DRL) to achieve dynamic VNF splitting in an O-RAN scenario by as much as 63% compared to a Greedy algorithm approach. Therefore, it is crucial to understand the power consumption of each individual component, including the RU, supporting infrastructure for both distributed and centralized units (whether physical or virtualized), and the xHaul. A recent report by NGMN for Green Future Network[3] emphasizes the importance of hardware metering when standard COTS (Commercial-off-the-shelf) that would host some of these VNF with the aim of determining the energy consumption analyses how the cloud model could be harnessed to optimize the energy efficiency.

A. ANALYTICAL MODELS

Analytical models attempt to construct estimates for power consumption from equations based on physical principles. Our survey found three major power models for RAN networks. However, each covers a slightly different part of the system, and each makes different assumptions about which components have constant power consumption and which components vary with load. Fig. 3 will be used to illustrate which components are included in each of the three major models we cover. The top (green) box represents the RU model from [24], the middle (pink) box represents the base station base station (BS) model known as EARTH [25] and the bottom (yellow) box represents the mMIMO model from [26]. The models in these works are extremely detailed and here a high-level view is given.

The authors in [24] formulate a power model for an RU (green box in Fig. 2). This model accounts for current (3GPP Release-18) and future multiple-in multiple-out (MIMO) architecture considering the scaling effects on power consumption of discontinuous transmission and reception schemes, antenna muting and chip processing in addition to radiated transmit power. The mapping of components of this RU model can be found in Fig. 3 (green shaded box). The RU power consumption model [24, Eq.1] is presented as[4]:

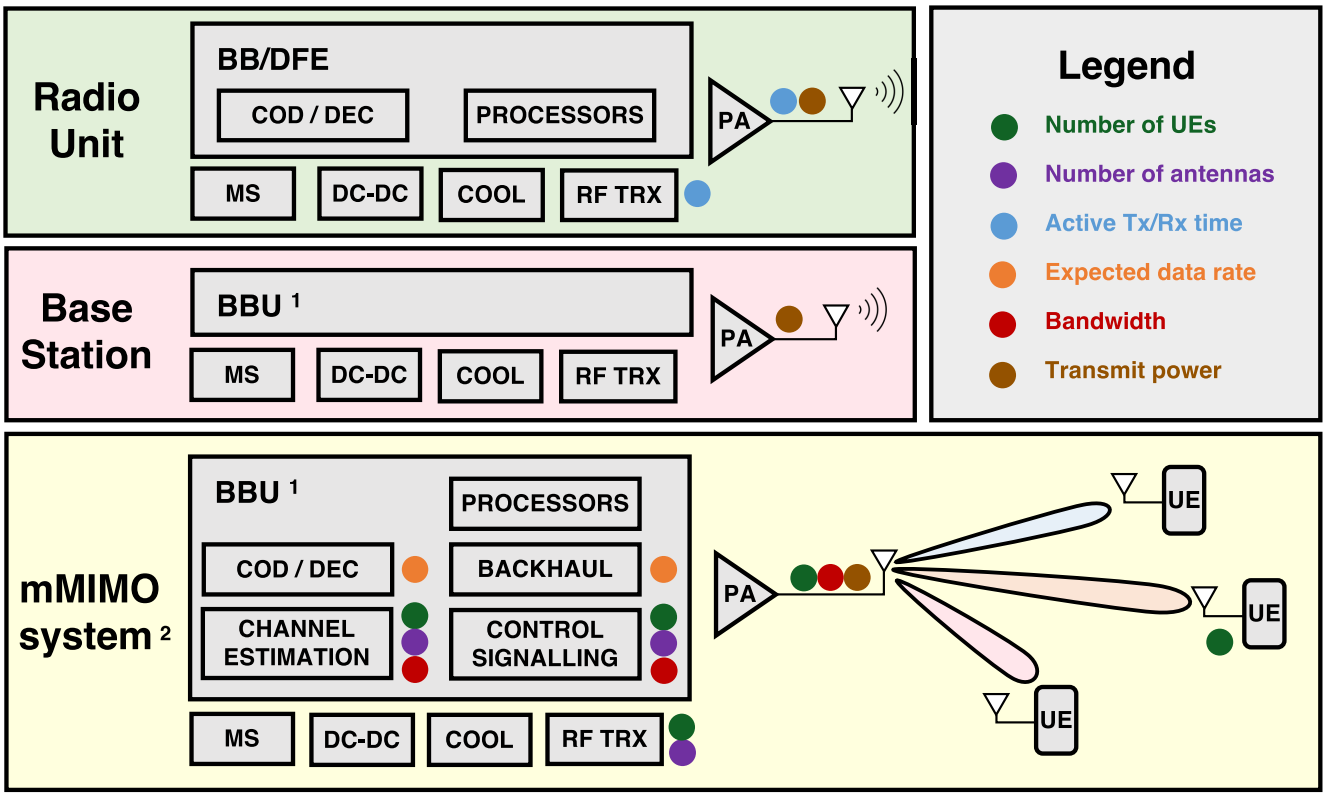
$$P_{\mathrm{RU}} = \beta M P_{\mathrm{Tx}}^{\mathrm{dyn}} + (1-\beta) M P_{\mathrm{Rx}}^{\mathrm{dyn}} + M C P_{\mathrm{static}} \qquad (1)$$

where M is the number of transceiver chains, β is the uplink-to-downlink ratio for time division duplexing (TDD), $P_{\mathrm{Tx}}^{\mathrm{dyn}}$ is the dynamic transmit power, $P_{\mathrm{Rx}}^{\mathrm{dyn}}$ is the dynamic receive power, C is the ratio of active computation power to total computation power, and P_{static} accounts for the load-independent power consumption of the digital front-end (DFE) and baseband (BB).

As noted in [24], the burden of processing functions for BB and DFE are moving towards an integrated RU, where it was once reserved for dedicated hardware called the baseband unit (BBU). The combination of RU + BBU, commonly referred to as a BS, is equivalent to the functions carried out in all three parts of a gNB, namely the RU, DU and CU, as seen in Fig. 2. In lieu of reference models for DU and CU power consumption, traditional distributed RANs power consumption models are commonplace. A cornerstone model, capturing more functions than the aforementioned RU model is the EARTH framework [25] which maps the radio frequency (RF) output power (P_{out}), measured at the

[3] https://www.ngmn.org/

[4] The original paper gives a more detailed equation, whereas we summarize terms here.



[1] The mMIMO BBU components are part of the Base Station model but are assumed load-independent and omitted here for brevity.
[2] This model focuses on the effective transmit power rather than just the radiated power.

FIGURE 3. A comparison of power consumption models from the literature focused on the radio unit, base station and Massive Multiple-Input and Multiple-Output (mMIMO) system. Components include the baseband-digital front end (BB/DFE), baseband unit (BBU), channel coding and decoding (COD/DEC), mains power supply losses (MS), direct current conversion losses (DC-DC), active cooling losses (COOL), radio frequency transceiver (RF TRX), power amplifier (PA) and user equipment (UE). Grey components without a dot indicate a component with load-independent power consumption. Components with a dot represent a dynamic power consumption, where the color represents the influencing factor.

antenna interface, to the total supply power of a BS. A visual representation of this *Base Station* model is shown (pink shaded box) in Fig. 3. Abstracting away physical hardware components (e.g. BBU, power amplifier (PA)) and conversion losses (e.g. cooling, mains supply) from their complex model, the power consumption for a fourth generation (4G) BS in [25, Eq. 1] is presented as:

$$P_{\text{BS}} = \begin{cases} M \cdot (P_0 + \Delta_{ld} P_{\text{out}}), & 0 < P_{\text{out}} \leqslant P_{\text{max}} \\ M \cdot P_{\text{sleep}}, & P_{\text{out}} = 0 \end{cases} \tag{2}$$

where M is the total number of BS antennas, P_0 is power consumption independent of RF output, Δ_{ld} is the gradient of the power consumption dependent on RF output power, P_{out} is RF output power and P_{sleep} is the sleep mode power consumption. It should be highlighted that when comparing (1) and (2), $P_0 \neq P_{\text{static}}$, as the assumptions of what constitutes a load-independent factor of the power models differ, as illustrated in Fig. 3.

Building on [25], the authors in [27] developed a tractable power model by factoring in PA output range and transmission bandwidth. The *GreenTouch* framework in [28] and [29] further adopts a five-layer approach towards flexibility for future enabling technologies. The opportunities to reduce power consumption within this modeling approach suggest three strategies for reducing power consumption, such as reducing the RF output power, reducing load-independent power consumption and maximizing P_{sleep}, which we cover in Section IV.

Technological enablers for NG-RAN, such as massive multiple-input multiple-output (mMIMO) and network function virtualisation (NFV), challenge assumptions for power consumption of past models. For example, the authors in [30] highlight the need for more sophisticated models when considering mMIMO systems which increase the complexity of the BS model. In particular, they assert that power consumption within the BBUs, RF transceiver chains (TRXs) and PAs varies with the number of antennas and UEs.

In [26], the authors derive the circuit power (P_{CP}) of a mMIMO from the number of antennas M, number of users K, effective transmit power P_{out} and gross rate $\bar{R}$, for different linear processing schemes. A high-level view of the mMIMO system power consumption model from [30, Eq. 21] may be summarized as,

$$P_{\text{sys}}^{\text{mMIMO}} = P_{\text{out}} + P_{\text{CP}}\left(M, K, \bar{R}\right), \tag{3}$$

The power consumption of hardware supporting the signaling between nodes is dependent on technology such as microwave radio, passive optical network (PON) and Ethernet [9] in addition to factors such as network topology, capacity and activity. In [31], the authors present an analytical model for the power consumption of the xHaul as the sum of 1) power consumption as a function of the bandwidth of the common public radio interface (CPRI) and Ethernet circuits between an access site and the central office; 2) power consumption of the radio stations and servers. In [32], the authors study how increasing RAN coverage and capacity affects the xHaul power consumption and which xHaul parameters impact RAN power consumption. Specifically, they model *power consumption* of a xHaul switch as:

$$\mathrm{PC} = P_{\mathrm{standby}} + (P_{\mathrm{bit}} \cdot N_{\mathrm{bits}}) + (P_{\mathrm{pk}} \cdot N_{\mathrm{packets}}), \quad (4)$$

where P_{standby} is the power consumption of a switch on standby, P_{bit} is power consumption per bit traversing the switch, N_{bits} is the number of bits, P_{pk} is the power consumption of per packet and N_{packets} is the number of packets. While the power consumption of the xHaul is not the main focus of this review, we recognize the multiplicative effect of data volume on the power consumption of the transport network and on the wisdom of efforts towards more efficient data transmission and scheduling approaches, as discussed later in Section IV.

B. EMPIRICAL MODELS

Empirical models are data driven and attempt to estimate power consumption from measurements of the system. We encountered two types of empirical model in this survey. The first type uses power consumption data from product information sheets provided by manufacturers enriched with features from traffic profiles or mobility. The second type conducts studies on testbeds to examine proposed designs and presents results from experimental methods, such as measurements from probes or power meters.

In [33] regression is used to analyze energy consumption data drawn from energy consumption sensors in 3G and 4G network deployments across 60 sites in three countries. They conclude to a good approximation a linear model relates traffic volume and emitted power and higher-order models lead to over-fitting. More recently [34] develop a power model for 5G multicarrier mMIMO active antenna units (AAUs), where a single power amplifier can support multiple carriers using multicarrier power amplification (MCPA) technology and deep dormancy or symbol, channel or carrier level shutdown. They initially explore a data-centric approach using an artificial neural network (ANN) and derive an analytical power consumption model based on the results. Compared to power models that do not account for MCPA effects [30], the proposed analytical model is described as being 1.5 times more accurate while maintaining a low mean absolute error of 5.6% compared to the ANN model.

The data from [33] is also the basis for [35], which presents field measurements on data and visitor volumes. Combining these with parameters for different RATs (Radio Access Technologies), including 5G RUs from Nokia product data sheets, they calculate and extrapolate the base station power consumption in dense urban and suburban areas of Finland. Compared with a measurement campaign of the same base stations, the proposed theoretical model for 5G is better at predicting energy consumption in the dense urban area, with the caveat that there are more users of the same type in that area.

The discussed models lack good open 5G data. They use 4G data [33], normalize the power consumption values [34] (to maintain commercial security) or speculate on the power dynamics of 5G hardware based on manufacturer reported spectral efficiency [35]. This demonstrates a lack of open research with clear reporting of empirical power consumption within real networks. As an alternative approach, the gray literature (*outside of formal commercial or academic publication*) can provide an intuition of peak power consumption of different types of 5G base stations. For example, actual power consumption from an anonymous operator shows that a 5G BS under full load consumes approximately 1.4 kW [36] for vendor equipment supporting one band, whereas another source reports 4.7 kW [37] for a different vendor supporting 3–4 bands. To put this into context, the power consumption of a consumer workstation PC from a leading vendor [38] ranges from 170 – 300 W [39], which is comparable to the reported BBU power in [36] and [37].

As BBU processing becomes disaggregated and workloads delegated to a DU or CU, these new nodes must cater for future growth. It is unsurprising, then, that datasheets for commercial servers advertised as suitable for DU workloads report peak power consumption between 300–1800 W [40], [41], [42]. The EARTH power consumption model [25] is popular but predates the move toward virtualised base stations (vBSs) and may not capture the power implications. Motivated by this, the authors in [43] approximate vBS power consumption derived from experimental results from uplink transmissions in a testbed. Virtualization tackles the problem of over-provisioning, allowing resources to be scaled to the user demand and afford resilience. When considering vBS it is important to know the cost of virtualization.

In [44] the authors provide three open datasets, including the energy consumption of a vBS as a function of a range of parameters within an O-RAN compliant testbed. Similarly, the authors in [45] measure the power consumption (wattage) of software implemented physical layer (PHY) for 5G NR using Intel's running average power limit (RAPL) machine specific registers (MSR), they measure CPU and dynamic random access memory (DRAM) power, estimating the measurement overhead as $\leqslant 1\%$. This proves a valuable study to dimension the energy consumption of a software-defined NG-RAN approach.

When NG-RAN functions are virtualized it is important they still satisfy latency constraints. This is mitigated by having a dynamic functional split [46], which allows baseband processing to move closer to the cell site when

required. Measuring the impact of different functional splits, the authors in [47] profile the energy consumption in an Open Air Interface testbed. Specifically, they attempt to profile the power consumption of a DU and CU by varying the CPU clock frequency and channel bandwidth. They find cases where CPU clock frequency could be reduced for use cases where full-buffer traffic is employed. These results are based on a single user and connecting to a RU modeled using software-defined radio (SDR). Therefore, the applicability of these results when scaled up to operational network volumes and optimized cloud radio access network (C-RAN) datacenters, remains an unanswered question. Moreover, since the processing is decoupled from hardware (which vary between architectures), an extension of this study to quantify the computational complexity, per layer of the radio stack in the DU and CU, as a function of throughput, would provide a useful future-proof contribution. For example, what would be the empirical computational load (in floating point operations per second (FLOPS)) to run radio link control (RLC) processing while ensuring a data rate of 100 Mbps?

Testbeds also provide a way to measure the energy consumption of the xHaul. Considering the power consumption of access networks, the study [48] presents energy consumption figures for digital subscriber line (DSL), hybrid fiber coaxial (HFC) networks, PON, fibre-to-the-node (FTTN) and point-to-point (PtP) optical. The study found that optical networks are the most energy-efficient. Later studies by [49] show that energy consumption does not grow proportionally with the number of ports, and [50] show that high-capacity routers and switches use 80-90% of their maximum power whilst at idle load. More recently, in [51], the authors provide a measurement methodology for power profiling based on a linear model for rate adaptation. They provide testbed measurements for two types of 24-port 1GbE switches (one with fixed ports, another with modular) and three routers (edge, fixed aggregation and modular chassis aggregation). Although the xHaul is not the main focus of this survey, we note that there is a need to integrate the heterogeneity of transport network technologies, into the power modeling for NG-RAN for a more accurate representation of the energy impact.

C. POWER MODELING SUMMARY

This survey covers all the major RAN power modeling papers the authors could locate, but it is notable what was missing from the literature. In many ways, it is not unexpected that no papers were found that unambiguously showed the net power consumption for a 5G deployment. Some parameter fitting was done against small deployments of 3G and 4G systems. The most likely explanation is that the information needed to do this is extremely hard to obtain and would be commercially sensitive. In the case of analytical models, it means that necessary parameters are not well-known and, while the power models can be used as part of a modeling package, the uncertainties in the absolute value of the result may be large. In the case of empirical models, it means that unobscured results for 5G systems have, at best, been tested against small test beds. The formal academic literature could not provide even an order of magnitude estimate for the power consumption of a single BS in a ''typical'' installation. Looking outside formal publications we were able to find two estimates of 1.4kW (for a single band) and 4.7kW (for three or four) for a 5G BS under full load, however, it is highly unsatisfactory to resort to such untrustworthy sources. Following such sources further did not seem to fall into the scope of a survey of academic literature.

IV. ENERGY SAVING TECHNIQUES

This section investigates the techniques researchers have used to reduce power usage in RAN networks. We categorize techniques into time, spatial, frequency and power domains. We provide a brief overview of approaches within each domain that show promising gains in network energy efficiency. Table 2 provides an at-a-glance summary of the papers considered. Numerical comparison of results between papers was quickly discovered to be insurmountable for a number of reasons. Different authors use different metrics to measure energy saving/efficiency. Some studies allow energy efficiency to be traded against degraded user experience whereas others assume the user experience must remain at least as good. Finally, the studies are done with different modeling assumptions such as path-loss models and layouts of BS and UE. For these reasons, it is not possible to look at relative gains between two papers and deduce which is better at improving energy efficiency simply by comparing the claimed saving.

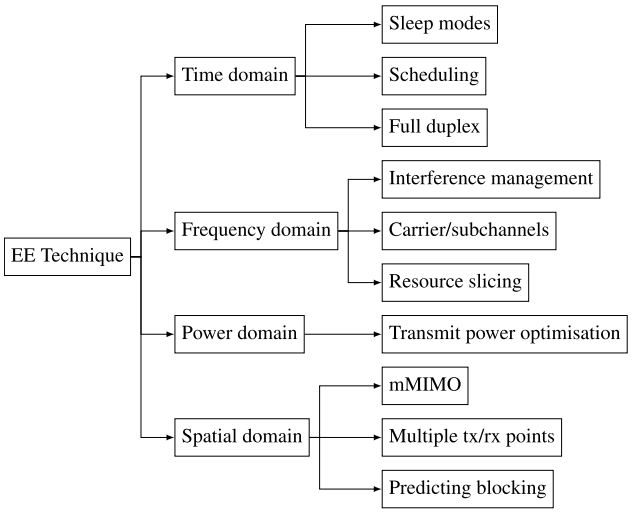


FIGURE 4. Taxonomy of power saving techniques in this survey.

Techniques can be split by settings altered in the modeled RAN or by the AI techniques used to alter those settings. We have split the techniques into four broad areas of resources within wireless communication. *Time domain* techniques primarily work by moving resources in time.

Frequency domain techniques optimize by changing the frequencies at which signals are sent. Works based on how UEs, BS and the signals between them are positioned in physical space. Obviously, some studies will use more than one of these areas in optimization, jointly optimizing transmission power and frequency use. Where a study could fit in more than one section we have tried to fit it according to the main technique used in the primary result presented by the authors. Fig. 4 shows our taxonomy of techniques based on this split.

There is a great deal of interest in using the data collected in 5G networks to optimize power reduction and 3GPP [52] highlights AI/ML based solutions to reduce network energy. The techniques they highlight include cell deactivation/sleep (power domain), coverage modification (power and spatial domain) and traffic offloading (spatial domain). New abilities unlocked by 5G enable new techniques that can be used to save energy. For example, having the DU/CU as a virtual appliance removes the dependence on application-specific network hardware, with efficient software implementations that can run on lower-cost general-purpose processing platforms. This also allows functions to dynamically move between hosts and scale computation resources based on the performance requirements or network demand. Scaling with network load helps to improve energy efficiency by dynamically powering down resources during idle periods [53].

A. TIME-DOMAIN

Sleep modes have naturally attracted a lot of research interest. As mobile networks are designed to handle peak capacity, some base stations remain powered on outside peak hours despite being underutilized. As indicated in (2) sleep modes (SMs) are a way to dynamically switch base stations between an active state and power off components in an idle state, depending on user attachment status. This helps to reduce energy consumption in mobile networks by deactivating unnecessary components of the radio transceiver chains when traffic is low. With advanced sleep modes in 5G, a base station is progressively put into deeper sleep modes during increasing periods of inactivity. While deeper sleep modes have lower power consumption, they also introduce longer reactivation delays, impacting user quality of service (QoS). Investigating the compromise between energy saving and delay, the authors in [53] show that dynamic adjustment of the time spent at each SM level can reduce the reactivation delay by 90% for low loads. They further validate a stochastic model to tune parameters in real time.

Reinforcement learning is a promising approach to achieving real-time optimization. The authors in [54] propose a traffic-aware DRL based sleep control approach for base stations in large-scale networks using precise mobile traffic forecasting that combines geographical, temporal and semantic spatial (cosine similarity across traffic loads) correlation. They demonstrate that their approach can achieve a 20% reduction in cost, with energy being the most significant contributing factor, compared to an autoregressive integrated moving average (ARIMA) forecasting model [76].

Operating at the millimeter wave (mmWave) spectrum provides wider bandwidth and, therefore, data rates, but signals do not propagate as far, leading to smaller coverage areas. In contrast to macro cells, which cover wider areas, small cells offer increased capacity per geographic area. Small cells are densely deployed to meet the needs of 5G networks. However, overlaps in the coverage area for a small cell in a macro cell region or between neighboring small cells can cause increased interference. In [55], the authors investigate the impact of small cells on the overall performance of 5G networks. They focus on end-users quality of service (QoS) constraints and account for inter-cell interference in a heterogeneous network. The authors propose a distributed Q-learning algorithm that controls the activities of small cells based on their interference level, expected throughput, and buffer size. The findings suggest that, under low traffic loads, moving users from small cells to macro cells can reduce delay and energy consumption for the cluster. However, this may increase the overall network energy consumption at the cost of the average user throughput.

In another reinforcement learning approach [56], the authors use a state-action-reward-state-action (SARSA) algorithm to set a sleep mode policy while studying the impact of the wake-up delay of the sleep mode level on the end-to-end user packet latency for uplink traffic. Results show that by increasing the latency threshold to 5 ms and defining low traffic load at 5%, a 56% reduction in energy consumption is realized.

In [57], the authors utilized a deep Q-learning network (DQN) to conserve 5-10% more power across all levels of user demand at low loads of up to 50 Mbps. However, the researchers found it unfeasible to meet user quality-of-service demands beyond this point. The DQN method involves many state-action pairs, which increases computational complexity and reduces system performance.

Scheduling is crucial in NG-RAN to manage increasing data volumes while meeting latency demands and reducing power consumption. Two studies, [58] and [59], address this challenging task. Researchers in [58] formulated a Markovian model that efficiently schedules proactive caching and on-demand transmission, analyzing the average delay and power consumption. On the other hand, [59] proposed policies that motivate the fact that data processing time strongly depends on the transmission modulation and coding scheme (MCS) index. These policies allow the radio scheduler to set the MCS index for users' transmission based on the radio conditions and the BBU pool's ability to process users' data. Furthermore, they propose heuristics to reduce power consumption compared to non-coordination heuristics. Considering the economics of RAN operation, [60] proposed a profit-based algorithm that optimizes task scheduling and resource allocation for C-RANs towards maximizing profit margins for network operators. This highlights the need for algorithms adapting to network

TABLE 2. Summary of EE techniques.

Ref	Year	Optimization	Key Findings or Contributions
			Time domain techniques
[53]	2021	Stochastic	Reduces sleep mode reactivation delay by 90% for low loads, allowing more time in sleep mode for energy conservation.
[54]	2021	Reinforcement Learning	20% reduction in cost of sleep control compared to ARIMA model.
[55]	2022	Distributed Q-learning	Sleep mode policy optimization with QoS constraints, 5 times more efficient than macro cell offloading.
[56]	2023	Reinforcement learning	Optimal sleep mode management for trade-off of energy consumption with 5% energy savings with less than 1 ms latency.
[57]	2022	Deep Q-network	Proposed sleep mode activation saves 5-10% more power than baseline.
[58]	2021	Heuristic	Low complexity proactive scheduling algorithm minimizing queuing delay with power constraints.
[59]	2023	Heuristic	Reduce wasted power by 48% from retransmissions by coordinated scheduling of radio and computing resources.
[60]	2021	Linear Programming	Scheduling algorithms maximize profit margins for network operators.
[61]	2023	Analytical	Reduce fronthaul capacity demand by duplexing thereby lower energy requirements.
[62]	2022	Quadratic Programming	Serves multiple users efficiently over long distances using duplexing.
			Frequency domain techniques
[63]	2023	Reinforcement Learning	Resource allocation strategy to enhance energy efficiency in ultra-dense networks.
[64]	2023	Lyapunov framework	Optimize interference by sequential user scheduling and power allocation approximation algorithm in Edge SON architecture.
[65]	2022	Heuristic	Resource allocation strategy reduces power consumption up to 93.8% (low-load) and 64% (high-load) in C-RANs
[66]	2021	Fractional programming	Proposed algorithm reduces computing time cost compared to baseline.
[67]	2023	Quasi-Newton	Resource allocation optimized using second-order optimizer for RAN Slicing, compared to dense neural network (NN) approach.
			Power domain techniques
[68]	2023	Deep Neural Network	Proposed DNN for power allocation reduces computational complexity, while providing similar performance.
[69]	2023	Deep Neural Network	Proposed resource allocation scheme achieves 96% EE of the optimal solution, with reduced computation time.
[70]	2020	Reinforcement Learning	Proposed power control algorithm for improved throughput for cell-edge users in urban macro scenarios.
[71]	2021	Stochastic	Proposed transmit power minimization with imperfect CSIT, serves more users while satisfying minimum QoS constraints.
			Spatial domain techniques
[72]	2021	Analytical	Dynamic scaling down of Massive MIMO antenna array size to reduce energy consumption by 30%
[73]	2022	Analytical	Optimizes the number of spatial layers and band activation against user rate requirement while minimizing macro cell power consumption.
[74]	2021	Combinatorial	Transmission reception point selection algorithm improves weighted sum energy efficiency in uplink mmWave network over existing methods
[75]	2022	Lyapunov framework	Polynomial complexity beam activation and user scheduling algorithm, saves 65% of average RRH energy consumption in 5G CoMP networks.
[16]	2022	Deep Neural Network	LiDAR-aided link blockage prediction enables proactive hand-off with lower delay and more efficient use of network resources

performance and economic conditions to ensure sustainable growth.

Full duplex systems handle simultaneous data transmission and reception and this has been a catalyst for technological advancements, leading to increased data throughput and enhanced energy efficiency. In [61], the authors employed stochastic geometry to analyze a C-RAN-enabled full-duplex (FD) cellular network, revealing that strategic downlink (DL) power control significantly boosts the mean rate and mitigates the substantial fronthaul capacity demands in C-RAN, thus conserving energy. On the other hand, researchers in [62] propose a hybrid full-duplex transmission model tailored for 5G networks. The model combines single-mode fiber with free-space optics for mmWave signal transmission. It utilizes variable quadrature amplitude modulation to efficiently serve multiple users over long distances.

B. FREQUENCY-DOMAIN

Interference Management (IM) involves avoiding or minimizing interference in a wireless network. In a heterogeneous network (HetNet), small cells (SCs) can generate interference or be affected by interference from a Macro Base Station (MBS) or other nearby SCs. In dense deployments, interference management is crucial for energy efficiency, and intelligent algorithms may help. For example, [63] proposes a reinforcement learning-based resource allocation algorithm to enhance energy efficiency in ultra-dense

networks, employing Q-value approximation to tackle the problem of large state spaces and reduce convergence time. Meanwhile, an edge self-organising network (SON) architecture is proposed in [64], integrating centralized and distributed approaches to manage cellular networks with an algorithm that uses Lyapunov optimization for interference management towards performance improvements in real 5G networks. Finally, the work in [65] focuses on a two-level hybrid resource allocation framework in C-RANs that significantly reduces power consumption by up to 93.8% in low-load conditions, utilizing an admission control algorithm and a heuristic-based RRH-BBU mapping algorithm to optimize the number of users and manage interference, considering both BBU capacity and user QoS constraints. These studies demonstrate innovative approaches to optimizing network performance while prioritizing energy conservation and effective interference management.

Carrier/subchannel optimization provides another avenue for energy saving. RS describes the idea where user messages are segmented into common and private components at the transmitter, while at the receiver partial decoding of interference is used and the remainder is treated as noise. With a focus on RS, a transmit scheme for device-to-device (D2D) underlaid cache-enabled C-RANs is proposed by the authors in [77]. They focus on maximizing the sum rate while adhering to power and fronthaul cost constraints through user grouping, dynamic clustering, beamforming, RS ratio, and subcarrier allocation. They present algorithms for each subproblem, leading to convergence at a stationary point. The proposed technique achieves a 22% gain in sum-rate versus D2D random scheduling, but the specific amount of energy saved is not quantified.

Resource slicing is a term referring to how resource blocks in a RAN are allocated (at heart this involves both the time and frequency domain but here we have included it as a frequency domain technique). Conscious of the energy impact of their previous works [78], researchers in [67] looked at energy optimization using RAN slicing. They introduce KPIC-Lite, a neural network-based solution that consumes 700 to 1000 times fewer computational resources than previous models while maintaining performance in most tested scenarios. They offer a new loss function for better convergence and efficient use of a second-order optimizer to reduce computational resource usage. However, the specific energy savings related to RAN slicing operations are not explicitly quantified.

C. POWER-DOMAIN

Transmit power optimization is a technique to control networks and reduce power consumption. As users move further away from an access point, the transmit power must increase to ensure the signal can reach the receiver with sufficient strength. However, transmit power reduction strategies must be carefully managed to minimize the loss of signal-to-noise-plus-interference ratio (SINR), which would impact performance within the high density 5G networks. Recent studies have explored the potential of AI to optimize power allocation, such as the deep learning-based resource allocation scheme presented in [68]. This scheme includes a subchannel allocation algorithm and a power allocation strategy that uses deep neural networks specifically designed for the DL in heterogeneous nonorthogonal multiple access (NOMA) networks. Another study on ultra-dense small cell networks, [69], proposes a deep learning-based approach to maximize energy efficiency. Their method uses a neural network to determine the activation of small cell base stations, user association, and transmit power. It aims to achieve near-optimal energy management with less than 4 ms computation time across all considered cases, notably within the rigid latency constraints of 3GPP requirements [79]. Another study on dense 5G networks [70] proposes a data-driven approach based on deep reinforcement learning for DL power control to improve interference at the cell edge. In contrast to treating interference as an unwanted artifact, the authors in [71] consider rate splitting for interference mitigation, in addition to their primary focus power of transmit power minimization under imperfect channel state information at the transmitter (CSIT). They show that compared to the conventional ''treating interference as noise'' approach, RS uses a lower sum total transmit power for the same number of users in a C-RAN system.

D. SPATIAL-DOMAIN

Massive MIMO provides fine-grained spatial control of signals using multiple antennas. The term 'spatial elements' in this context refers to these antennas, a critical part of the system. The radio resource control (RRC) protocol manages the configuration of these radio resources. This protocol allows for periodic updates to the configurations, enabling changes in the number of antenna ports or elements that are actively used.

User equipment (UE), such as smartphones, play a vital role by providing the base station with Channel State Information-Reference Signals (CSI-RS). These signals convey the UE's understanding of the channel conditions. Presently, UEs can support various CSI-RS configurations [80], each corresponding to a different quantity of antenna ports or elements. This versatility permits the base station to dynamically adjust which spatial elements are engaged for transmitting data to the UE based on the channel state information (CSI) reports, thereby optimizing the communication to suit the prevailing channel conditions. These CSI-RS and corresponding CSI reports are tailored to specific segments of the available bandwidth, known as bandwidth parts (BWPs).

Towards spatial domain optimization, [72] propose and evaluate dynamic massive MIMO muting, which is a technique that can be used to scale down the active antenna array size when traffic demand is low, hence reducing energy consumption. Whereas the authors in [73] propose a spatial

and spectral resource allocation for energy-efficient massive MIMO 5G networks. Specifically, they consider spatial optimization by selecting the number of active antennas. The results from [73] highlight that a single spatial layer per physical resource block (PRB) achieves the lowest energy consumption in low-load scenarios.

Multiple transmission-reception points (TRPs) allow optimization of energy efficiency by more intelligently routing radio signals between UE and BS. Here, there is a capacity to adapt the number of TRPs actively transmitting and receiving signals and channels to a UE. Considering this, researchers in [74] study the joint transmission reception point (TRP) selection and resource allocation problem to maximize energy efficiency under imperfect channel state information CSI for an uplink mmWave network. In contrast, [75] focuses on the combinatorial beam activation and user scheduling problem; they propose an approximation algorithm to save 65% of average remote radio head (RRH) energy consumption for the same average queue backlogs compared to baseline algorithms, which do not consider energy consumption and queue backlog.

Predicting signal blockage can be used to increase efficiency through a better understanding of how signals actually propagate in a physical space. In [16] the authors use LiDAR-aided mobile blockage prediction in real-world mmWave systems. Here, spatial elements are considered to predict the physical location and movement of obstacles that can block line-of-sight (LOS) paths. This allows alternate signal paths to be used when a signal path is predicted to be weak. By predicting the blockages with high accuracy, their proactive scheme allows for lower delay and more efficient use of network resources.

E. SUMMARY OF AI TECHNIQUES FOR RAN EFFICIENCY

We have seen that a large number of levers are available for pulling to increase energy efficiency in a network. We have also seen that a large number of AI techniques can be applied to each. A frustration in this survey is the near impossibility of comparing between techniques in the published literature. A lack of reference models and common scenarios makes it irresponsible to compare a claim of x% saving in one paper with y% saving in another. While this problem will always remain difficult it could be mitigated by including test scenarios with set parameters that could be replicated between papers as a baseline. However, this relies on those scenarios containing sufficient modeling detail that they can capture the optimization details the researchers wish to model. A further problem is in the reporting of optimization efficiency. The computational requirements of the proposed schemes were orders of magnitude apart but it is unclear how to compare them. Some authors give asymptotic estimates of the algorithmic performance which is a good starting point but certainly not a panacea.

V. OPERATIONAL ENERGY COST OF AI AND ML

In this section, we review the factors that impact the operational energy cost of AI techniques[5] and consider tools that can help with this. Specifically, we highlight supporting literature for the costs of model inference (as this is the part of the model that will be run continually) considering aspects of data, software and hardware. AI forms an essential part of the future of NG-RAN [81], particularly in optimizing network energy usage. The workflow of an AI model includes training, testing, and deployment [82]. Models deployed in the NG-RAN require input of parameters and state information from the local node (e.g. gNB), UE and neighboring NG-RAN nodes [52]. The output of the model inference is then used to make predictions or decisions that are hoped to increase the performance and energy efficiency of the RAN. However, an open question is how the energy saving from improved efficiency compares with the energy cost of running the AI/ML pipeline. We will also consider in Section V-B the computational costs within the RAN of virtualizing network functions.

A. AI COSTS IN GENERAL

Insights into the costs of running AI in the NG-RAN context may be gained from studying the costs in a more general setting. A recent study [4], reports the energy footprint of data processing of a recommendation model (RM) accounts for 31% of the AI end-to-end pipeline, based on data center electricity use. In this section, we discuss the salient factors that impact the AI power consumption, with a summary provided in Table. 3, and commentary on the limitations of the studies.

The computational load of a model is primarily governed by the complexity of the model and the types of operations it must carry out. Models with a larger number of parameters impose a higher computation load to evaluate their relationships, translating into a higher power consumption. For example, in [84], the authors highlight the impact of modifying datasets to improve energy efficiency of algorithms. Notably, they observe that decreasing the feature set and volume of data points can achieve nearly 70% energy reduction at a negligible accuracy loss for most algorithms, after factoring in an average of 5% for data preprocessing overhead. Whereas this study focused on model training, the same group later underscore the gaps in the AI pipeline [85], emphasizing that model training is far less frequent than model inference. One effective way to improve model inference efficiency is an optimizer like *Clover* [87], which uses a mix of high and low quality models, alongside GPU partitioning to maintain high accuracy, to match computational load to available resources. In [83], the authors analyze the inference costs of computer vision (CV) and natural language processing (NLP) models, and

[5] As previously discussed, AI here is a catchall and should be considered to include ML.

TABLE 3. AI power consumption factors.

Ref	Factor	Impact	Limitation
[83]	Number of parameters	Correlated, but severity not as strong as anticipated.	Only computer vision and natural language processing models were evaluated.
[84]	Volume of data points	Approx. 70% energy saving	Focuses on data for model training, not inference.
[85]	Number of inferences	Multiplicative effect on power consumption.	Does not account for the power saved by reducing human effort.
[86]	Model updates	Energy burden from retraining.	Model portability efficiency unexplored.
[5]	Sparsity	Mixture of experts (MoE) up to 10x lower energy consumption.	MoE is not widely explored in wireless communication.

conclude that energy costs of model inference with respect to the number of parameters are not as rapidly increasing as previously thought [3]. They attribute this to improvements in both AI-optimized hardware and also to efforts that are invested for improving algorithmic efficiency in the years after an algorithm is widely deployed.

Improvements in algorithm design such as model pruning and quantization [88] are shown to reduce complexity for fixed energy requirements. Choosing efficient AI models, e.g. opting for sparse models, can reduce computation by up to a factor of ten [5]. Model scalability is improved by training large, sparsely-activated NNs [4], achieving higher accuracy at a lower operational energy footprint. It is worth noting that the model type dictates the start of the inference phase, with supervised models requiring completed training to begin [89], unlike reinforcement learning models. In [86] the authors note that in the wireless network context, significant changes in system state will often require updates to model parameters and the calculation of new solutions which may have a big impact on energy demand. A consideration for future studies might therefore consider the energy impact of the frequency model inference is executed or the conditions which trigger them that give the greatest benefits.

The potential gains from running computation in a cloud, are explored in [5], where experiments show a reduction of computation energy costs by 50% compared to on-premises. The authors attribute this to the massive investment in the custom design and operation of data centers by cloud providers. However, not all cloud data centers are equally efficient [90]. As discussed in Section V-B, the choice of datacenters in the RAN context will be heavily constrained by latency requirements.

The execution of an AI model with greater computational complexity has a higher energy requirement [87], though many other factors (e.g. number of iterations and CPU frequency) may come into play [47]. Algorithmic complexity is not always reported and not reported consistently. Those authors in Section IV who did report algorithmic complexity used Big-O notation, a derivative of FLOPS or a custom cost function to do so. This mixture of reporting adds weight to the assertion in [83] that future studies need detailed and consistent reporting of measures.

A barrier to reporting may be a lack of awareness of tools available. Tools such as *ptflops* [91] or *EAIBench* [92] which calculates an energy consumption benchmark for models, do not yet satisfy the need for a robust and mature tool for energy consumption. In [93], the authors highlight a tool that predicts the energy and carbon footprint of DL models that use hardware acceleration, such as graphics processing units (GPUs). They show that the GPU consumes approximately 50-60% of total energy spent during training, with the remaining energy use being the aggregate of CPU and DRAM. In [94] the authors present a tool to estimate the carbon footprint of a computational task reliably for a variety of hardware. Their *GreenAlgorithms* calculator calculates the carbon footprint as the product of energy needed and carbon intensity.

B. COMPUTATION COSTS FOR NG-RAN

Doing computation in an NG-RAN setting has implications that are different from simply doing a similar computation in a general setting. The 3GPP framework has strict latency requirements for 5G systems [79]. The 3GPP framework places responsibility for data preparation (e.g. cleaning, formatting and transformation) [52], on the inference model. A related 3GPP study [89] on AI management emphasizes the importance of selective data usage and the filtering of low-quality data since excessive, irrelevant data increases storage and processing load. However, the O-RAN Alliance [95] proposes that AI model optimization that is to interact in the Near Real-Time setting must do its computation close (in terms of latency) to the BS. Making computation (AI inference and/or data processing) either co-located with the BS or physically distributed has drawbacks such as loss of the efficiencies of cloud computing in a datacenter (see Section V).

Studies [96] and [97], estimate the computation requirements of BBU functions with respect to parameters such as MCS, signal-to-noise-ratio (SNR) and bandwidth. The authors in [98] argue that BBU processing will significantly impact energy consumption when considering densely deployed small cells with low transmit power in a 5G network as a replacement for high-power macro cells with larger bandwidths. The study uses Landauer's principle [99] to model the computational power of a small cell and macro cell BBU. Simulations reveal the computational power ratio (computation power required over the total power at a base

station) of 5G small cells is more than 50%. However, more recently [24] highlights the trend for integration of baseband (BB) functions into the RU, wherein latency critical parts of BB processing reside in at the radio site and delay tolerant processing may be offloaded to virtualized DUs or CUs.

As previously mentioned, the move to vBSs means we need to understand the power consumed by modeling those functions that are virtualized. A recent study [100] models the computational power needed to provision a virtual RAN. The authors note that the share of BBU processing time is limited to 3 ms per subframe, owing to the standardized hybrid automatic-repeat-request (HARQ) feature. As a result, they argue that virtual resources must be sufficiently allocated to meet this requirement and provide a model to approximate the processing time based on the CPU frequency, number of PRB and MCS index. Based on an O-RAN testbed, they show that a CPU with advanced vector extension support, requires a minimum clock speed of 2 GHz. Looking at the virtualized BS as a whole, the authors in [43] study how SNR, MCS and bandwidth parameters affect CPU power consumption of a general purpose processor (GPP). Observing nonlinear effects of SNR on power consumption, the authors explain that higher noise necessitates more iterations of the turbo decoder to process the signal thereby increasing the computational load.

Lightweight models tailored specifically to NG-RAN needs could reduce the energy consumed by AI models. For example, in [14], the authors use a *block-wise* training approach to reduce the complexity of the path and orientation prediction, with subsequent intelligent reflective surfaces (IRS) angle optimization and beamforming. Simulations show the proposed ANN can reduce transmit power by up to 40% with two IRSs in a system with 10 energy users. Moreover, these results serve as an example of how wise choices in ML design can help to ease the impact on NG-RAN power consumption through efficient computation methods.

C. FUTURE RESEARCH OPPORTUNITIES

The research on energy usage of AI, particularly within the RAN context, highlighted some significant research gaps. It is not using the current literature to compare the energy cost of running AI models for energy efficiency compared with the energy savings created by those models. Studies that use AI for energy reduction do not consistently report complexity and do not take advantage of existing tools that can estimate energy usage. Most give little attention to the trade-off between model complexity (and hence energy consumption) and inference accuracy. One possible solution relies on the report card approach outlined in [101]. This approach has been used to report the carbon emissions with NLP applications leading the reporting.

Most research encountered did not align with industry standards and specifications particularly those outlined by standard bodies like the 3GPP and O-RAN. For example, solutions need to meet certain latency requirements and it was far from clear that this was generally the case. While there was a body of research into the efficiency gains of vBS and optimizations that this makes possible, there is a lack of clarity on the energy requirements from that virtualization.

VI. CONCLUSION

This paper examined the literature related to energy efficiency in next-generation mobile networks, focusing on the RAN. To this end, the power consumption of the RAN is first studied to form a baseline understanding of the power consumption of different functions in the RAN and how these vary with different traffic loads. Energy efficiency in the RAN is often addressed by optimizing RAN resources; this paper proposed four categorization to analyses this process. These are defined based on the location of the degrees of freedom in the optimization process including time, frequency, power, and space. Recent advances and successes of AI have led to a surge in research that employs AI to address these optimization objectives which are often impossible to solve analytically.

The first challenge encountered in this study is the difficulty of conducting a fair and correct comparison of power saving capabilities of the surveyed works. This is due to the lack of reference models and common test scenarios with set parameters that would be used to rank the power-saving capabilities of each of these works.

Another gap found in the literature is the lack or inconsistent reporting of AI complexity, and henceforth power consumption, thus the failure to answer the question of "At what AI power consumption cost is RAN power saving of the proposed method achieved and is worthwhile?"

Despite a recent increase in awareness of the power consumption of general AI, a great research gap remains in its application to RAN energy efficiency optimization and, therefore, in the energy-aware design of AI for RAN. The degrees of freedom in the RAN energy-efficiency problem are limited by rigid requirements from the standards such as latency and quality of service which adds constraints to the design of energy-aware AI.

There is no denying that AI has a critical role to play in the evolution of mobile networks as recognized by standardization groups (e.g., 3GPP and O-RAN) nonetheless, a significant amount of research is needed to bridge the gaps identified in this paper to drive an effective and efficient pathway for this role.

ACKNOWLEDGMENT

The authors thank Keith Briggs for contributions to editing.

REFERENCES

[1] *Mobile Net Zero: State of the Industry on Climate Action 2023*, GSMA, London, U.K., Feb. 2023.
[2] P. Soldati, E. Ghadimi, B. Demirel, Y. Wang, M. Sintorn, and R. Gaigalas, "Approaching AI-native RANs through generalization and scalability of learning," *Ericsson Technol. Rev.*, vol. 2023, no. 3, pp. 2–12, Mar. 2023.
[3] A. de Vries, "The growing energy footprint of artificial intelligence," *Joule*, vol. 7, no. 10, pp. 2191–2194, Oct. 2023.

[4] C.-J. Wu et al., "Sustainable AI: Environmental implications, challenges and opportunities," in *Proc. Mach. Learn. Syst. (MLSys)*, Santa Clara, CA, USA, Jan. 2022, pp. 795–813.

[5] D. Patterson, J. Gonzalez, U. Hölzle, Q. Le, C. Liang, L.-M. Munguia, D. Rothchild, D. R. So, M. Texier, and J. Dean, "The carbon footprint of machine learning training will plateau, then shrink," *Computer*, vol. 55, no. 7, pp. 18–28, Jul. 2022.

[6] R. T. Rodoshi, T. Kim, and W. Choi, "Resource management in cloud radio access network: Conventional and new approaches," *Sensors*, vol. 20, no. 9, p. 2708, May 2020.

[7] D. López-Pérez, A. De Domenico, N. Piovesan, G. Xinli, H. Bao, S. Qitao, and M. Debbah, "A survey on 5G radio access network energy efficiency: Massive MIMO, lean carrier design, sleep modes, and machine learning," *IEEE Commun. Surveys Tuts.*, vol. 24, no. 1, pp. 653–697, 1st Quart., 2022.

[8] B. Brik, K. Boutiba, and A. Ksentini, "Deep learning for B5G open radio access network: Evolution, survey, case studies, and challenges," *IEEE Open J. Commun. Soc.*, vol. 3, pp. 228–250, 2022.

[9] A. I. Abubakar, O. Onireti, Y. Sambo, L. Zhang, G. K. Ragesh, and M. A. Imran, "Energy efficiency of open radio access network: A survey," in *Proc. IEEE 97th Veh. Technol. Conf. (VTC-Spring)*, vol. 13, Florence, Italy, Jun. 2023, pp. 1–7.

[10] L. M. P. Larsen, H. L. Christiansen, S. Ruepp, and M. S. Berger, "Toward greener 5G and beyond radio access networks—A survey," *IEEE Open J. Commun. Soc.*, vol. 4, pp. 768–797, 2023.

[11] B. Mao, F. Tang, Y. Kawamoto, and N. Kato, "AI models for green communications towards 6G," *IEEE Commun. Surveys Tuts.*, vol. 24, no. 1, pp. 210–247, 1st Quart., 2022.

[12] E. García-Martín, C. F. Rodrigues, G. Riley, and H. Grahn, "Estimation of energy consumption in machine learning," *J. Parallel Distrib. Comput.*, vol. 134, pp. 75–88, Dec. 2019.

[13] M. M. Azari, S. Solanki, S. Chatzinotas, O. Kodheli, H. Sallouha, A. Colpaert, J. F. M. Montoya, S. Pollin, A. Haqiqatnejad, A. Mostaani, E. Lagunas, and B. Ottersten, "Evolution of non-terrestrial networks from 5G to 6G: A survey," *IEEE Commun. Surveys Tuts.*, vol. 24, no. 4, pp. 2633–2672, 4th Quart., 2022.

[14] K. W. S. Palitharathna, A. M. Vegni, and H. A. Suraweera, "SLIVER: A SLIPT-enabled IRS-assisted VLC system for energy optimization," in *Proc. IEEE 20th Int. Conf. Mobile Ad Hoc Smart Syst. (MASS)*, Sep. 2023, pp. 143–151.

[15] V. K. Papanikolaou, S. A. Tegos, K. W. S. Palitharathna, P. D. Diamantoulakis, H. A. Suraweera, M.-A. Khalighi, and G. K. Karagiannidis, "Simultaneous lightwave information and power transfer in 6G networks," *IEEE Commun. Mag.*, vol. 62, no. 3, pp. 16–22, Mar. 2024.

[16] S. Wu, C. Chakrabarti, and A. Alkhateeb, "LiDAR-aided mobile blockage prediction in real-world millimeter wave systems," in *Proc. IEEE Wireless Commun. Netw. Conf. (WCNC)*, Austin, TX, USA, Apr. 2022, pp. 2631–2636.

[17] M. Rahim, T. L. Nguyen, G. Kaddoum, and T. N. Do, "Multi-IRS-Aided terahertz networks: Channel modeling and user association with imperfect CSI," *IEEE Open J. Commun. Soc.*, vol. 5, pp. 836–855, 2024.

[18] *NR and NG-RAN Overall Description (Release 17)*, 3GPP, Sophia Antipolis, France, Oct. 2023.

[19] *Architecture Description (Release 17)*, 3GPP, Sophia Antipolis, France, Oct. 2023.

[20] S. T. Arzo, R. Bassoli, F. Granelli, and F. H. P. Fitzek, "Study of virtual network function placement in 5G cloud radio access network," *IEEE Trans. Netw. Service Manage.*, vol. 17, no. 4, pp. 2242–2259, Dec. 2020.

[21] N. Gkatzios, M. Anastasopoulos, A. Tzanakaki, and D. Simeonidou, "Optimized placement of virtualized resources for 5G services exploiting live migration," *Photonic Netw. Commun.*, vol. 40, no. 3, pp. 233–244, Dec. 2020.

[22] L. M. M. Zorello, M. Sodano, S. Troia, and G. Maier, "Power-efficient baseband-function placement in latency-constrained 5G metro access," *IEEE Trans. Green Commun. Netw.*, vol. 6, no. 3, pp. 1683–1696, Sep. 2022.

[23] E. Amiri, N. Wang, M. Shojafar, and R. Tafazolli, "Energy-aware dynamic VNF splitting in O-RAN using deep reinforcement learning," *IEEE Wireless Commun. Lett.*, vol. 12, no. 11, pp. 1891–1895, Nov. 2023.

[24] S. Wesemann, J. Du, and H. Viswanathan, "Energy efficient extreme MIMO: Design goals and directions," *IEEE Commun. Mag.*, vol. 61, no. 10, pp. 132–138, Oct. 2023.

[25] G. Auer, V. Giannini, C. Desset, I. Godor, P. Skillermark, M. Olsson, M. A. Imran, D. Sabella, M. J. Gonzalez, O. Blume, and A. Fehske, "How much energy is needed to run a wireless network?" *IEEE Wireless Commun.*, vol. 18, no. 5, pp. 40–49, Oct. 2011.

[26] E. Björnson, J. Hoydis, and L. Sanguinetti, "Massive MIMO networks: Spectral, energy, and hardware efficiency," *Found. Trends Signal Process.*, vol. 11, nos. 3–4, pp. 154–655, 2017.

[27] H. Holtkamp, G. Auer, V. Giannini, and H. Haas, "A parameterized base station power model," *IEEE Commun. Lett.*, vol. 17, no. 11, pp. 2033–2035, Nov. 2013.

[28] C. Desset, B. Debaillie, and F. Louagie, "Towards a flexible and future-proof power model for cellular base stations," in *Proc. 24th Tyrrhenian Int. Workshop Digit. Commun.-Green ICT (TIWDC)*, Genoa, Italy, Sep. 2013, pp. 1–6.

[29] B. Debaillie, C. Desset, and F. Louagie, "A flexible and future-proof power model for cellular base stations," in *Proc. IEEE 81st Veh. Technol. Conf. (VTC Spring)*, May 2015, pp. 1–7.

[30] E. Björnson, L. Sanguinetti, J. Hoydis, and M. Debbah, "Optimal design of energy-efficient multi-user MIMO systems: Is massive MIMO the answer?" *IEEE Trans. Wireless Commun.*, vol. 14, no. 6, pp. 3059–3075, Jun. 2015.

[31] V. Eramo, M. Listanti, F. G. Lavacca, P. Iovanna, G. Bottari, and F. Ponzini, "Trade-off between power and bandwidth consumption in a reconfigurable Xhaul network architecture," *IEEE Access*, vol. 4, pp. 9053–9065, 2016.

[32] L. M. P. Larsen, S. Ruepp, M. S. Berger, and H. L. Christiansen, "RAN design guidelines for energy efficient 5G mobile Xhaul networks," in *Proc. 14th Int. Conf. Commun. (COMM)*, Bucharest, Romania, Jun. 2022, pp. 1–6.

[33] A. Capone, S. D'Elia, I. Filippini, A. E. C. Redondi, and M. Zangani, "Modeling energy consumption of mobile radio networks: An operator perspective," *IEEE Wireless Commun.*, vol. 24, no. 4, pp. 120–126, Aug. 2017.

[34] N. Piovesan, D. López-Pérez, A. De Domenico, X. Geng, H. Bao, and M. Debbah, "Machine learning and analytical power consumption models for 5G base stations," *IEEE Commun. Mag.*, vol. 60, no. 10, pp. 56–62, Oct. 2022.

[35] J. Huttunen, M. Pärssinen, T. Heikkilä, O. Salmela, J. Manner, and E. Pongracz, "Base station energy use in dense urban and suburban areas," *IEEE Access*, vol. 11, pp. 2863–2874, 2023.

[36] Dappworks. (May 2020). *Front Line Data Study About 5G Power Consumption*. Dappworks. [Online]. Available: https://dappworks.com/front-line-data-study-about-5g-power-consumption-you-need-to-know-about-5g

[37] C. Dongxu. (Jul. 2020). *5G Power: Creating a Green Grid That Slashes Costs, Emissions & Energy Use*. [Online]. Available: https://www.huawei.com/en/huaweitech/publication/89/5g-power-green-grid-slashes-costs-emissions-energy-use

[38] Gartner. (Jan. 2023). *Gartner Says Worldwide PC Shipments Declined 28.5% in Fourth Quarter of 2022 and 16.2% for the Year*. [Online]. Available: https://www.gartner.com/en/newsroom/press-releases/2023-01-11-gartner-says-worldwide-pc-shipments-declined-28-percent-in-fourth-quarter-of-2022-and-16-percent-for-the-year

[39] *ThinkStation P3 Ultra User Guide*, Lenovo, Quarry Bay, Hong Kong, Sep. 2023.

[40] *HPE ProLiant DL110 Gen10 Plus Telco Server Data Sheet*, HPE, Houston, TX, USA, Jan. 2024.

[41] *5G DU SYS-111E-FWTR-1U Specsheet*, Supermicro, San Jose, CA, USA, Jan. 2024.

[42] E. Rodriguez, "DELL PowerEdge XR8000r product environmental compliance," DELL Technol., Round Rock, TX, USA, Tech. Rep., May 2023.

[43] J. A. Ayala-Romero, I. Khalid, A. Garcia-Saavedra, X. Costa-Perez, and G. Iosifidis, "Experimental evaluation of power consumption in virtualized base stations," in *Proc. IEEE Int. Conf. Commun. (ICC)*, Montreal, QC, Canada, Jun. 2021, pp. 1–6.

[44] J. X. Salvat, J. A. Ayala-Romero, L. Zanzi, A. Garcia-Saavedra, and X. Costa-Perez, "Open radio access networks (O-RAN) experimentation platform: Design and datasets," *IEEE Commun. Mag.*, vol. 61, no. 9, pp. 138–144, Sep. 2023.

[45] G. N. Katsaros, R. Tafazolli, and K. Nikitopoulos, "On the power consumption of massive-MIMO, 5G new radio with software-based PHY processing," in *Proc. IEEE Globecom Workshops (GC Wkshps)*, Rio de Janeiro, Brazil, Dec. 2022, pp. 765–770.

[46] S. Matoussi, I. Fajjari, S. Costanzo, N. Aitsaadi, and R. Langar, "5G RAN: Functional split orchestration optimization," *IEEE J. Sel. Areas Commun.*, vol. 38, no. 7, pp. 1448–1463, Jul. 2020.

[47] U. Pawar, A. K. Singh, K. Malde, B. R. Tamma, and A. A. Franklin, "Understanding energy consumption of cloud radio access networks: An experimental study," in *Proc. IEEE 3rd 5G World Forum (5GWF)*, Bangalore, India, Sep. 2020, pp. 407–412.

[48] J. Baliga, R. Ayre, K. Hinton, and R. S. Tucker, "Energy consumption in wired and wireless access networks," *IEEE Commun. Mag.*, vol. 49, no. 6, pp. 70–77, Jun. 2011.

[49] E. Granell, S. Andrade-Morelli, E. Ruiz-Sánchez, and J. Lloret, "Energy consumption study of network access switches to enhance energy distribution," in *Proc. IEEE Globecom Workshops*, Anaheim, CA, USA, Dec. 2012, pp. 1496–1501.

[50] A. Vishwanath, K. Hinton, R. W. A. Ayre, and R. S. Tucker, "Modeling energy consumption in high-capacity routers and switches," *IEEE J. Sel. Areas Commun.*, vol. 32, no. 8, pp. 1524–1532, Aug. 2014.

[51] A. Francini, S. Fortune, T. Klein, and M. Ricca, "A low-cost methodology for profiling the power consumption of network equipment," *IEEE Commun. Mag.*, vol. 53, no. 5, pp. 250–256, May 2015.

[52] "Study on enhancement for data collection for NR and EN-DC," 3GPP, Sophia Antipolis, France, Tech. Rep. 37.817, Apr. 2022.

[53] M. Meo, D. Renga, and Z. Umar, "Advanced sleep modes to comply with delay constraints in energy efficient 5G networks," in *Proc. IEEE 93rd Veh. Technol. Conf. (VTC-Spring)*, Helsinki, Finland, Apr. 2021, pp. 1–7.

[54] Q. Wu, X. Chen, Z. Zhou, L. Chen, and J. Zhang, "Deep reinforcement learning with spatio-temporal traffic forecasting for data-driven base station sleep control," *IEEE/ACM Trans. Netw.*, vol. 29, no. 2, pp. 935–948, Apr. 2021.

[55] A. E. Amine, J.-P. Chaiban, H. A. H. Hassan, P. Dini, L. Nuaymi, and R. Achkar, "Energy optimization with multi-sleeping control in 5G heterogeneous networks using reinforcement learning," *IEEE Trans. Netw. Service Manage.*, vol. 19, no. 4, pp. 4310–4322, Dec. 2022.

[56] S. Malta, P. Pinto, and M. Fernández-Veiga, "Using reinforcement learning to reduce energy consumption of ultra-dense networks with 5G use cases requirements," *IEEE Access*, vol. 11, pp. 5417–5428, 2023.

[57] A. Iqbal, M.-L. Tham, and Y. C. Chang, "Convolutional neural network-based deep Q-network (CNN-DQN) resource management in cloud radio access network," *China Commun.*, vol. 19, no. 10, pp. 129–142, Oct. 2022.

[58] C. Li, W. Chen, and K. B. Letaief, "Joint scheduling of proactive caching and on-demand transmission traffics over shared spectrum," *IEEE Trans. Commun.*, vol. 69, no. 12, pp. 8319–8334, Dec. 2021.

[59] M. Sharara, S. Hoteit, P. Brown, and V. Vèque, "On coordinated scheduling of radio and computing resources in cloud-RAN," *IEEE Trans. Netw. Service Manage.*, vol. 20, no. 3, pp. 2990–3003, Sep. 2023.

[60] C.-C. Hu, "Profit-based algorithm of joint real-time task scheduling and resource allocation in C-RANs," *IEEE Internet Things J.*, vol. 8, no. 2, pp. 941–950, Jan. 2021.

[61] A. M. Kundu and T. V. Sreejith, "Downlink power control in C-RAN enabled full duplex cellular networks," *Phys. Commun.*, vol. 60, Oct. 2023, Art. no. 102154.

[62] D.-N. Nguyen, L. Vallejo, V. Almenar, B. Ortega, P. T. Dat, S. T. Le, J. Bohata, and S. Zvanovec, "Full-duplex transmission of multi-Gb/s subcarrier multiplexing and 5G NR signals in 39 GHz band over fiber and space," *Appl. Opt.*, vol. 61, no. 5, p. 1183, Feb. 2022.

[63] N. Sharma and K. Kumar, "Energy efficient clustering and resource allocation strategy for ultra-dense networks: A machine learning framework," *IEEE Trans. Netw. Service Manage.*, vol. 20, no. 2, pp. 1884–1897, Jun. 2023.

[64] P. Yoon, J. Hong, S. Ahn, Y. Cho, J. Na, and J. Kwak, "ULTIMA: Ultimate balance of centralized and distributed benefits for interference management in 5G cellular networks," *IEEE Access*, vol. 11, pp. 85694–85710, 2023.

[65] F. Marzouk, J. P. Barraca, and A. Radwan, "Interference and QoS-aware resource allocation considering DAS behavior for C-RAN power minimization," *IEEE Can. J. Electr. Comput. Eng.*, vol. 45, no. 4, pp. 442–453, Fall. 2022.

[66] J. Zhou, Y. Sun, R. Chen, and C. Tellambura, "Rate splitting multiple access for multigroup multicast beamforming in cache-enabled C-RAN," *IEEE Trans. Veh. Technol.*, vol. 70, no. 12, pp. 12758–12770, Dec. 2021.

[67] A. Oliveira and T. Vazão, "Towards green machine learning for resource allocation in beyond 5G RAN slicing," *Comput. Netw.*, vol. 233, Sep. 2023, Art. no. 109877.

[68] D. Kim, S. Kwon, H. Jung, and I.-H. Lee, "Deep learning-based resource allocation scheme for heterogeneous NOMA networks," *IEEE Access*, vol. 11, pp. 89423–89432, 2023.

[69] W. Lee, H. Lee, and H.-H. Choi, "Deep learning-based network-wide energy efficiency optimization in ultra-dense small cell networks," *IEEE Trans. Veh. Technol.*, vol. 72, no. 6, pp. 8244–8249, Jun. 2023.

[70] S. Saeidian, S. Tayamon, and E. Ghadimi, "Downlink power control in dense 5G radio access networks through deep reinforcement learning," in *Proc. IEEE Int. Conf. Commun. (ICC)*, Dublin, Ireland, Jun. 2020, pp. 1–6.

[71] A. A. Ahmad, H. Dahrouj, A. Chaaban, T. Y. Al-Naffouri, A. Sezgin, J. S. Shamma, and M.-S. Alouini, "Power minimization using rate splitting with statistical CSI in cloud-radio access networks," *Frontiers Commun. Netw.*, vol. 2, pp. 1–19, Sep. 2021.

[72] P. Frenger and K. W. Helmersson, "Massive MIMO muting using dual-polarized and array-size invariant beamforming," in *Proc. IEEE 93rd Veh. Technol. Conf. (VTC-Spring)*, Helsinki, Finland, Apr. 2021, pp. 1–6.

[73] S. Marwaha, E. A. Jorswieck, D. López-Pérez, X. Geng, and H. Bao, "Spatial and spectral resource allocation for energy-efficient massive MIMO 5G networks," in *Proc. IEEE Int. Conf. Commun. (ICC)*, Seoul, South Korea, May 2022, pp. 135–140.

[74] Y. Liu, X. Fang, and M. Xiao, "Joint transmission reception point selection and resource allocation for energy-efficient millimeter-wave communications," *IEEE Trans. Veh. Technol.*, vol. 70, no. 1, pp. 412–428, Jan. 2021.

[75] Y. Kim, J. Jeong, S. Ahn, J. Kwak, and S. Chong, "Energy and delay guaranteed joint beam and user scheduling policy in 5G CoMP networks," *IEEE Trans. Wireless Commun.*, vol. 21, no. 4, pp. 2742–2756, Apr. 2022.

[76] H.-W. Kim, J.-H. Lee, Y.-H. Choi, Y.-U. Chung, and H. Lee, "Dynamic bandwidth provisioning using ARIMA-based traffic forecasting for mobile Wimax," *Comput. Commun.*, vol. 34, no. 1, pp. 99–106, Jan. 2011.

[77] J. Zhou, Y. Sun, C. Tellambura, and G. Y. Li, "Joint user grouping, sparse beamforming, and subcarrier allocation for D2D underlaid cache-enabled C-RANs with rate splitting," *IEEE Trans. Veh. Technol.*, vol. 71, no. 4, pp. 3792–3806, Apr. 2022.

[78] A. Oliveira and T. Vazão, "Mapping network performance to radio resources," in *Proc. Int. Conf. Inf. Netw. (ICOIN)*. Los Alamitos, CA, USA: IEEE Computer Society, Jan. 2022, pp. 298–303.

[79] *Service Requirements for the 5G System (Release 18)*, document TS 22.261 v18.12.0, 3GPP, Dec. 2023.

[80] "Study on network energy savings for NR," 3GPP, Sophia Antipolis, France, Tech. Rep. 38.864, Dec. 2022.

[81] X. Lin, "An overview of 5G advanced evolution in 3GPP release 18," *IEEE Commun. Standards Mag.*, vol. 6, no. 3, pp. 77–83, Sep. 2022.

[82] S. Amershi, A. Begel, C. Bird, R. DeLine, H. Gall, E. Kamar, N. Nagappan, B. Nushi, and T. Zimmermann, "Software engineering for machine learning: A case study," in *Proc. IEEE/ACM 41st Int. Conf. Softw. Eng., Softw. Eng. Pract. (ICSE-SEIP)*, Montreal, QC, Canada, May 2019, pp. 291–300.

[83] R. Desislavov, F. Martínez-Plumed, and J. Hernández-Orallo, "Trends in AI inference energy consumption: Beyond the performance-vs-parameter laws of deep learning," *Sustain. Comput., Informat. Syst.*, vol. 38, Apr. 2023, Art. no. 100857.

[84] R. Verdecchia, L. Cruz, J. Sallou, M. Lin, J. Wickenden, and E. Hotellier, "Data-centric green AI an exploratory empirical study," in *Proc. Int. Conf. ICT Sustainability (ICT4S)*, Plovdiv, Bulgaria, Jun. 2022, pp. 35–45.

[85] R. Verdecchia, J. Sallou, and L. Cruz, "A systematic review of green AI," *WIREs Data Mining Knowl. Discovery*, vol. 13, no. 4, p. e1507, Jul. 2023.

[86] X. Lin, "An overview of the 3GPP study on artificial intelligence for 5G new radio," 2023, *arXiv:2308.05315*.

[87] B. Li, S. Samsi, V. Gadepally, and D. Tiwari, "Clover: Toward sustainable AI with carbon-aware machine learning inference service," in *Proc. Int. Conf. High Perform. Comput., Netw., Storage Anal.* Denver, CO, USA: ACM, Nov. 2023, pp. 1–15.

[88] M. Zawish, N. Ashraf, R. I. Ansari, and S. Davy, "Energy-aware AI-driven framework for edge-computing-based IoT applications," *IEEE Internet Things J.*, vol. 10, no. 6, pp. 5013–5023, Mar. 2023.

[89] "Study on artificial intelligence / machine learning (AI/ML) management (Release 18)," 3GPP, Sophia Antipolis, France, Tech. Rep. 28.908 v1.2.0, Apr. 2023.

[90] M. Brown. (May 2023). *Digging Into Data Center Efficiency, PUE and the Impact of HCI*. [Online]. Available: https://www.nutanix.dev/2023/05/04/digging-into-data-center-efficiency-pue-and-the-impact-of-hci

[91] V. Sovrasov. (Dec. 2023). *Ptflops: A Flops Counting Tool for Neural Networks in Pytorch Framework*. [Online]. Available: https://github.com/sovrasov/flops-counter.pytorch
[92] F. Zhang, C. Lan, L. Wang, F. Tang, S. Dai, J. Wang, J. Ma, and J. Zhan, "EAIBench: An energy efficiency benchmark for AI training," in *Proc. 14th BenchCouncil Int. Symp.*, A. Gainaru, C. Zhang, and C. Luo, Eds., Nov. 2023, pp. 19–34.
[93] L. F. W. Anthony, B. Kanding, and R. Selvan, "Carbontracker: Tracking and predicting the carbon footprint of training deep learning models," 2020, *arXiv:2007.03051*.
[94] L. Lannelongue, J. Grealey, and M. Inouye, "Green algorithms: Quantifying the carbon footprint of computation," *Adv. Sci.*, vol. 8, no. 12, May 2021, Art. no. 2100707.
[95] *AI/ML Workflow Description and Requirements*, O-RAN Alliance, O-RAN WG2, Jul. 2021.
[96] P. Rost, S. Talarico, and M. C. Valenti, "The complexity–rate tradeoff of centralized radio access networks," *IEEE Trans. Wireless Commun.*, vol. 14, no. 11, pp. 6164–6176, Nov. 2015.
[97] N. Nikaein, "Processing radio access network functions in the cloud: Critical issues and modeling," in *Proc. 6th Int. Workshop Mobile Cloud Comput. Services*. Paris, France: ACM, Sep. 2015, pp. 36–43.
[98] X. Ge, J. Yang, H. Gharavi, and Y. Sun, "Energy efficiency challenges of 5G small cell networks," *IEEE Commun. Mag.*, vol. 55, no. 5, pp. 184–191, May 2017.
[99] R. Landauer, "Irreversibility and heat generation in the computing process," *IBM J. Res. Develop.*, vol. 5, no. 3, pp. 183–191, Jul. 1961.
[100] S. Khatibi, K. Shah, and M. Roshdi, "Modelling of computational resources for 5G RAN," in *Proc. Eur. Conf. Netw. Commun. (EuCNC)*, Ljubljana, Slovenia, Jun. 2018, pp. 1–5.
[101] J. Castaño, S. Martínez-Fernández, X. Franch, and J. Bogner, "Exploring the carbon footprint of hugging face's ML models: A repository mining study," in *Proc. ACM/IEEE Int. Symp. Empirical Softw. Eng. Meas. (ESEM)*, vol. 1, Oct. 2023, pp. 1–12.

KISHAN STHANKIYA (Graduate Student Member, IEEE) is currently pursuing the Eng.D. degree with the Data-Centric Engineering Centre for Doctoral Training and Networks Research Group, Queen Mary University of London. He received the H.E. Certificate in biosciences from King's College London, U.K., in 2015. He was a professionally accredited Infrastructure Consultant with expertise across enterprise projects. His research interests include next-generation radio access, sustainability, and machine learning.

NAGHAM SAEED (Senior Member, IEEE) received the B.S. degree in computer and control and the M.S. degree in mechatronics from the University of Technology, Baghdad, Iraq, in 1992 and 1999, respectively, and the Ph.D. degree from the Wireless Networks and Communications Centre, Brunel University, London, U.K., in 2011. Her Ph.D. research was optimizing mobile ad hoc wireless communication networks by introducing an intelligent mobile ad hoc network system based on AI. She is currently an Associate Professor in electrical and electronic engineering and the Head of the Industrial Internet of Things Research Group, University of West London. Her research interests include expert systems for smart cities, wherein she applies AI algorithms to drive modeling and optimization. She is recognized as a C.Eng. by the Engineering Council, a member of the IET, the Vice Chair of the IEEE U.K. and Ireland Section (2024–2025), the Elect-Chair of the IEEE U.K. and Ireland Section (2026–2027), and the Past Chair of the IEEE Women in Engineering U.K. and Ireland Section (2023).

GREG MCSORLEY received the B.A. degree in ancient history and archaeology from the University of Manchester, U.K., in 2012, and the M.S. degree in geospatial information systems from Ulster University, Belfast, in 2022. Previously, his roles include management, archaeology, and data analytics. He joined BT Research, in 2022, where he is currently an AI and Sustainability Researcher. His research interests include spatial analysis and net-zero initiatives.

MONA JABER (Senior Member, IEEE) received the B.E. degree in computer and communications engineering and the M.E. degree in electrical and computer engineering from the American University of Beirut, Lebanon, in 1996 and 2014, respectively, and the Ph.D. degree from the 5G Innovation Centre, University of Surrey, in 2017. Her Ph.D. research was on 5G backhaul innovations. She was a Telecommunication Consultant in various international firms with a focus on the radio design of cellular networks, including GSM, GPRS, 3G, and 4G. She led the IoT Research Group, Fujitsu Laboratories of Europe, from 2017 to 2019, where she researched IoT-driven solutions for the automotive industry. She is currently a Lecturer in IoT with the School of Electronic Engineering and Computer Science, Queen Mary University of London. Her research interests include zero-touch networks, the intersection of ML and IoT in the context of sustainable development goals, and IoT-driven digital twins.

RICHARD G. CLEGG is a Senior Lecturer in networks with the School of Electronic Engineering and Computer Science, Queen Mary University of London. He is a Research Lead with Pometry company, that develops software for temporal networks. His research interests include complex networks and the statistics of network measurements.