The Two Faces of AI in Green Mobile Computing: A Literature Review

Siemers, Wander ; Sallou, June; Cruz, Luis

DOI
10.1109/SEAA60479.2023.00053

Publication date
2023

Document Version
Final published version

Published in
Proceedings of the 2023 49th Euromicro Conference on Software Engineering and Advanced Applications (SEAA)

2023 49th Euromicro Conference on Software Engineering and Advanced Applications (SEAA) | 979-8-3503-4235-2/23/$31.00 ©2023 IEEE | DOI: 10.1109/SEAA60479.2023.00053

The Two Faces of AI in Green Mobile Computing: A Literature Review

Wander Siemers*, June Sallou*, Luís Cruz*
*Delft University of Technology, The Netherlands - wandersiemers@me.com, { j.sallou, l.cruz }@tudelft.nl

***Abstract*—Artificial intelligence is bringing ever new functionalities to the realm of mobile devices that are now considered essential (e.g., camera and voice assistants, recommender systems). Yet, operating artificial intelligence takes up a substantial amount of energy. However, artificial intelligence is also being used to enable more energy-efficient solutions for mobile systems. Hence, artificial intelligence has two faces in that regard, it is both a key enabler of desired (efficient) mobile functionalities and a major power draw on these devices, playing a part in both the solution and the problem. In this paper, we present a review of the literature of the past decade on the usage of artificial intelligence within the realm of green mobile computing. From the analysis of 34 papers, we highlight the emerging patterns and map the field into 13 main topics that are summarized in details.**

Our results showcase that the field is slowly increasing in the past years, more specifically, since 2019. Regarding the double impact AI has on the mobile energy consumption, the energy consumption of AI-based mobile systems is under-studied in comparison to the usage of AI for energy-efficient mobile computing, and we argue for more exploratory studies in that direction. We observe that although most studies are framed as solution papers (94%), the large majority do not make those solutions publicly available to the community. Moreover, we also show that most contributions are purely academic (*28 out of 34 papers*) and that we need to promote the involvement of the mobile software industry in this field.

***Index Terms*—mobile software, energy consumption, artificial intelligence**

I. Introduction

Artificial intelligence (AI) is bringing ever more new functionalities to the realm of mobile devices that are now considered essential (e.g., camera and voice assistants, recommender systems). Users have increasing expectations of the processing power and capabilities of their mobile devices. Contemporary smartphones have highly advanced image processing systems, smart integrated assistants, and offer gigabit speeds over their radios. All of these features, and many more, are enabled by artificial intelligence [1], [2], [28].

Yet, operating artificial intelligence takes up a substantial amount of energy. Modern artificial intelligence techniques, such as deep learning, can have very high energy consumption, both on dedicated servers [3] and on mobile devices [4].

Beyond the realm of artificial intelligence, mobile computing has long been concerned with energy efficiency due to the limited power capacity of smartphones [5], [6]. Less obviously, artificial intelligence techniques themselves can also be used to reduce mobile energy consumption. For example, by optimizing data transmission [28], or location services [29]. Hence, artificial intelligence has two faces in this problem: it is both 1) a key enabler of desired (efficient) mobile features and 2) a major power draw on these devices, playing a part in both the solution and the problem. In this paper, we provide a comprehensive overview of both of these aspects of mobile energy use and artificial intelligence by reviewing the associated literature. The goal of this review is to understand the characteristics of the literature on mobile energy consumption involving artificial intelligence.

Our literature review yields 34 papers from 2013 until late 2022. We identify and pinpoint thirteen different topics being addressed by the literature. Our results showcase a growing interest in the intersection between AI and Mobile Computing Energy since 2019. Most studies revolve around solution papers (***32** out of 34 papers*) and only ***6** out of 34 papers* display the participation of authors with an industry affiliation. We argue that it is quintessential that contributions in this field come with a replication package and that proposed solutions are made available to the public. Finally, although topics such as Approximate Computing and Benchmarking have been marginally covered by the literature (***2** out of 34 papers*), we expect them to be relevant to the challenges posed in this field.

The contributions of this paper are three-fold:

- An analysis of the field of AI in Green Mobile Computing, covering publications per year, study type, industry involvement, level of study, and tool provision.
- A mapping of the field into different topics, with the respective summary of existing contributions.
- A replication package that provides all the collected data for each paper, that can be used in future reviews.[1]

The remainder of this paper is structured as follows. We describe the detailed methodology in Section II. Following this methodology, we present our results in section III. We then discuss these results and their related impacts in the research community in Section IV. The threats to validity of our study can be found in Section V. We then treat related work in section VI and discuss how our work differs from those studies. Lastly, we highlight the conclusions of the literature review in section VII.

II. Methodology

In this section, we present the methodology we rigorously followed while carrying out this review. We follow the guidelines for conducting literature reviews in software engineering research presented by several dedicated publications [7]–[10].

[1]Replication package: https://zenodo.org/record/8172245

2376-9521/23/$31.00 ©2023 IEEE
DOI 10.1109/SEAA60479.2023.00053

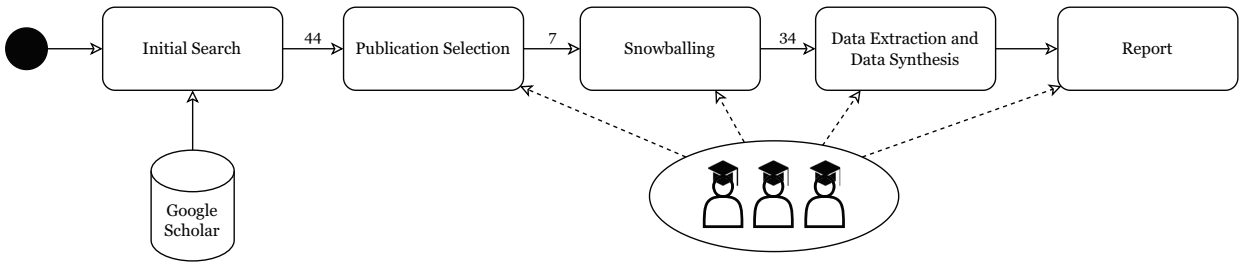

Fig. 1: Process overview of the systematic literature review.

A. Research Goal and Question

The goal of this review is to understand the characteristics of the literature on mobile energy consumption involving artificial intelligence. It can be directly translated into a research question (RQ), which states as follows:

RQ. *What are the characteristics of the state-of-the-art research regarding Artificial Intelligence in Green Mobile Software?*

We are interested in determining the trends in publishing in this field (e.g., the evolution of the number of publications, the study environment, and tool provision), as well as the trends in the studies themselves (e.g., topics, types of AI usage).

B. Research Process

The different steps of the research process are illustrated in Figure 1. It begins with an initial search conducted through an automatic query on the Google Scholar database, which is then augmented by a bidirectional snowballing process. The subsequent sections provide a detailed description of each step.

1) Initial Search: To gather an initial set of publications, we define a search query that is executed on the *Google Scholar* database. The query states as follows:

Listing 1: Search query

```
("AI-Based" OR "machine learning" OR
"Artificial Intelligence") AND "mobile"
AND ("energy" OR "efficient" OR "green")
```

The query is designed to retrieve publications whose titles match keywords related to the concepts of AI (*i.e., AI-Based*, or *machine learning*, or *Artificial Intelligence*), mobile computing (*i.e., mobile*), and energy efficiency (*i.e., energy efficient*, or *green*). We focus on titles to target literature whose main topics deal with green mobile computing and AI, restricting the inital pool of results but attempting to increase the relevance of the matches. We perform this search using the *Publish or Perish*[2] software, searching on *Google Scholar* and matching only on title keywords (*i.e.,* by using the "title words" option). We select *Google Scholar* for several reasons: (i) it indexes results from various libraries, allowing us to perform only a single search, (ii) as reported in the set of guidelines by Wohlin [8], the use of such an indexer is a suitable choice for identifying the initial set of literature for snowballing processes, (iii) the query results can be automatically extracted from the indexer using Publish or Perish.

The initial search was performed on the October 19th, 2022, and provided 44 results using Publish or Perish[3].

2) Publication Selection: The initial set of publications undergoes a selection process based on predefined selection criteria. Inclusion (**I**) and exclusion (**E**) criteria are established to ensure the included papers in the primary study set are not only relevant to our research question but also of high quality.

- **I-1.** The study regards mobile devices (smartphones/tablets)
- **I-2.** The study regards energy consumption
- **I-3.** The study regards artificial intelligence[4]: either by treating how AI can be used to reduce mobile energy consumption or by treating the energy consumption of mobile AI itself
- **I-4.** The study regards the software level
- **E-1.** The study is not written in English
- **E-2.** The study is not accessible
- **E-3.** The study is not peer-reviewed
- **E-4.** The study is in the form of citations, patents, editorials, tutorials, books, extended abstracts, thesis, etc.
- **E-5.** The study is not a primary study, such as a review paper.
- **E-6.** The study was published before 2012

The first three inclusion criteria ensure that the selected papers deal with the topic of artificial intelligence in mobile energy consumption (**I-1** to **I-3**). The fourth criterion (**I-4**) guarantees that the studies focus on the software level. The goal is to exclude papers that focus exclusively on hardware-specific approaches (*e.g.,* involving peculiar hardware components to improve the mobile energy consumption while using AI). For the exclusion criteria, the first two criteria (**E-1** and **E-2**) assure that we can extract data from the papers. The two following criteria (**E-3** to **E-5**) make sure the papers constitute scientific primary studies. We include an additional criterion, **E-6**, to exclude papers published before 2012. This decision is based on the rapid changes in both AI and mobile technology

[2]Publish or Perish software, https://harzing.com/resources/publish-or-perish

[3]Performing the query with Publish or Perish is the equivalent of the Advanced Google Scholar Search with the query "*allintitle: mobile energy OR efficient OR green "AI Based" OR "machine learning" OR "Artificial Intelligence*"".

[4]We have used a broad interpretation of the concept 'AI', including papers proposing simple techniques like regression.

over the past decade. Hence, we believe that studies before that date are not particularly relevant, especially when considering the relatively new trend of energy efficiency in the topic.

The publications in the initial set are distributed among the three authors for assessment based on the selection criteria. Each author independently evaluates the assigned publications using adaptive reading depth [9]. Moreover, regular meetings are conducted to facilitate discussions regarding the selection process and to minimize any potential personal biases.

After applying the in/exclusion criteria to the 44 retrieved papers of the initial set, we identify 7 primary studies.

3) Snowballing: To address the limitations of our initial query and ensure a comprehensive representation of the literature on AI and mobile computing energy, we employ a bidirectional snowballing technique. This approach aims to supplement the primary studies by retrieving papers that may have been missed by the title-only search query. Following the guidelines presented by Wohlin [8], we conduct a single iteration of snowballing, encompassing both backward and forward passes for the papers in the initial set. During this process, we thoroughly examine all studies that either cite (forward snowballing) or are cited (backward snowballing) by the primary studies already included in our analysis.

To maintain objectivity and rigour, the three authors independently explore different primary studies, identifying additional studies that align with the predefined selection criteria for inclusion. Throughout this exploration, any doubts or disagreements that arise during the assessment are addressed through discussions among the authors. These discussions serve as a mechanism to mitigate subjective biases and ensure a collective resolution for all assessments.

Finally, the snowballing process results in the addition of 27 new primary studies, leading to a total of 34 primary studies, which are examined in this literature review.

4) Data Extraction: Once the final set of primary studies is completed, we proceed to the systematic data extraction step to answer our research question (cf. **RQ** in Section II-A).

To establish the specific data fields to be extracted from the primary studies, the authors independently review them initially and annotate the relevant characteristics that address the research question. These characteristics are subsequently subject to thorough discussions and refinement through open coding techniques [10], [11]. By engaging in open coding, the authors ensure a more precise identification and categorization of the study attributes. This iterative process allows for the final determination of the fields to be utilized during the subsequent data extraction phase.

Once the data fields are finalized, the authors re-examine the papers meticulously, rigorously analysing them to extract the data corresponding to the identified fields of interest. This data extraction process involves a comprehensive and systematic approach, ensuring that the desired information is accurately captured from the primary studies.

The fields used for the data extraction are the following:

- **Publication Year**
- **Study Type:** The type of study the paper is presenting: either a *position* on AI in Green Mobile Software, a *solution* to tackle an issue on the topic, or an *observational* study;
- **Category of AI Role:** The role AI has regarding Green Mobile Computing. It can either be the use of AI for improving the energy efficiency of mobile computing, or the study of energy consumption of AI-based mobile systems;
- **Topic:** The topic the primary study is focusing on. For instance, context adaptation, in which the mobile software execution is readjusted according to the context, to improve the energy efficiency;
- **Level of Study:** It corresponds to the scale at which the mobile software is studied (either at the level of the *device*, or of the *system*);
- **Industry Involvement:** The involvement of industry in the authoring of the study, which can be either exclusively academic, exclusively industrial or a mix;
- **Tool Provision:** The availability of the tool(s) to handle AI in Green Mobile Computing presented in the study (if applicable).

5) Data Analysis: Along the whole process, the authors discussed the emerging codes generated during data extraction to ensure that they are congruent with each other, meet the research objective and answer the question being addressed.

Regarding the field of **Topic**, the approach of *open coding* was used to group the different keywords into a coherent hierarchical structure [10], [11]. For the rest of the fields, the keywords were already pre-set in advance. In the case of **Study Type**, the different options were *position*, *observational*, and *solution*. As for **Category of AI Role**, they were *AI4E*, to translate the use of AI to make the mobile software more energy efficient, and *EofAI*, for the fact that AI was involved in the design of the mobile software itself, and was studied regarding energy efficiency. The **Level of Study** field was given the keywords *Device* and *System*, to represent the scale at which the study was focusing. The **Industry Involvement** was based on the terms *Academic* for academic authorship only, *Industrial* for industrial authorship only, and *Mix* for mixed authorship. In the case of **Tool Provision**, the options were reduced to either *Yes* and *No*. Finally, for the **Publication Year**, the year was directly extracted from the publication date.

III. Results

In this section, we present the results collected with our systematic literature review on the roles of artificial intelligence in mobile energy efficiency.

A. Publication Years

For the past decade, we can observe a trend of an increase in the number of papers dealing with Artificial Intelligence with respect to mobile energy consumption. Figure 2 shows an apparent increase from year 2019, going from approximatively 2 papers per year before 2019, to 6 papers being published per year after 2019. As the number of papers remains small, that

increase needs to be interpreted with caution. Furthermore, it should be noted that the results for the year 2022 do not correspond to the full year and may not be representative of the actual research output, as the automated initial search was executed in October 2022.

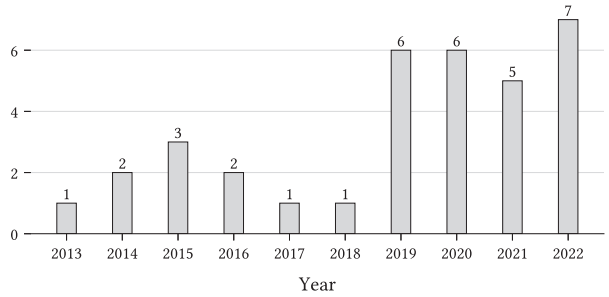

Fig. 2: Publication of papers per year.

> ***Publication Trends:*** *The research topic of artificial intelligence in green mobile software has been gaining in popularity since 2019, and appears to be on track to continue doing so.*

B. Types of Study

The literature on the mobile energy consumption related to artificial intelligence from the last decade is predominantly composed of *solution* papers with ***32*** *out of 34 papers*. We can notice that, in comparison, there are only 2 *observational* papers and no *position* papers. Note that study types are mutually exclusive, i.e., a single paper has only one study type.

> ***Type of Study:*** *The majority of the literature consists of studies proposing solutions, with a very small number of observational studies.*

C. Category of AI Role

Around 68% of the papers (*i.e.,* ***23*** *out of 34 papers*) deal with the use of artificial intelligence to tackle the mobile energy consumption (cf. *AI4E* in Figure 3), while the rest (*i.e.,* ***11*** *out of 34 papers*) studies the energy consumption of the use of artificial intelligence in mobile devices (cf. *EofAI* in Figure 3).

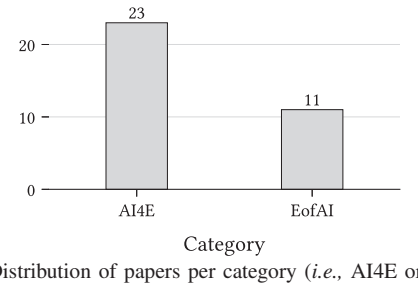

Fig. 3: Distribution of papers per category (*i.e.,* AI4E or EofAI).

> ***Category of AI Role:*** *The use of AI to make mobile software more energy-efficient is being studied twice as much as the energy consumption of AI-based mobile software.*

D. Topics

We identified 13 different topics that are covered by the literature on the use of artificial intelligence in mobile in the context of energy efficiency. Figure 4 presents the distribution of the studies according to the topics. The most popular topic is *Offloading*, addressed by 9 papers, followed by *Context Adaptation* (8), and *Federated Learning* (7). Since papers are not exclusive to a single topic, these top-3 topics alone cover 70% of the papers in this review. Below, we pinpoint each topic with a short summary and the respective number of publications.

Offloading: ***9*** *out of 34 papers*. Approach to delegate the execution of a resource-intensive task from a lower-powered device, often a mobile device, to a different, usually more powerful, device or service [12]. Frequently, a mobile device offloads to the cloud, in which case the technique is also called *Mobile Cloud Computing* (MCC) [30]–[32]. The papers covering this topic study how to optimise the offloading process by employing AI. For instance, Markov Decision Processes (MDPs) are used to optimize offloading schedules in [30], [33]. The objective of using these MDPs is to minimize the long-term cost, in which energy expended is included. Both traditional value iteration [30] and approaches [33] using a Deep Q-Network (DQN) (first proposed in [14]) have been used to determine policies efficiently. In particular, Zhang *et al.* [30] propose to tackle the problem of intermittence in connection, when the user mobile is out of communication range with the offloading cloud) system.

Context Adaptation: ***8*** *out of 34 papers*. The goal of the primary studies involving the topic is to improve the energy efficiency of mobile software by readjusting its execution according to the context. Some papers are using the user of the mobile device as their reference for the definition of the context, while others deal with the device itself as the reference for context of execution.

For instance, Machidon *et al.* [34] propose to optimise the resolution of video depending on (user) context. The authors observe that the personality traits of the user and the user's motion affect the perception of mobile video. The authors then use this observation to build a predictor for the desired resolution of a mobile video. Both a Random Forest and Mean Regressor are applied, with the Random Forest being the most accurate. However, the energy saved is not quantified, and importantly, the energy use of the predictor itself is not measured.

Regarding the mobile context perspective, Donohoo *et al.* [29] exploit the spatio-temporal and device context and use AI to predict and adapt device wireless data and location interface configurations so that energy consumption in mobile devices is optimised.

Nawrock *et al.* [35] focus on the context of the device in a heterogeneous environment to provide users with customized recommendations of products or services through the use of recommender systems (based on the use of AI).

Federated Learning: ***7*** *out of 34 papers*. The application of AI to train a model across several decentralised devices with their own local subset of data, without the necessity of sharing data to all the involved devices.

The associated papers are looking at improving the communication and training execution strategies to improve the overall energy efficiency of such mobile software.

For example, a custom distributed learning system is provided by Deng *et al.* [36]. The authors propose multiple algorithms to schedule training efficiently while allowing setting limits on the energy consumption of the clients. It reaches higher accuracy scores than its baselines.

Shahidinejad *et al.* [37] considers a network full of devices, in which offloading decisions are interdependent. Each device learns using Deep Reinforcement Learning and shares its results with peers. The authors report a modest (2%) energy consumption improvement compared to not offloading.

Accuracy-Energy Trade-Off: ***5*** *out of 34 papers*. The highest accuracy in AI-based software can lead to high energy consumption. Trading off some accuracy for better energy performances is a means to make mobile software greener.

For instance, Zheng *et al.* [38] identify the logical trade-off between energy use and accuracy applied to Federated Learning: selecting the clients of the network with the largest data sets, the accuracy is increased, but the energy use grows accordingly since more training time and computation are needed on a larger data set. The authors use POMPDP, a deep reinforcement learning technique, to optimize the ratio between training accuracy and energy consumption. They report a 51.8% improvement in the ratio between accuracy and energy consumption.

Other application domains are explored, such as cellular networks. In that direction, Ruiz *et al.* [39] use Discontinuous Reception (DRX), an energy-saving technique applied in 3G and 4G networks that turns off cellular hardware regularly to reduce energy consumption. The authors consider DRX on voice communications traffic. Using the fact that human speech contains periods of silence, the authors propose employing a Gaussian Process (GP) to optimize the duration for which cellular radios can remain off without interrupting speech. Energy savings of up to 30% can be achieved compared to basic DRX schemes.

Scheduling: ***5*** *out of 34 papers*. The scheduling, when the execution of a task is performed, can impact the energy efficiency of the mobile software.

Multiple solutions for learning optimal training schedules exist. Deep Q-Networks (DQN) can be used to learn *resource budgets* for mobile clients [40], [41] to deal with the constraints of mobile devices. This solution does not require any advance knowledge of network dynamics. However, DQN can suffer from a problem called *over-optimistic value estimation* because it uses the same Q-value for selecting and evaluating an action [13]. Therefore, Anh *et al.* [40] use a variant of DQN, called DQNN, which uses two neural networks to circumvent this problem. Huang *et al.* [42] propose a very similar approach with two neural networks but runs it on the client instead of the server.

Monitoring: ***4*** *out of 34 papers*. Covering monitoring approaches to study the energy efficiency of mobile software.

Papers provide solutions to profile of energy consumption during runtime.

In a first study, Aggarwal *et al.* [43] rely on dynamic analysis of test cases to estimate the fluctuation of power use caused by modifications in the software. More specifically, they use the traces of system calls during application execution to gather metrics while software is undergoing a use case test. In a second paper, Aggarwal *et al.* [44] provide a tool called *GreenAdvisor* to help mobile software practitioner with analysing the energy profile of their code, and to predict the impact of code changes on that energy profile.

With the study made by Pandey *et al.* [45], the different tasks executed in a network of mobile devices are profiled with regard to resource utilisation. AI is then used to analyse the profiles and to perform the best match between task to be executed and devices in order to improve the energy efficiency of the tasks.

Finally, Wang *et al.* [46] address the energy profiling of mobile software involving augmented reality. The authors measure the energy consumption of a mobile Convolutional Neural Network (CNN) for computer vision and compare it to the energy consumption of an offloaded version of the model. They conclude that some light models can be run on recently released smartphones with low latency. Latency might even be lower than offloading the model to a remote server.

Deployment: ***2*** *out of 34 papers*. Dealing with how the mobile software is being deployed.

Li *et al.* [47] propose to employ AI to support intelligent network resource management strategies in mobile cloud computing to optimise the resource allocation.

In contrast, Tang *et al.* [60] address tail energy in their study. Tail energy is the energy that a cellular interface uses after data transmission has ended because it remains in a high-power state. This accounts for a significant fraction of total data transmission energy. The pattern of data transmission may be learned by artificial intelligence to allow the cellular interface to enter a low-power state sooner. Instead of training a model on-device, which might suffer from high resource use and the cold start problem, the system in [60] uses a client-server architecture. Mobile devices, therefore, send transmission records to a central server which trains an MLP classifier with two classes: 1) *can be delayed* and 2) *should be transmitted now*. Energy usage can be reduced by 20% compared to always immediately sending requests.

Other: ***6*** *out of 34 papers*. Studies addressing a relevant topic with only a single publication in total: Benchmarking [48], Approximate Computing [34], Model Design [49], Recommenders [50], Energy Measurement [51], and Resource Management [47].

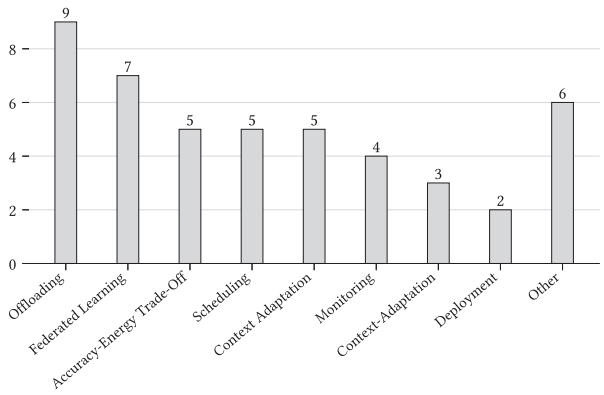

Fig. 4: Topics addressed in the studies.

Topics: *The literature on AI in Green Mobile Software covers 13 main topics. It focuses mainly on Offloading, Context Adaptation and Federated Learning, representing 70% of the papers. In contrast, the following topics are currently under-explored: Benchmark, Approximate Computing, Model Design, Recommenders, Energy Measurement, and Resource Management.*

E. Level of Study

The mobile energy consumption can be addressed at different levels depending on the context of study. The majority of the papers (***22*** *out of 34 papers*) focus their studies at the system level. Rather than the energy consumption of a single mobile device, it is the consumption of a set of devices in a more complex setting, such as a network, and related tasks (*e.g.,* transmission of data). The rest of the papers, around a third of them (***12*** *out of 34 papers*), tackle energy-consuming (AI) tasks executed on a single mobile device.

Level of Study: *Approximately two thirds of the literature focus on Green Mobile software at the system-level, compared to the device-level. More attention is given to the topic in the context of a network of mobile devices.*

F. Industry Involvement

Regarding the industry involvement in the publications on artificial intelligence in mobile energy consumption (cf. Section II), an overview of the authorship of the primary studies is rendered in Figure 5.

From the figure, we can notice that the vast majority of the papers are authored by academic researchers (***28*** *out of 34 papers*), while 6 papers have authors being a mix of researchers with an academic and industrial background (cf. *Mix* in Figure 5). Additionally, we observe that there is no publication being exclusively authored by industrial researchers.

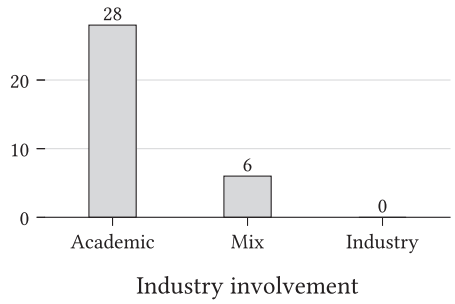

Fig. 5: Distribution of papers according to the type of industry involvement.

Industry Involvement: *The presence of industrial researchers among the authors is still scarce.*

G. Tool Provision

From the set of primary studies, only very few papers (***2*** *out of 34 papers*) provide available tools to address the mobile energy consumption involving artificial intelligence. *GreenAdvisor*, a Java-based tool, is made available by Aggarwal *et al.* [44]. It advises Android application developers on change in energy consumption of their app with change in app code, based on system call tracing[5]. Nawrocki *et al.* [31] provide a TypeScript library for Mobile Cloud Computing code offloading with Machine Learning[6].

Tool Provision: *Although many studies provide solutions to address mobile energy consumption involving AI, only a small portion of them make the solution-based tools readily available online.*

IV. Discussion

A clear picture emerges from the analysis of publishing trends on mobile energy consumption involving AI. As shown in Section III-A, a recent increase in the number of papers published on the topic can be observed. We can speculate on a few reasons for this. AI in general has seen a great increase in interest over the past few years, driven by major technical developments in deep learning, such as shown by Mnih *et al.* in reinforcement learning [14], Goodfellow *et al.* in the development of the Generative Adversarial Network (GAN) [15], and the recent development of models like Minerva, which can solve undergraduate-level mathematical and engineering problems [16]. Secondly, AI on mobile devices has become more feasible both due to generic improvements to hardware and software, but especially due to the inclusion of so-called accelerators, such as Apple's Neural Engine [17] and Google's TPU [18]. Finally, energy consumption and efficiency are becoming increasingly important across all sectors due to ever-escalating global warming. Therefore, as this issue becomes

[5]https://github.com/kaggarwal/GreenAdvisor
[6]https://github.com/Hoobie/ml-offloading

more pressing, one can expect an increase in the interest in energy efficiency improvements in technology.

Considering the explored topics, in general, a fair body of research on artificial intelligence and mobile energy use exists. However, the works in this body are often highly specific to a specific technology, such as *Federated Learning*. However, even Federated Learning itself is partly an accounting trick: although the model is only trained partially on each local device, thereby reducing energy consumption per training device, this energy use is simply spread over multiple devices. Accounting for the communication energy required by each device to get the model from the server and send it back, a net training energy consumption calculation might turn out poorly for federated learning.

We also noticed that having access to up-to-date benchmarks is a major challenge in this field. For example, Deng *et al.* [36] was published in 2022 but uses baselines from 2016 and 2018. This can be an indicator that the field is still evolving rapidly, or that it is more difficult to publish such studies when involving AI in Green Mobile Software.

Regarding the two faces of AI and its involvement in mobile energy consumption, it can be noted that the number of papers in this review on the topic of using AI to reduce energy consumption was twice as high as the number of papers on reducing the energy consumption of AI itself. All this work cited or was cited by Anh *et al.* [40] or McIntosh *et al.* [51], both of which were published years ago already. In general, a lack of broad research on this topic can be observed, except for papers addressing Federated Learning energy use. However, as discussed earlier in the discussion, some of the benefits of Federated Learning can be seen as an accounting trick. The body of work focused on using AI to reduce energy consumption is extensive, but it remains confined to a relatively limited range of research topics. Networking optimization and offloading techniques appear to be prominent areas with intense competition driving innovative approaches. However, other areas, such as quantifying the impact of code changes in mobile apps or optimizing video playback, have received less exploration. Future research in these areas could contribute significantly to further reducing mobile energy consumption, which is crucial given the projected growth in the number of mobile devices in the coming years.

Thanks to the findings of this review, it is evident that the systematic provision of tools addressing AI and Green Mobile Software is far from optimal. While many studies propose solutions to address mobile energy consumption using AI, only a small portion of them make their solution-based tools readily available online. This observation suggests two potential explanations: (i) the rapid evolution of research in this field, which leads to quickly outdated results and renders tools less meaningful, or (ii) an immaturity in the research field that requires stronger empirical support as a foundation for the development of tools.

Finally, from our results, the literature seems to be mainly involving the academic community, as the majority of the primary studies presents an academic authorship only. Again, although there is a predominance of solutions papers, only a very few industrial researchers are authors of papers dealing with AI in Green Mobile Software. This can be related to the fact that few tools are made available online. We argue that, especially as mobile devices are getting more and more prevalent, to have a real impact on the energy efficiency of mobile software, and to update the related practises, we need to promote the involvement of the mobile software industry in this field.

V. Threats to Validity

In this section, we discuss the threats to validity of our research. To ensure the quality of the results, we established a well-defined research protocol before performing the data collection. In addition, we adhered to a set of well-accepted snowballing guidelines for literature reviews [7]–[10]. To lower potential sources of bias, crucial considerations that emerged during the research were discussed among the authors. Despite adhering to a systematic literature review approach, potential threats to validity remain. The remainder of this section addresses four types of validity: internal, external, construct, and conclusion validity.

1) Internal validity: To address internal validity threats, we applied a systematic snowballing and open coding process. We used existing guidelines, such as [8], [10], [11], to avoid inventing new methods and prevent us from making non-obvious but consequential methodological mistakes. However, even though the choices were discussed among the authors, some bias in paper selection and coding might remain.

2) External validity: The main threat to external validity of this study is the lack of representativeness of the papers considered. We think this threat can manifest in three main ways: 1) The primary studies, found using an academic search engine, are not representative of the topic we attempt to study. We mitigate this threat by using Google Scholar, which searches very broadly among publishers of literature. 2) The set of studies is incomplete. We mitigate this threat by performing snowballing, both forward and backward. However, no snowballing process can guarantee a complete set of literature, and only one iteration was performed. 3) The potential non-relevance of the papers. We have mitigated this threat by using in- and exclusion criteria that address quality, such as the requirement of being peer-reviewed.

3) Construct validity: To ensure the relevance of the primary papers in addressing our research questions, we applied meticulously crafted inclusion and exclusion criteria. Subsequently, bidirectional snowballing was employed to expand the set of relevant primary papers. However, we conducted only one iteration of snowballing, and further iterations could have potentially resulted in the inclusion of additional papers.

4) Conclusion validity: Sources of bias from our analyses are addressed by following a strict and clearly defined process based on public guidelines [8], [10], [11]. Lastly, we documented all the data throughout the whole review process and made them available through a replication package for reproducibility and replicability purposes (cf. Section 1).

VI. Related work

Several secondary studies have examined related topics. We briefly discuss them here.

In a recent systematic review [19], the focus is on Green AI, specifically the energy consumption of AI itself. Interestingly, the authors found that studies on mobile computing comprise only a small fraction (around 4%) of the Green AI literature. Our review goes beyond mobile computing to explore how AI can reduce mobile energy consumption.

Measurement methods for mobile energy are discussed by Khan *et al.* [20], where 21 techniques (hardware and software-based) are described and compared. While previous work [21] covers similar ground, it was published in 2015, analysing a different period. Some techniques mentioned in these papers have been used to measure energy consumption in certain AI applications, but they may not be applicable to other AI systems like federated learning, which we focus on here.

General energy management for mobile devices is addressed in [22]. Other reviews cover specific areas like video streaming [23] and processing unit management [24]. In contrast, our work specifically examines AI's role in energy management within mobile software.

Studies on energy consumption in mobile cloud computing (MCC) [12] and [25] compare MCC frameworks and review multimedia application papers, respectively. However, our focus is exclusively on works that discuss AI applications.

Shi *et al.* [26] explore mobile edge AI techniques, including model compression, federated learning, and offloading. While our work intersects with this topic, we emphasize using AI to save energy rather than reducing AI's energy use.

Federated learning is extensively surveyed in [27], with a focus on communication costs in 5G networks. Our work intersects partially, considering the use of these AI techniques to save energy.

VII. Conclusion

In this paper, we provide an overview of the literature on mobile energy consumption and artificial intelligence. Our research questions reflect on the publication trends in this area and the characteristics of 34 papers. We describe two main branches in the literature: 1) papers looking at the energy consumption of mobile AI applications and 2) papers focusing on applying AI to reduce mobile energy consumption.

We identify groups of papers that consider similar topics or use similar techniques. We pinpoint main research directions, such as offloading and networking optimization to save energy on mobile devices and the analysis of the energy consumption of federated learning. However, other areas, such as approximation computing, have been less investigated.

For researchers, this paper provides an overview of this research area and it points to promising directions for future research. It is also relevant for stakeholders in the mobile computing industry, as we identify potential solutions that arise from the deployment of artificial intelligence models in mobile apps. It also helps in identifying areas where further research and investment are needed.

References

[1] A. Ignatov, K. Byeoung-Su, R. Timofte, and A. Pouget, "Fast camera image denoising on mobile gpus with deep learning, mobile ai 2021 challenge: Report," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 2515–2524.

[2] M. Schuster, "Speech recognition for mobile devices at google," *Pacific Rim International Conference on Artificial Intelligence*. Springer, 2010, pp. 8–10.

[3] H. Zhu, M. Akrout, B. Zheng, A. Pelegris, A. Jayarajan, A. Phanishayee, B. Schroeder, and G. Pekhimenko, "Benchmarking and analyzing deep neural network training," in *2018 IEEE International Symposium on Workload Characterization (IISWC)*. IEEE, 2018, pp. 88–100.

[4] J. Liu, J. Liu, W. Du, and D. Li, "Performance analysis and characterization of training deep learning models on mobile device," in *2019 IEEE 25th International Conference on Parallel and Distributed Systems (ICPADS)*. IEEE, 2019, pp. 506–515.

[5] L. Cruz and R. Abreu, "Performance-based guidelines for energy efficient mobile applications," in *2017 IEEE/ACM 4th International Conference on Mobile Software Engineering and Systems (MOBILESoft)* 2017, pp. 46–57.

[6] ——, "Catalog of energy patterns for mobile applications," *Empirical Software Engineering*, vol. 24, pp. 2209–2235, 2019.

[7] B. Kitchenham, "Procedures for performing systematic reviews," *Keele, UK, Keele University*, vol. 33, no. TR/SE-0401, p. 28, 2004.

[8] C. Wohlin, "Guidelines for snowballing in systematic literature studies and a replication in software engineering," in *Proceedings of the 18th international conference on evaluation and assessment in software engineering*, 2014, pp. 1–10.

[9] K. Petersen, S. Vakkalanka, and L. Kuzniarz, "Guidelines for conducting systematic mapping studies in software engineering: An update," *Information and software technology*, vol. 64, pp. 1–18, 2015.

[10] P. Mayring *et al.*, "Qualitative content analysis," *A companion to qualitative research*, vol. 1, no. 2, pp. 159–176, 2004.

[11] D. Maurer and T. Warfel, "Card sorting: a definitive guide. boxes and arrows, 2004," 2008.

[12] R. Somula and R. Sasikala, "A research review on energy consumption of different frameworks in mobile cloud computing," *innovations in computer science and engineering*, pp. 129–142, 2019.

[13] H. Van Hasselt, A. Guez, and D. Silver, "Deep reinforcement learning with double q-learning," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 30, no. 1, 2016.

[14] V. Mnih, K. Kavukcuoglu, D. Silver, A. A. Rusu, J. Veness, M. G. Bellemare, A. Graves, M. Riedmiller, A. K. Fidjeland, G. Ostrovski *et al.*, "Human-level control through deep reinforcement learning," *nature*, vol. 518, no. 7540, pp. 529–533, 2015.

[15] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial networks," *Communications of the ACM*, vol. 63, no. 11, pp. 139–144, 2020.

[16] A. Lewkowycz, A. Andreassen, D. Dohan, E. Dyer, H. Michalewski, V. Ramasesh, A. Slone, C. Anil, I. Schlag, T. Gutman-Solo *et al.*, "Solving quantitative reasoning problems with language models," *arXiv preprint arXiv:2206.14858*, 2022.

[17] D. Banerjee, "A microarchitectural study on apple's a11 bionic processor," *Arkansas State University: Jonesboro, AR, USA*, 2018.

[18] M. Gupta, "Google tensor is a milestone for machine learning," Oct 2021. [Online]. Available: https://blog.google/products/pixel/introducing-google-tensor/

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

[20] M. U. Khan, S. Abbas, S. U.-J. Lee, and A. Abbas, "Measuring power consumption in mobile devices for energy sustainable app development: A comparative study and challenges," *Sustainable Computing: Informatics and Systems*, vol. 31, p. 100589, 2021.

[21] R. W. Ahmad, A. Gani, S. H. A. Hamid, F. Xia, and M. Shiraz, "A review on mobile application energy profiling: Taxonomy, state-of-the-art, and open research issues," *Journal of Network and Computer Applications* vol. 58, pp. 42–59, 2015.

[22] S. Pasricha, R. Ayoub, M. Kishinevsky, S. K. Mandal, and U. Y. Ogras, "A survey on energy management for mobile and iot devices," *IEEE Design & Test*, vol. 37, no. 5, pp. 7–24, 2020.

[23] A. Deshpande, "Exploring energy consumption issues for video streaming in mobile devices: a review," *International Journal of Advanced Engineering Research and Science*, vol. 4, no. 1, p. 236986, 2017.
[24] Y. G. Kim, J. Kong, and S. W. Chung, "A survey on recent os-level energy management techniques for mobile processing units," *IEEE Transactions on Parallel and Distributed Systems*, vol. 29, no. 10, pp. 2388–2401, 2018.
[25] N. Parajuli, A. Alsadoon, P. Prasad, R. S. Ali, and O. H. Alsadoon, "A recent review and a taxonomy for multimedia application in mobile cloud computing based energy efficient transmission," *Multimedia Tools and Applications*, vol. 79, no. 41, pp. 31 567–31 594, 2020.
[26] Y. Shi, K. Yang, Z. Yang, and Y. Zhou, *Mobile Edge Artificial Intelligence: Opportunities and Challenges*. Academic Press, 2021.
[27] D. Shi, L. Li, R. Chen, P. Prakash, M. Pan, and Y. Fang, "Towards energy efficient federated learning over 5g+ mobile devices," *IEEE Wireless Communications*, 2022.

PRIMARY STUDIES

[28] M. L. Memon, M. K. Maheshwari, N. Saxena, A. Roy, and D. R. Shin, "Artificial intelligence-based discontinuous reception for energy saving in 5g networks," *Electronics*, vol. 8, no. 7, p. 778, 2019.
[29] B. K. Donohoo, C. Ohlsen, S. Pasricha, Y. Xiang, and C. Anderson, "Context-aware energy enhancements for smart mobile devices," *IEEE Transactions on Mobile Computing*, vol. 13, no. 8, pp. 1720–1732, 2013.
[30] Y. Zhang, D. Niyato, and P. Wang, "Offloading in mobile cloudlet systems with intermittent connectivity," *IEEE Transactions on Mobile Computing*, vol. 14, no. 12, pp. 2516–2529, 2015.
[31] P. Nawrocki, B. Sniezynski, and H. Slojewski, "Adaptable mobile cloud computing environment with code transfer based on machine learning," *Pervasive and Mobile Computing*, vol. 57, pp. 49–63, 2019.
[32] P. Akki and V. Vijayarajan, "Energy efficient resource scheduling using optimization based neural network in mobile cloud computing," *Wireless Personal Communications*, vol. 114, no. 2, pp. 1785–1804, 2020.
[33] X. Chen, H. Zhang, C. Wu, S. Mao, Y. Ji, and M. Bennis, "Performance optimization in mobile-edge computing via deep reinforcement learning," in *2018 IEEE 88th Vehicular Technology Conference (VTC-Fall)*, pp. 1–6, IEEE, 2018.
[34] O. Machidon, J. Asprov, T. Fajfar, and V. Pejović, "Context-aware adaptation of mobile video decoding resolution," *Multimedia Tools and Applications*, pp. 1–32, 2022.
[35] P. Nawrocki, B. Sniezynski, J. Kolodziej, and P. Szynkiewicz, "Adaptive context-aware energy optimization for services on mobile devices with use of machine learning considering security aspects," in *2020 20th IEEE/ACM International Symposium on Cluster, Cloud and Internet Computing (CCGRID)*, pp. 708–717, IEEE, May 2020.
[36] Y. Deng, S. Gu, C. Jiao, X. Bao, and F. Lyu, "Making resource adaptive to federated learning with cots mobile devices," *Peer-to-Peer Networking and Applications*, vol. 15, no. 2, pp. 1214–1231, 2022.
[37] A. Shahidinejad, F. Farahbakhsh, M. Ghobaei-Arani, M. H. Malik, and T. Anwar, "Context-aware multi-user offloading in mobile edge computing: a federated learning-based approach," *Journal of Grid Computing*, vol. 19, no. 2, pp. 1–23, 2021.
[38] J. Zheng, K. Li, N. Mhaisen, W. Ni, E. Tovar, and M. Guizani, "Exploring deep reinforcement learning-assisted federated learning for online resource allocation in privacy-preserving edgeiot," *IEEE Internet of Things Journal*, 2022.
[39] D. E. Ruíz-Guirola, C. A. Rodríguez-López, S. Montejo-Sánchez, R. D. Souza, and M. A. Imran, "DRX-based energy-efficient supervised machine learning algorithm for mobile communication networks," *IET Commun.*, vol. 15, pp. 1000–1013, Feb. 2021.
[40] T. T. Anh, N. C. Luong, D. Niyato, D. I. Kim, and L.-C. Wang, "Efficient training management for mobile crowd-machine learning: A deep reinforcement learning approach," *IEEE Wireless Communications Letters*, vol. 8, no. 5, pp. 1345–1348, 2019.
[41] N. Q. Hieu, T. A. Tran, C. L. Nguyen, D. Niyato, D. I. Kim, and E. Elmroth, "Deep reinforcement learning for resource management in blockchain-enabled federated learning network," *IEEE Networking Letters*, vol. 4, no. 3, pp. 137–141, 2022.
[42] H. Huang, Y. Yang, Z. Jiang, and Z. Zheng, "Worker-centric model allocation for federated learning in mobile edge computing," *IEEE Transactions on Green Communications and Networking*, 2022.
[43] K. Aggarwal, C. Zhang, J. C. Campbell, A. Hindle, and E. Stroulia, "The power of system call traces: Predicting the software energy consumption impact of changes," in *Proceedings of 24th Annual International Conference on Computer Science and Software Engineering*, CASCON '14, (USA), p. 219–233, IBM Corp., 2014.
[44] K. Aggarwal, A. Hindle, and E. Stroulia, "Greenadvisor: A tool for analyzing the impact of software evolution on energy consumption," in *2015 IEEE international conference on software maintenance and evolution (ICSME)*, pp. 311–320, IEEE, 2015.
[45] M. Pandey, B. D. Cruz, M. Le, Y.-W. Kwon, and E. Tilevich, "Here, there, anywhere: Profiling-driven services to tame the heterogeneity of edge applications," in *2021 IEEE International Conference on Smart Data Services (SMDS)*, pp. 61–71, IEEE, 2021.
[46] H. Wang, B. Kim, J. Xie, and Z. Han, "Energy drain of the object detection processing pipeline for mobile devices: Analysis and implications," *IEEE Transactions on Green Communications and Networking*, vol. 5, no. 1, pp. 41–60, 2020.
[47] L. Li, Y. Wei, L. Zhang, and X. Wang, "Efficient virtual resource allocation in mobile edge networks based on machine learning," *Journal of Cybersecurity*, vol. 2, no. 3, p. 141, 2020.
[48] M. Szabó, "Machine learning on android with oracle tribuo, smile and weka," in *Proceedings of the 1st Conference on Information Technology and Data Science*, 2021.
[49] S. Bhattacharya and N. D. Lane, "Sparsification and separation of deep learning layers for constrained resource inference on wearables," in *Proceedings of the 14th ACM Conference on Embedded Network Sensor Systems CD-ROM*, pp. 176–189, 2016.
[50] P. Nawrocki *et al.*, "Learning agent for a service-oriented context-aware recommender system in heterogeneous environment," *Computing and Informatics*, vol. 35, no. 5, pp. 1005–1026, 2016.
[51] A. McIntosh, S. Hassan, and A. Hindle, "What can android mobile app developers do about the energy consumption of machine learning?," *Empirical Software Engineering*, vol. 24, no. 2, pp. 562–601, 2019.
[52] J. Kang, Z. Xiong, D. Niyato, Y. Zou, Y. Zhang, and M. Guizani, "Reliable federated learning for mobile networks," *IEEE Wireless Communications*, vol. 27, no. 2, pp. 72–80, 2020.
[53] S. Chu, J. Li, J. Wang, Z. Wang, M. Ding, Y. Zhang, Y. Qian, and W. Chen, "Federated learning over wireless channels: Dynamic resource allocation and task scheduling," *IEEE Transactions on Cognitive Communications and Networking*, 2022.
[54] H. Eom, R. Figueiredo, H. Cai, Y. Zhang, and G. Huang, "Malmos: Machine learning-based mobile offloading scheduler with online training," in *2015 3rd IEEE International Conference on Mobile Cloud Computing, Services, and Engineering*, pp. 51–60, IEEE, 2015.
[55] J. Zhou, G. Feng, T.-S. P. Yum, M. Yan, and S. Qin, "Online learning-based discontinuous reception (drx) for machine-type communications," *IEEE Internet of Things Journal*, vol. 6, no. 3, pp. 5550–5561, 2019.
[56] A. Azari, F. Salehi, P. Papapetrou, and C. Cavdar, "Energy and resource efficiency by user traffic prediction and classification in cellular networks," *IEEE Transactions on Green Communications and Networking*, vol. 6, no. 2, pp. 1082–1095, 2021.
[57] P. Brand, B. Hackenberg, J. Falk, and J. Teich, "Grant prediction-based dynamic power management for 5g to reduce mobile device energy consumption," in *2022 International Wireless Communications and Mobile Computing (IWCMC)*, pp. 647–652, IEEE, 2022.
[58] J. Ren, L. Gao, H. Wang, and Z. Wang, "Optimise web browsing on heterogeneous mobile platforms: a machine learning based approach," in *IEEE INFOCOM 2017-IEEE Conference on Computer Communications*, pp. 1–9, IEEE, 2017.
[59] P. Nawrocki, B. Sniezynski, J. Kolodziej, and P. Szynkiewicz, "Adaptive context-aware energy optimization for services on mobile devices with use of machine learning considering security aspects," in *2020 20th IEEE/ACM International Symposium on Cluster, Cloud and Internet Computing (CCGRID)*, pp. 708–717, IEEE, 2020.
[60] Z. Tang, S. Guo, P. Li, T. Miyazaki, H. Jin, and X. Liao, "Energy-efficient transmission scheduling in mobile phones using machine learning and participatory sensing," *IEEE Transactions on Vehicular Technology*, vol. 64, no. 7, pp. 3167–3176, 2014.
[61] M. L. Memon, M. K. Maheshwari, D. R. Shin, A. Roy, and N. Saxena, "Deep-DRX: A framework for deep learning–based discontinuous reception in 5G wireless networks," *Trans. Emerging Telecommun. Technol.*, vol. 30, p. e3579, Mar. 2019.