Internet of Things and Cyber-Physical Systems 3 (2023) 71–92

Contents lists available at ScienceDirect

Internet of Things and Cyber-Physical Systems

journal homepage: www.keaipublishing.com/en/journals/internet-of-things-and-cyber-physical-systems

Edge AI: A survey

Raghubir Singh [a], Sukhpal Singh Gill [b,*]

[a] Department of Computer Science, University of Bath, Bath, United Kingdom
[b] School of Electronic Engineering and Computer Science, Queen Mary University of London, London, United Kingdom

ARTICLE INFO

Keywords:
Edge AI
Edge computing
Artificial intelligence
Fog computing
Machine learning
Cloud computing

ABSTRACT

Artificial Intelligence (AI) at the edge is the utilization of AI in real-world devices. Edge AI refers to the practice of doing AI computations near the users at the network's edge, instead of centralised location like a cloud service provider's data centre. With the latest innovations in AI efficiency, the proliferation of Internet of Things (IoT) devices, and the rise of edge computing, the potential of edge AI has now been unlocked. This study provides a thorough analysis of AI approaches and capabilities as they pertain to edge computing, or Edge AI. Further, a detailed survey of edge computing and its paradigms including transition to Edge AI is presented to explore the background of each variant proposed for implementing Edge Computing. Furthermore, we discussed the Edge AI approach to deploying AI algorithms and models on edge devices, which are typically resource-constrained devices located at the edge of the network. We also presented the technology used in various modern IoT applications, including autonomous vehicles, smart homes, industrial automation, healthcare, and surveillance. Moreover, the discussion of leveraging machine learning algorithms optimized for resource-constrained environments is presented. Finally, important open challenges and potential research directions in the field of edge computing and edge AI have been identified and investigated. We hope that this article will serve as a common goal for a future blueprint that will unite important stakeholders and facilitates to accelerate development in the field of Edge AI.

1. Introduction

As IT developed after 2000, Cloud Computing was established as a novel computing infrastructure for the Internet based on highly resourced data centres. Interest in and adoption of cloud computing services has increased to the extent that global cloud IP traffic will account for more than 90% of total data centre traffic by 2020 [1]. The main advantages of the cloud computing paradigm remain "unlimited" storage capacity and computing resources, reduced capital expenditure and minimized carbon footprints [2].

However, this technology faces key issues: security, speed of services and slow connections, which are often combined as low bandwidth/high latency and jitter as mobile devices offload computational and processing capacity to cloud computing services [3,4]. These challenges have been exacerbated by the continued proliferation of mobile and fixed Internet-connected devices [5]. Problems of high latency and narrow bandwidths with reduced Quality of Experience (QoE) for users led to proposals to re-imagine the cloud: rather than being thought of as a homogeneous entity; the cloud would have a distinct "edge" separate from the core in which large-scale processing and storage would occur [6]. Devices could, therefore, communicate with local servers unless a need arose for contact with the cloud's core competencies [7]. This view was first articulated as the challenge to the rapidly increasing reliance on mega-data centres for hosting cloud computing [8]. These authors argued that geo-diverse multiple data centres would provide a superior model for applications such as email distribution, using "local" servers to filter out spam and blocking undesirable forms of traffic closer to their points of origin [9]. This formed, in effect, the first proposal for "edge" computing based on the deployment of "micro data centres" (mDCs) as advanced by Microsoft, Inc., which can be seen as a highly distributed cloud focused on mobile users and connected devices and requiring the installation of a global infrastructure of hardware sites, each with a limited number of servers (up to 10 per centre) and supplied with several terabytes of memory [8,10].

Shortly afterwards, the paradigm of the "cloudlet" for small numbers of casual and transient mobile device users in locations such as coffee shops and restaurants etc., was articulated [11]. A third concept developed from the increasing availability and use of fixed Internet-connected sensors (the "Internet of Things (IoT)") requiring fast responses; this was structured as the "Fog Computing" (FC) paradigm [12].

* Corresponding author. School of Electronic Engineering and Computer Science, Queen Mary University of London, London, E1 4NS, United Kingdom.
E-mail addresses: rs3022@bath.ac.uk (R. Singh), s.s.gill@qmul.ac.uk (S.S. Gill).

https://doi.org/10.1016/j.iotcps.2023.02.004
Received 8 February 2023; Received in revised form 21 February 2023; Accepted 27 February 2023
Available online 3 March 2023

Since 2011, mobile vendors have brought powerful smart mobile devices to change the fundamentals of how people interact with IT and telecommunications [13]. Due to significantly increased demands of mobile devices such as smartphones and tablets and because intensive mobile applications require high levels of processing and rely on remote data centres, accessing mobile services at "anytime, anywhere" increasingly clashes with users' QoE and their sense of personal privacy and control [14].

By its very nature, Edge Computing must be accessible by (and respond to) a heterogeneous collection of devices in wireless networks: Wi-Fi, 3G, 4G, 5G and beyond [9]. To ensure the key essentials of low latency and high bandwidth in this highly flexible and highly changeable system, wireless interference must be minimized [15]. Fig. 1 presents the overall history of the edge computing paradigms from World Wide Web (WWW) to Edge Artificial Intelligence or Edge AI.

Mobile-access Edge Computing (MEC) initially emerged as an edge computing paradigm where a mobile user does not need to access cloud computing for data or computing capabilities in remote data centres but can use "edge" computing resources. The fundamentals were discussed in a white paper published by the European Telecommunications Standards Institute (ETSI) in 2014 [15]. The concept of MEC is simply to provide mobile and cloud computing services within close proximity of the mobile user, i.e. the provision of computing power in a delocalized manner close to mobile users (smartphones, tablets, etc.), aiming to decrease latency, achieve as high throughput as possible and provide direct access to real-time network information [16]. The renaming of MEC as Multi-access Edge Computing reflects aims for applications development in 2017 for non-mobile devices; this is a crucial change of direction and its full implications will be discussed later in this survey.[1]

Fig. 2 illustrates the four Edge Computing paradigms in three-tier hierarchies and shows where actual functionality can be implemented either at the end device or at the edge network [17]. FC end devices such as CCTVs can do some processing and send useful data to fog nodes to the fog's core. A simple Cloudlet server at the business premise can perform processing itself rather than at the end devices or in combination with end devices. An mDC processes multiple users' requests locally [18]. Lastly, the concept of Mobile Cloud Computing (MCC) repeats many features of Cloud Computing but, because of the constraints of mobile devices (processing power and battery life), data processing is forwarded to cloud data centres and is therefore not Edge Computing.

Fig. 1. Evolution of edge computing.

1.1. The birth of edge AI

With the latest innovations in Artificial Intelligence (AI) efficiency, the proliferation of IoT devices, and the rise of edge computing, the potential of edge AI has now been unlocked [19]. This has resulted in hitherto inconceivable applications for edge AI, such as assisting radiologists in diagnosing diseases, driving automobiles on the highway, and even fertilizing plants [20]. Numerous experts and companies are discussing and adopting edge computing, which has its roots in the 1990s with the advent of content delivery networks that employ edge servers located near customers to provide web and gameplay videos [21]. Nearly every single industry today has tasks that might be better served by using edge AI [22]. In reality, edge applications are leading the charge for the future generation of AI computing, which will have positive effects on our everyday lives in many contexts, including at home, in the workplace, in the classroom, and when travelling [23].

AI at the edge is the use of AI in real-world devices. Edge AI refers to the practice of doing AI computations near the users at the network's edge instead of centralised location like a cloud service provider's data centre or a company's own private data warehouse [24]. The Internet's worldwide reach means that any region might be considered its periphery [25]. It might be everything from a storefront to a factory to a hospital to the ubiquitous traffic signals, driverless equipment, and mobile phones. Increased automation is a goal of businesses across all sectors since it leads to greater efficiency, productivity, and security [26]. Computer programmes may assist with this by learning to spot patterns and reliably carry out the same actions over and over [27]. But since the world is chaotic and human activities span an unlimited number of scenarios, it is hard to completely represent them in a set of rules and algorithms [28]. Edge AI has advanced to the point where robots and gadgets may now function with the "intelligence" of human cognition regardless of where they be. Intelligent IoT applications powered by AI can adapt to new situations and learn to execute the same or identical tasks successfully [29].

Recent developments in key dimensions have made it possible to successfully deploy AI models at the edge [30]. Ultimately, the foundations for generalised machine learning have been laid by advances in neural networks and other areas of AI [31]. Successful training of AI models and deployment of these models into operation at the edge is something that many organizations are realizing [32]. AI at the edge needs massively dispersed computing capacity [33]. To run neural networks, recent developments in massively parallel GPUs have been applied. The proliferation of IoT-connected devices is largely responsible for the current era's unprecedented growth in data volume [34]. Now that we have the data and the devices required to implement AI models at the edge, we can start doing so in practically every facet of the industry, thanks to the proliferation of sensors, smart cameras, robotics, and other data-gathering tools [35]. IoT is also benefiting from the improved speed, reliability, and security that 5G/6G is bringing to the battlefield [36].

[1] Multi-access Edge Computing: http://www.etsi.org/news-events/news/1180-2017-03-news-etsi-.

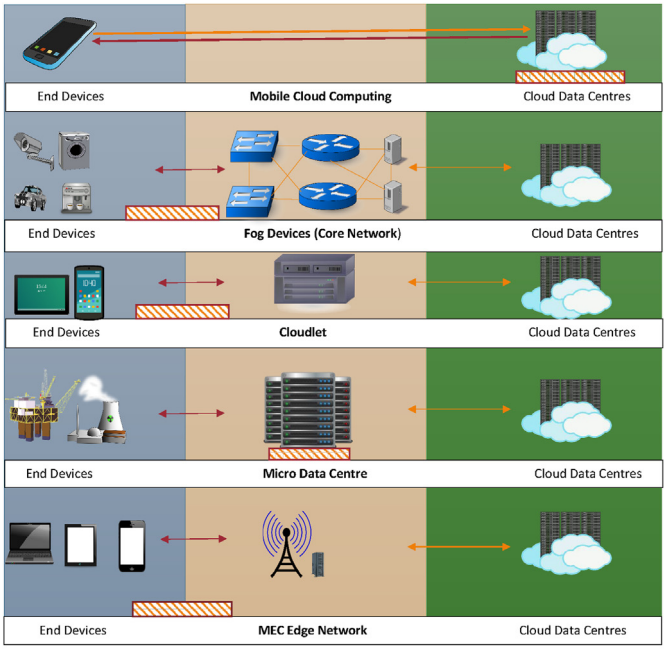

Fig. 2. Three-tier relationship between users/devices and cloud computing and four paradigms of intervening Edge Computing as well as Mobile Cloud Computing.

1.2. Motivation

Each of these four edge computing concepts, Cloudlet, FC, MEC and mDCs, shares a perceived vision of the future of the Internet that urgently requires addressing the mismatch between cloud computing (with its finite number of distant data centres) and the increasing number of mobile users competing for access with a relentless increase in edge devices via edge computing [37]. Overlap between the four concepts as they have been discussed in the literature has, however, resulted in significant degradation of the boundaries between the different approaches. Further, because of the network performance improvements offered by Edge Computing, it is considered as an enabling technology for improving performance by leveraging Edge AI in the "last mile" of wireless networks [38].

1.3. Comparison with related surveys

Existing survey articles [39–41] introduced detailed surveys on edge computing by focusing on only IoT applications but [41] has also discussed about resource management in edge computing. Further, another survey paper [42] has given a detailed study about resource management in edge computing. Another work [43] only focuses on edge computing and gives various future directions. Further, we have identified three relevant surveys [19,21,23] which are mainly focuses on Edge AI. Most of the survey articles on Edge AI have been published before 2021 except [23], but this [23] work is just an overview on Edge AI or a visionary work which offers only future directions in many relevant areas but not in-depth. In order to analyse, update, and integrate the existing research and explore prospective trends and futuristic views in the field of edge computing, especially Edge AI, there is a requirement for a new innovative study as this area of edge computing is continuing to expand towards Edge AI. In addition to the findings of earlier studies [19,21,23, 39–43], the insights of this latest research present a novel and inventive strategy for evaluating and pinpointing the most important research gaps in the literature. Table 1 compares our survey with existing ones based on different criteria.

1.4. Our contributions

In this context, this survey's contributions are the following.

1) Analyzes Edge Computing paradigms, especially Edge AI and discusses cloudlets, FC, MEC and mDCs at various stages of development.
2) Presents Edge Computing and Edge AI-based application use cases being investigated in proof of concept studies.
3) Discusses models and initial commercial service offerings.
4) Demonstrates how the FC and MEC paradigms have been unified in a multi-access approach to heterogeneous networks that employ WiFi technologies and which will evolve to utilize 5G and Long-Term Evolution (LTE).
5) Compares Edge AI with cloudlets, FC, MEC and mDCs based on basic computing characteristics along with iewpoints of applications, functionalities and technologies.
6) Discusses the Edge AI approach to deploying AI algorithms and models on edge devices, which are typically resource-constrained devices located at the edge of the network.
7) Highlights future research directions and offers open challenges for future readers.

1.5. Article structure

The rest of this paper is structured as illustrated in Fig. 3. Section II explores the background to individual variants proposed for implementing Edge Computing, including Edge AI. Section III analyses the benefits and experimental demonstrations of computation offloading to

Table 1
Comparison of our work with existing survey articles.

Works	Edge	Edge AI	IoT Applications	Standardization	Performance Comparisons	Resource Management	Year
[39]	x		x				2017
[40]	x		x				2019
[41]	x		x			x	2019
[42]	x					x	2021
[43]	x						2021
[19]	x	x				x	2020
[21]	x	x					2020
[23]	x	x[a]	x				2022
Our Survey (this work)	x	x	x	x	x	x	2023

Abbreviations: × : = method supports the property.
[a] = just an Overview/Visionary.

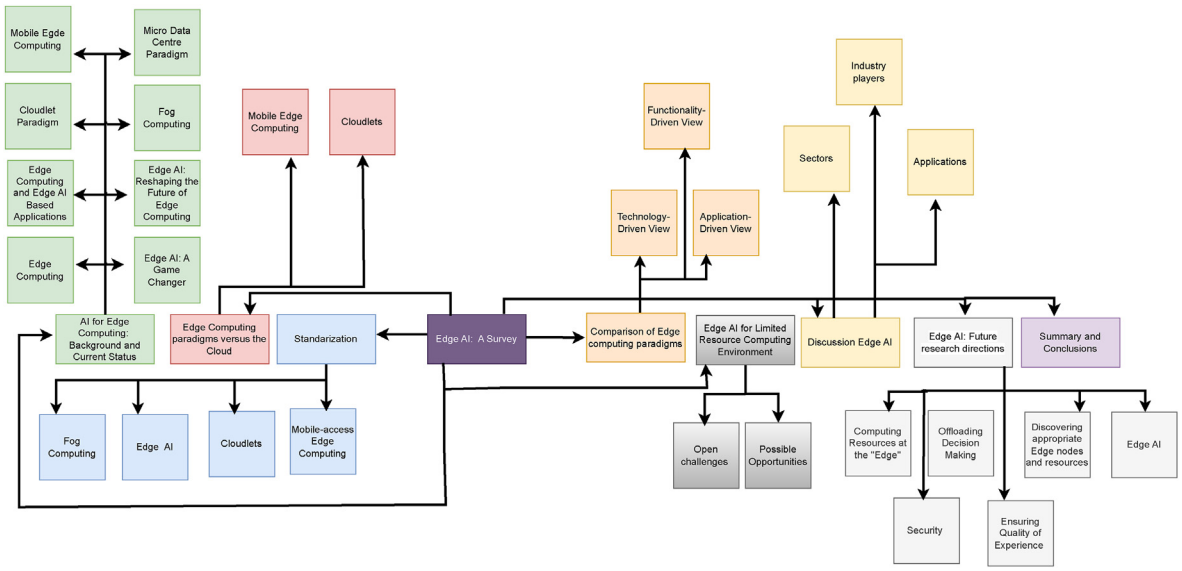

Fig. 3. The Organization of this Survey.

edge servers, work mostly done for cloudlets and mobile edge computing. Section IV discusses standardization initiatives for Edge Computing paradigms and Edge AI. Section V offers a novel analysis that compares the different Edge Computing paradigms and Edge AI from the viewpoints of applications, functionalities and technologies. Section VI gives a detailed overview of Edge AI for finite computing resources at the edge network and possible opportunities. Section VII explores the potential advantages of Edge AI at the edge. Section VIII highlights the trends of the transition from Edge Computing to Edge AI and discusses present and future Edge AI research challenges. Finally, Section IX summarizes and concludes the article and draws the future of Edge AI towards moderen IoT applications. The list acronyms used in this survey are given in Appendix A.

2. Background

Edge computing proposes moving data processing capability away from distant consolidated data centres in the cloud to servers usually (but not exclusively) physically closer to the end user in order to support high QoE applications for heterogeneous mobile device users and fixed Internet-connected streaming devices via wireless networks [43]. Between 2009 and 2014, four paradigms for Edge Computing Cloudlets, Fog Computing, Mobile Edge Computing and Micro Data Centres emerged as practical contenders for implementation of the Edge Computing concept [44]. Unfortunately, the expanding literature dealing with these distinct concepts often demonstrates overlap and mutual occupation of the diverse conceptual spaces [45]. While previous surveys have focused on individual Edge Computing paradigms, a detailed comparative analysis of the various approaches for implementing Edge Computing paradigms in terms of definitions, architectures, paradigm evolution and application use cases, viewed from the standpoints of applications, functionalities and technologies for 5G and the IoT [46]. Distinct sectors of Edge Computing can be clearly defined that address different markets with different business models and technologies [47]: Fog Computing as the optimal solution for IoT applications to minimize the time required for time-critical processing and Big Data analytics; Mobile Edge Computing as centred on applications intended for large numbers of mobile devices, Micro Data Centres as the optimum for industrial applications in specific locations or at temporary sites while cloudlets are dedicated either (as originally envisioned) to small-enterprise scenarios or can be expanded to publicly funded schemes with free access and open source software [48]. In 2017, Mobile Edge Computing was retitled Multi Access Edge Computing to recognize the large IoT component in proof-of-concept studies and to harmonize Edge Computing systems and applications in highly heterogeneous wireless networks [49]. With the latest innovations in AI efficiency, the proliferation of IoT devices, and the rise of edge computing, the potential of edge AI has now been unlocked [50,51].

2.1. Edge computing

The key objective of Edge Computing is to put resources within close proximity to the users and sources of data and information to help overcome cloud computing's recognized weaknesses of high latency, jitter and narrow bandwidth [11,52]. These performance parameters are particularly important and relevant for wireless access networks in the context of a user's computing devices as well as in the IoT [53].

- *Agility of Services:* Nomadic mobile devices and fixed IoT sensors generate enormous amounts of data sent to the central Cloud. However, due to the centralized approach of the Cloud, this has lacked various important features such as contextual and location awareness [47]. If Edge Computing processes data at the edge network, then context and location awareness are much more readily obtainable and achievable [54].
- *Low latency:* Reducing the time required when a packet travels from a node to the destination is critical in high processing applications such as augmented reality and gaming where mobile users expect uninterrupted services from the content provider [55].
- *Coherence:* The Edge Computing architecture can determine where to offload data, either on the local device or the edge network [49]. Smart sensors make decisions and this improves the performance of the overall network and sends only useful data to the cloud [56]. For example, a CCTV camera captures and transmits information only when movement occurs in close proximity to the camera [57].
- *No Single Point of Failure:* Edge Computing stores the limited amount of resources that allows applications to control computing, offloading and networking resources to achieve the high level of efficiency and performance [58]. Additionally, the architecture of Edge Computing provides a distributed approach if the primary edge network resources pool fails, “instantaneously” redirecting traffic to alternative edge network resources [32]. With emerging technologies, such as software-defined networking (SDN) and Network Function

Virtualization (NFV), this enables reliability and robustness of the network and improves integration with existing IoT environments [59].

2.2. Edge AI: A game changer

Practically every sector has felt the effects of AI's role in digital transformation [60]. AI and Machine Learning (ML) are being used by industries to streamline processes, enhance the customer experience, and save expenses [61]. In addition to AI development, the advent of edge computing has been crucial in enabling businesses to have data evaluated and managed at the Edge immediately [62]. The term "edge" refers to a type of distributed computing that places processing, data storage, and energy generation at the scene of an event [13].

Smart cities and smart factories would not have been possible without edge computing, which has emerged as an essential part of the Industrial Internet [50]. Smart cities rely on edge computing, which allows sensors on traffic lights and other public utilities to communicate with smartphones and apps. It also helps sensor-equipped machines in "smart factories" report on its status and performance in real-time [63]. The potential to deploy AI on the Edge has sprung up and flourished in recent years, thanks to the growing complexity of activities that must be done by edge devices [59].

There was a time when edge devices couldn't be used for training and deploying machine learning and deep learning models [64]. However, with the introduction of more powerful computing, such edge devices' ability to deal with a variety of AI-optimized tasks has grown substantially in recent years [65]. It has also made it possible to handle data in real time, with a heightened level of privacy and security [66].

Tiny devices are possible in today's generation and there are several examples of smart technology, such as thermostats, doorbells, house and car cameras, and Augmented reality (AR) and Virtual Reality (VR) glasses [67]. Industrial robots, automobiles, smart buildings, and oil rigs are all examples of huge devices. Self-driving cars, autonomous robots, AI-powered household appliances (from vacuum cleaners to drones) and mobile devices (from smartphones to security cameras) are all examples of Edge AI devices that can run an optimized AI on their system [68]. So, thanks to Edge and AI, machines and gadgets can instantly understand, learn, and act on data and information [69].

Data gathering, preparation, and analysis may happen in near real-time without having to wait for the model output to be downloaded from the cloud [70]. When AI models are put on the Edge instead of in the cloud, we allow Edge AI devices to analyse data more quickly, make decisions more quickly, increase the security of data processing, and improve the user experience [71].

Many sectors might anticipate major changes due to AI-powered edge computing [50]. International Data Corporation (IDC) predicts that by 2025, there will be 150 billion intelligent edge devices in use [72]. It's true that certain forms of edge computing are in use already, but that number is expected to explode in the near future. By 2025, Gartner expects 75% of enterprise-generated data will be created and processed outside of the conventional data centre or cloud, thanks to the expansion of the IoT [73]. As of 2019, the worldwide edge computing industry is estimated to be worth $3.5 billion. By 2027, that figure might have increased to $43.4 billion [74]. This suggests that AI computation on such devices will grow at an exponential rate, allowing consumers to engage with compiled code at the source in a safe and optimal manner [75].

The aforementioned figures demonstrate that the massive digital change brought about by edge computing powered by AI is only getting started. Given the complexity of hybrid cloud infrastructure and accompanying applications, there are several obstacles that must be overcome before such a transition can be implemented [76]. The rewards are undeniable, but getting started on the path today is crucial.

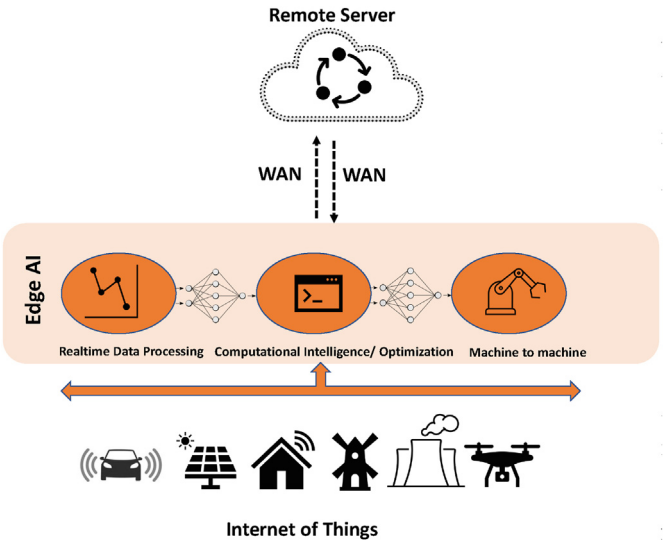

Fig. 4. Edge AI: Reshaping the future of edge computing.

2.3. Edge AI: Reshaping the Future of Edge Computing

Edge AI refers to the use of AI algorithms and techniques on devices located at or near the edge of a network, such as on a mobile device, sensor, or IoT device, rather than relying on a central cloud server for processing [77]. This allows for faster and more efficient data processing, as well as improved privacy and security [78]. Fig. 4 shows that machine learning algorithms and techniques can be deployed on local devices or at the edge of a network to minimize the processing time [79]. For example, real-time data processing approach at the edge for applications where the processing speed is critical, such as video and audio streaming, online gaming, and financial trading [80].

Furthermore, Edge AI could be provided with the computational intelligence to develop intelligent systems that can perform tasks that typically require human-level intelligence, such as decision-making, problem-solving, pattern recognition, and learning [79]. Computational intelligence encompasses various subfields, including machine learning, neural networks, fuzzy systems, evolutionary computation, and swarm intelligence [81]. These techniques are often used in applications such as robotics, data mining, control systems, and optimization problems. The goal of computational intelligence is to create intelligent systems that can adapt to changing environments and learn from experience without being explicitly programmed [82]. Machine-to-machine (M2M) approach could be considered when direct communication is required between devices or machines, without the need for human intervention [83]. This approach can be achieved through a variety of wireless and wired communication technologies, such as Bluetooth, Wi-Fi, and cellular networks. Examples of M2M applications include smart home automation, industrial automation, and remote monitoring and control systems [84].

2.4. Edge computing and edge AI based applications

In this section, we discuss the most popular Edge Computing and Edge AI-based applications.

2.4.1. Computation offloading at the edge

With the emergence of IoT technology and high mobile-intensive applications such as 4 K surveillance cameras, virtual reality and gaming applications etc., requirements for high processing power have increased. However, due to hardware limitations of wireless devices' computation, offloading must be performed to increase functionality, but overloaded offloads can degrade performance [85]. The solution is an

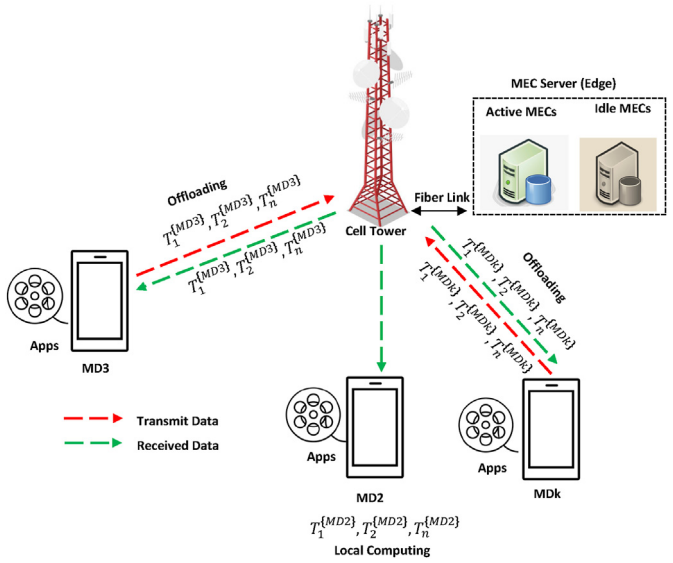

Fig. 5. The model of computation offloading at edge network.

edge server: a wireless device can transfer a series of tasks to the edge server for offloading. Therefore, this substantially reduces the latency and improves the bandwidth between the wireless users and the edge server. However, computational offloading still faces many challenges, in particular, how a wireless device will decide which tasks need to perform locally or which to hand over to the edge server. Another critical issue is on the edge server side: what happens if the edge server starts to overload due to the high demand from the wireless devices? This is considered further in Section VII. The basic architecture of computation offloading with a MEC network is shown in Fig. 5 shows that multiple users communicate with a base station and thence to a Edge server.

2.4.2. Context aware applications

Previously, when the Internet user frequently web surfed to see favourite content, to fulfil the user's request, the Internet provider used historical information in their database, but increasingly Internet service providers can provide users' favourite content via geographical locations or analysed information from the application [86]. With Edge Computing, content providers can host services at the edge of the network with accurate user location within a radio access network [16]. This can improve the QoE for mobile users [5].

2.4.3. Smart transport

Many cities are trying to implement various forms of this; for instance, when poor traffic decisions are made by traffic controllers, adverse weather conditions and delays caused by road re-construction all add to traffic congestion and inefficient fuel usage [87]. With Edge Computing, edge servers can automatically manage city traffic via data collected from intelligent sensors at traffic lights and CCTV cameras on highways [88]. Each sensor detects car movements and makes decisions accordingly, and traffic lights can react to this processed information. Another critical scenario is intelligent car parking, where users will access information about urban car parking spaces according to their geographical location [89]. Currently, an intelligent transportation environment uses cloud computing, where all the processing is done at remote data centres, but cloud computing lacks key safety features in smart transportation [56]. For example, if a driverless car needs to stop in case of a dangerous situation, it has to upload the data to the cloud, which then performs a computing process and sends the "stop" command to the car when the car finally acts upon the instruction [90]. A more rapid solution is to bring computation capability close and Edge Computing can provide limited computation capability to make quick (lower latency) decisions.

2.5. Cloudlet paradigm

The concept of the cloudlet was derived from two basic premises: firstly, mobile devices (excluding laptops and notebooks) were "resource poor", i.e. compared to static PCs and laptops/notebooks, mobile devices have little computing power; secondly, providing a "data center in a box" offered a small number of mobile device users in a private business (for example, a coffee store) the ability to leverage computing power [11]. "Computing power" was – in this original context – a suite of open-access software options incorporating Linux applications for word processing, spreadsheet data processing, etc.: AbiWord, GIMP, Gnumeric, Kpresenter, PathFind and SnapFind. These resources would be combined in a maintenance-free Virtual Machine (VM) environment with post-use clean-up for users accessing the cloudlet transiently via short-range wi-fi connections. The cloudlet paradigm has evolved in an open-source platform, OpenStack++, with cloudlet discovery and just-in-time provisioning. Elijah is the implementation of cloudlet-based Mobile Computing and the Elijah-related source code uses github.[2]

Cloudlets as a business concept have failed to gain traction and the default offering from small enterprises has become free Wi-Fi for mobile devices (including laptops and notebooks) [91–93]. The driver for this has undoubtedly been the near-pervasive use of social media by mobile users, the vast majority of whom have shown little taste or need for a "data centre in a box". This in turn has evolved the cloudlet concept into providing one link in a 3-tier hierarchy: mobile device/cloudlet/cloud [94].

Such an arrangement was made explicit in an application demonstrated for cloudlet computing in cognitive assistance [95]. This proof of concept study utilized Google Glass, streaming video from the Google Glass device to the cloudlet. An open-source platform for cognitive assistance uses virtual machine encapsulation connected to a publish-subscribe mechanism to efficiently share sensor data streamed from the wearable device [47]. The cloudlet can subsequently link with the "traditional" cloud for services including centralized error reporting, usage logging and pre-collection of data [96]. Logically, this hierarchical architecture has gradually linked cloudlets to other forms of Edge Computing, in particular as local solutions to high latency and low bandwidth problems for IoT applications [97]. This application case considered video images continuously recorded by, for example, dashboard cameras in vehicles but also considered personally uploaded video content from cameras on mobile devices; cloudlet systems were used to process information before metadata was sent to the cloud. The same approach would be applicable to fixed surveillance cameras operated by public bodies on city streets. This case study overlaps FC, as discussed in (Sec. II-F) and (with images uploaded from mobile devices) MEC (Sec II-G). The application for mobile device users was, however, an evolution of the cloudlet paradigm and the authors presented a semi-quantitative analysis of how to structure the cloudlet architecture for an urban environment, concluding that smaller edge cloudlets performed better for multiple users than a larger centralized metropolitan area network [9]. The implication of these findings was that an unspecified number of smaller cloudlet servers would be positioned to provide diffuse coverage to ensure a high QoE [5].

The cloudlet architectural concept has also been applied to reconfigure the "traditional" cloud to add an intervening layer between mobile users and the cloud [98]. This approach evolved the "data centre in a box" to a more ambitious device with only a "soft" state, i.e. cached from the cloud but able to buffer data from a mobile user. Additionally, the cloudlet device would be sufficiently powerful for resource-intensive applications from multiple mobile devices within one hop and would use cloud infrastructure and standards [91–93]. This architecture would move data centre processing "closer" to the mobile user but this does not imply a closer physical proximity but a "logical proximity", i.e. an

[2] QEMU: https://github.com/cmusatyalab/elijah-qemu

arrangement to minimize latency and jitter while maximizing bandwidth.

In practical terms, this closer proximity might be implementable by augmenting Wi-Fi access points by adding processing, memory and storage; a desirable side-effect of such an arrangement would be to extend the battery life of the expected mobile device by requiring less power usage for transactions with the cloudlet [98]. In effect, the authors were suggesting cloudlets as an intervening layer in mobile cloud computing – as opposed, for example, to the deployment of mDCs data centres (Sec. II-H). In other words, the original cloudlet in a fixed location has become one of many components in an intervening layer between the mobile user and the cloud but offering better performance for the mobile user, whether or not the user is connected to a geographically closer server (i.e. the "logical proximity" effect). The remote cloud is still available but as a last-resort or for delay-tolerant resource-intensive applications accessed by mobile users.

Finally, and inverting the cloudlet paradigm, the concept of a cloudlet formed from resource-rich mobile devices has been proposed [99]. The precise definition of resource-rich mobile devices was not made clear in this mathematical study but it can be assumed that high end tablets were included although no logical reason was proposed to exclude laptop/-notebook computers from the mobile cloudlet.

Articles discussed in (Section III-A) present case studies that have explored the practical implementation of cloudlets, i.e. application use cases. These have included linking mobile devices to large public screens [100], applications such as face and speech recognition, object identification, physical simulation and rendering and augmented reality [98] and cognitive assistance [95].

Such experimental demonstrations have shown the power of a local cloudlet in accelerating the completion of computing tasks by resource-poor mobile devices [91]. In other words, it is not the lack of effective functioning of cloudlets that have made the commercial take up of the cloudlet paradigm minimal, rather it is the lack of profitability in the concept - conceptual business model [92]. Eventually, mDCs may fill this gap for specific tasks in industrial practice [93].

This concept was at least partially considered in an earlier study where dynamic cloudlets were proposed from any mobile devices in the network that possessed the necessary computing resources [101]. This analysis proposed an infrastructure co-located with the Wi-Fi access point but also capable of discovery locally devices that could share computing resources [5]. The same fundamental idea has been elaborated into the FemtoCloud proposal, in which mobile devices with significant idle computing power can link via a client service installed on the devices [102]. A typical scenario might again be a coffee shop in which a broadly predictable customer base can be assumed (the same might also hold for forms of public transportation or even airport terminals). The authors concluded from a detailed mathematical analysis that the client service faced a major problem in scheduling computational events and tasks in a transient architecture of mobile processing capacity but the design architecture enables a useable configuration and users can be given incentives to participate if the concept were to be implemented [102].

2.6. Fog computing

The origins can be traced to a conference presentation by Cisco Systems, Inc. in 2012 [12]. The authors considered the problems inherent in devices accessing cloud computing resources and, like the cloudlet architecture, proposed the insertion of an extra layer between the end user device and the cloud: embedded systems and sensors linked first to a "field area network" that comprised the FC ("distributed intelligence") element which itself communicated with the Cloud [103]. This was articulated more persuasively in a subsequent Cisco document: "Analysing data close to the device that collected the data can make the difference between averting disaster and a cascading system failure" [6]. Furthermore, the same document states "Any device with computing, storage, and network connectivity can be a fog node. Examples include industrial controllers, switches, routers, embedded servers, and video surveillance cameras" [104]. This explicit linkage to fixed devices with connectivity and intelligent data analysing and data processing capabilities clearly distinguished FC from cloudlets [7]. The scenarios of interest identified were: networks of wireless sensors, connected vehicles, smart grid distribution networks and in any context where data is collected at the "edge": vehicles, ships, factory floors, roadways, railways, etc. [12,104]. The great value of this approach was noted as the dense geographical distribution with a local focus can analyse big data faster [105].

Smart traffic lights in vehicular networks, self-driving vehicles, smart meters monitoring domestic energy use and, pipeline monitoring, wind farms, closed loop control of industrial systems, and applications in the oil and gas sector were soon added to the list of things and fog functional relationships [106–108]. The relationship between FC and the cloud was succinctly described as: "Thus, the solution to this problem is a multi-tiered architecture (with at least 3 tiers) whereby an IoT application is deployed as follows: a part on the "thing" (e.g. a car), a 2nd part on the fog platform (e.g. a roadside cabinet or a router in a wireless access network or an LTE base station), and in the case of three tiers a final 3rd part in a datacentre of the main cloud (e.g. Amazon EC2)" [109]. Just as explicitly linking connected things, FC and the cloud is the definition of FC given by the Open Fog[3]: *"Fog computing is a system-level horizontal architecture that distributes resources and services of computing, storage, control and networking anywhere along the continuum from Cloud to Things*".

However, even by 2013, degradation of the link between the IoT and FC began to be evident. One of the features of FC was claimed to be the "great support for mobility", although "mobility" was not further defined [105]. [110] included 5G mobile devices along with the IoT, cyber-physical systems and data analytics in the applications open to FC. Similarly, mobile users were considered to use applications that could benefit from FC [111].

The title of an article in 2015 - "Fog Computing: Focusing on Mobile Users at the Edge" – unambiguously links mobile devices to FC [52]. Mobile users are equally a part of the view that FC enables the "everything-as-a-service" model [112]. Laptops and wearable devices have also been viewed as mobile devices connecting to "the fog" [113].

Nevertheless, focusing FC concepts on real-time data processing and analytics as opposed to the human user-generated demand for computing and processing power was emphasized in a proposal for "Edge-centric Computing" [14]. These authors also asserted that trust in the security of personal and socially sensitive data would be increased if the management of such sensitive data could be ensured at the edge rather than being centred in distant data centres [114].

Cisco introduced applications use cases of FC that include Fog Computing and the Internet of Everything (IoE), Video analytics optimisation.[4] Another work [115] discussed five likely areas for Fog Computing deployment: healthcare, smart grids, smart vehicles, urgent computing and augmented reality [116].

Several authors [106,117–121] studied smart grids based on meters in domestic and industrial settings to provide real-time data on power consumption are significant initiatives with both financial and environmental motives. Vehicular FC is a novel proposal to combine computational devices, both onboard vehicles and the power and resources of Edge Computing [122].

While augmented reality can be viewed as being more in the domain of MEC than of FC, urgent computing in disaster and emergencies is an FC application that can significantly speed up response times and optimize responses [115].

At present, there are many different offered definitions for the term FC [123,124]. The dilemma lies in extending the definition of "thing" to personal mobile devices, particularly smartphones and tablets. Once the

[3] OpenFog: https://opcfoundation.org/markets-collaboration/openfog/

[4] IoT, from Cloud to Fog Computing: https://blogs.cisco.com/perspectives/iot-from-cloud-to-fog-computing

Table 2
ETSI Sponsored proofs of concept (PoC) [130].

PoC	Application	Objective
PoC 1	Video User Experience Optimization via MEC	Recognizing paid video subscribers and assigning a higher priority to those streams
PoC 2	Edge Video Orchestration and Video Clip Replay	Enable the mobile user to receive live video streams from professional stadium cameras, choosing camera angles, etc., to enrich the fan experience
PoC 3	Radio aware video optimization in a fully virtualized network	Exploring video optimization to adjust the quality of video streams according to radio conditions of the users to give users an improved QoE.
PoC 4	Flexible IP-based Services	Accelerating the delivery of IP-based content and streaming media
PoC 5	Healthcare-Dynamic Hospital User & Alert Status management	Intelligent healthcare applications that can be used at hospitals by using wireless telecom services
Poc 6	Multi-Service MEC Platform for Advanced Service Delivery	multiple MEC platforms and applications residing on shared and common computing infrastructure to synergize service function chaining to enhance QoE and operator visibility
Poc 7	Video Analytics	Aiming to provide video surveillance to cities and other physical situations over an LTE network using MEC to analyse raw video streams from surveillance cameras.

concept of a network of fixed or moving sensors streaming data to an FC server is lost, the paradigm is radically altered, which is discussed next [9].

2.7. Mobile edge computing

The definition provided by the first paper on Mobile Edge Computing (MEC) was "Mobile-edge Computing provides IT and cloud-computing capabilities within the Radio Access Network (RAN) in close proximity to mobile subscribers" - this clearly differentiates MEC from Cloudlets (the business model based on free access) and FC (the focus on mobile devices in the IoT) while again aiming to bring cloud computing capabilities closer to the user base similar to all other Edge Computing paradigms [15]. The white paper also proposed setting up a new Industry Specification Group (ISG) inside the European Telecommunications Standards Institute (ETSI) to begin defining and devising industry specifications for multi-access MEC [13]. This highlights the role of telecommunications in MEC, which can be viewed as uniting the telecommunications industry with IT at the mobile network edge [120,125]. A variant term - "mobile edge cloud computing" - has also been used, but this is conceptually indistinguishable from MEC: "mobile-edge cloud computing can provide cloud-computing capabilities at the edge of pervasive radio access networks in close proximity to mobile users" [126].

As its core concept, MEC reconfigures the devices already deployed at the mobile edge as mobile access points, i.e. base stations forward traffic but also adds computing and storage capabilities to act as MEC servers [127]. Four distinct stakeholders contributed to this early vision of MEC: mobile users connecting to base stations, MEC servers and other hardware owned and maintained by network operators [128]. At the same time, Internet providers added connectivity to cloud elements (data centres and content distribution networks) in which application service providers host applications. This architecture aims to reduce latency, improve bandwidth and enhance scalability for mobile users while catalysing the development of entirely new services.

The original ETSI white paper [10] envisaged six use cases: active device location tracking, augmented reality content delivery, video analytics, "Radio Access Network" (RAN) aware content optimization, distributed content and domain name system caching and application-aware performance optimization. Of these, video analytics was presented as an IoT application use case in which video streams from cameras were utilized for public safety and smart city data collection [129]. The other five were, however, genuinely aimed at mobile device users. Table 2 shows the ETSI-sponsored proofs of concept (PoCs) that showed the viability of the MEC concept.

PoC-like studies outside ETSI at the Mobile-access Edge Computing Congress held in Munich (Germany) in September 2016 were presented, including Low Latency MEC Supercomputers for Connected Vehicles, Video Analytics and Ecommerce Transaction Management (IDT, Integrated Device Technology), Using MEC to Improve Road Safety and Traffic Control (Deutsche Telekom) and Mobile-access Edge Computing to Enhance Customer Experience in High Traffic Locations (EE). Some, if not all, of these, have strong IoT linkages.

The 5G MiEdge (Millimetre-wave Edge Cloud as an enabler for 5G ecosystem) project, funded by the European Union, is primarily focused on a millimetre wave 5G Radio access network.[5] This project has two main goals. Firstly, it considers mmWave access and MEC are combined to reduce the computation task at the edge of the network. The second goal is to develop a novel control plane to maximize resources for mobile users. The project will contribute to the standardization of mmWave access and Radio Access Network Centralised-plan in 3GPP and IEEE. Ultimately, it will demonstrate a joint 5G test-bed in the city of Berlin and the 2020 Tokyo Olympic Stadium. This project involves two private-sector participants: Intel Deutschland and Telecom Italia.

The Small Cells Coordination for Multi-tenancy and Edge Services[6] (SESAME) project is inter-sectoral and is focused on delivering novel architectures and standardized technologies for next-generation mobile communication. The strategy is to bring intelligence through Network Functions Virtualization (NFV) and Mobile-access Edge Computing in cellular network architectures.

Applications for Mobile Edge Computing were launched by Nokia.[7] The main aim of launching these applications is to enable businesses to get the benefit of applications which use low latency, lower costs and an improvement in bandwidth resource utilization [60]. One example of a MEC application from Nokia is video surveillance which enables security personnel the ability to analyse unusual activities from anywhere at any time [131].

Despite its short life history, MEC began to evolve rapidly. In addition to or place of base stations, MEC may utilize more cost-effective points in IP networks and may adopt Network Functions Virtualization (NFV) technologies in distributed MEC platforms [132]. However, another trend is that of incorporating IoT in MEC schemes and implementation scenarios [133]. This was at least partly envisaged by the original document for MEC, including machine-to-machine scenarios connecting sensors to MEC servers [15]; an IoT Gateway was subsequently discussed in an ETSI white paper [134]. This was unfortunate because confusion with Fog Computing was made possible and MEC and Fog Computing have subsequently been completely "mixed and matched" by various authors [52,112,135,136]. Nevertheless, different drivers have been emphasized when these two edge computing paradigms are compared, and a critical difference between FC and MEC is that wireless IoT networks can be viewed as the principal driver for FC, whilst low latency and resource efficiency in cellular networks are dominant considerations for MEC [137].

2.8. Micro data center paradigm

In contrast to other edge computing models, mDCs form more of a hardware solution for Edge Computing than a novel IT scenario [16]. As such, mDCs are compatible with FC, MEC and cloudlets but are highly

[5] 5G MiEdge: https://cordis.europa.eu/project/id/723171
[6] SESAME: https://cordis.europa.eu/project/id/671596
[7] Nokia Edge cloud: https://www.nokia.com/networks/portfolio/edge-cloud/

Table 3
Selected performance metrics for edge computing/cloud comparisons.

Cloud [100]	*Mean Latency (ms)*			
Asia	337			
US (East-West)	161–228			
Cloudlet [100]	*Mean Latency (ms)*			
EU	59–93			
US	186			
	Energy (J)	*Time-to-Completion (sec)*		*Throughput (KB/s)*
Cloud [141]	2052	2223		102
Cloudlet [141]	103	226		1002
	2013 [98]		*2016* [143]	
	Mean RTT (ms)	**Energy** (J)	**Mean RTT** (ms)	**Energy** (J)
Cloudlet	80	1.1	80	2.6
Cloud (US E-W)	260–420	3.1–5.1	320–420	4.4–6.1
Cloud (EU)	420	5.2	500	9.2
Cloud (Asia)	800	9.4	770	9.2

suitable for a completely different market sector, IIoT [138]. Ruggedized mDCs, able to be sited out of doors, are already available for installation in remote sites, for example, for oil/gas exploration, and any industrial application requiring sensor data, machine-to-machine communication and control and automation technologies will be amenable to this form of edge computing. Bahl [10] discusses mobile devices with wireless connections to mDCs achieving improved battery life between recharging events and high-end game stream but many applications are clearly IoT-related [10]. Juniper presents industrial scenarios in which locatable mDCs can be operated in extreme environments or on a temporary basis [132].

2.9. Multiple paradigms for multiple sectors

The above analysis emphasizes the diversity of applications being investigated for Edge Computing but demonstrates the different conceptual landscapes being considered [17]. At one extreme, a smart city that has thousands of Internet-connected fixed devices streaming information and sometimes requires local processing to filter out unnecessary data to so focus on situations that require "instant" responses. Alternatively, a private enterprise may employ mDCs in a geographically isolated environment to rapidly process sensor streams that have only temporary significance before they are moved. Mobile users may have high computational demands to maintain QoE, but this leads to novel services and applications, some of which are being actively investigated in ground-breaking studies of functionalities under "filled" conditions [139]. proposed the offloads framework, which automatically offloads the computation among the peers within the local network. This approach distributes the work overload between mobile users and reduces energy consumption, and increases battery longevity [140].

3. Performance of edge computing paradigms versus the cloud

This section discuss the performance of edge computing paradigms versus the cloud.

3.1. Cloudlets

Other than speculating on business models, the original presentation of the cloudlet concept left unanswered important questions concerning how much computing power would be required and the acceptable access time for any individual user [11]. More detailed analysis of the operating parameters of a cloudlet required specific application use cases and a seminal one was provided by a study of cloudlet functioning to facilitate the appropriation of a digital screen by a mobile device user [100]. The scenario envisioned was dramatic but realistic: a doctor (physician) was interrupted over dinner in a restaurant and asked to scrutinize a pathology slide in real-time as a surgical operation continued; a smartphone was used to access a large lobby screen to fully visualize the medical state of affairs. To simulate this, a simple display screen game was played using cloudlets and a commercial cloud to assess the impact of physical distance on user experience (Table 3).

To some extent at least, physical proximity must reduce latency; considerably longer latency delays have, under experimental conditions, been measured than the "ideal" latencies imposed by geographical distance (Table 3).

These positive results for cloudlets were extended in a study of cloudlet versus cloud performance using both Wi-Fi and radio access (LTE) means of communication from mobile devices [134]. With a variety of applications, the use of a local cloudlet greatly improved response times (in some instances by a factor of approximately five-fold); the greater the geographical distance to the cloud, the more the response time was elongated (Table 3). In addition, energy usage by the cloudlet was always reduced when compared with any cloud option, in some cases by factors of 7–10. In a separate practical test of cloudlet versus cloud, reductions in power usage by 50%, a nearly 10-fold reduction in delay time and an increased throughput of 10-fold (Table 3) were achieved using a single-device cloudlet with one mobile user [141].

Data are from Refs. [94,98,141] were captured using free and research-grade apps using study groups in the UK, Eastern US and Central Europe sending and receiving messages to and from laboratory cloudlets in the UK, Central Europe and the US or to and from four commercial clouds.

Nevertheless, the cloudlet model has been shown to suffer performance degradation if too many intermediary steps ("hops"), where bandwidth was reduced) were required between the mobile user and the cloudlet target in a simulation study [142]. With one or two cloudlet wireless "hops" used to transfer data, the cloudlet outperformed the cloud-based approach for application scenarios that included file editing and video streaming; with more intermediate steps, the cloud option performed better because of smaller request transfer delays [18].

3.2. Mobile edge computing

On a much larger scale – considering an area of 931 km^2 with approximately one base station per 2 km^2 and servicing 180,000 users generating daily traffic in excess of 10 TB, a "cloudlet network" has been simulated and mathematically analysed in Ref. [144]. The results suggest, however, that the problem addressed in that study was more related to MEC and/or mDCs than to cloudlets because so many devices were included in the network.

While these authors did not explicitly compare MEC and Cloud performance parameters, the results demonstrated advantages in an optimized "cloudlet network" in terms of traffic volume and reduction of latency [144].

Extrapolating from results with various definitions of cloudlets, it can be concluded from the published work that interposing edge servers between users and the Cloud is a viable means of improving QoE and overall system performance in terms of latency, bandwidth and consistency of service. Edge Computing is particularly applicable in the case of wireless networks where bandwidth is limited and latency could be high [145].

Nokia's MEC technology is claimed to improve up-link performance by the deployment of multiple base stations within a sports stadium. This scenario (in a sports stadium) does not convincingly address mobile device users because they were in a precisely defined location and were not "mobile" in the sense of freely moving in, for example, a city centre or a large airport. The use case study was, however, entirely within the definition of Multi-Access Edge Computing, which aims to converge telecommunication and IT services in multiple settings [146].

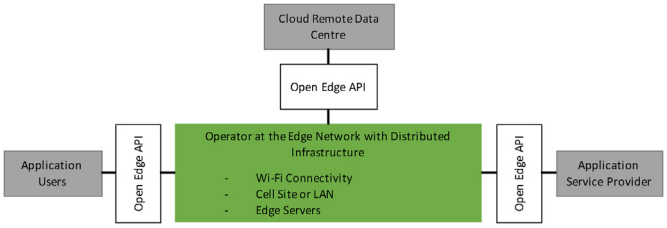

Fig. 6. Standardisation initiatives for Edge Computing development [137].

This approach optimized bandwidth and reduced uplink congestion as well as decreasing power requirements for uploading the content, thus making smartphones more energy efficient and with extended battery lifetimes.[8]

4. Standardization for edge computing paradigms and edge AI

Edge Computing projects and proposals are distributed across academic, industry and standardization activities and standardization initiatives have been active since 2009 [34]. The challenge for edge computing is to provide standardized means of access for users and novel service providers as well as an efficient interface to the cloud as shown in Fig. 6.

4.1. Cloudlets

The basis for Edge Computing using OpenStack with open sources to interpose a local facility between the cloud and the mobile edge is discussed extensively in the Open Edge Computing project.[9] Carnegie Mellon University, in the United States, is currently working with several companies to standardize the framework for cloudlets. In 2015, various organizations collaborated with each other and formed open source Open Edge Computing research. This research mainly focuses on cloudlets, and the application developers are trying to develop new applications such as offloading time-sensitive computations and to handoff the big data from one cloudlet to a second cloudlet through the Dynamic Configuration Application programming interface. In 2016, Open Edge Computing discussed gaps in the current raft of technologies and concluded that the aim should be the development of an open platform for Edge Computing aligned with the IT and telecommunications industries [55]. This would enable edge-enabled businesses to emerge, from drone support services to mobile app enhancements and virtual/mixed reality [147].

4.2. Mobile-access edge computing

ETSI's Mobile Edge Computing Industry Specification Group (MEC ISG) was formed in 2014 with the stated objective "to create a standardized, open environment which will allow the efficient and seamless integration of applications from vendors, service providers, and third-parties across multi-vendor Mobile-edge Computing platforms" [10].

The 3rd Generation Partnership Project (3GPP) united telecommunications standard development organizations to devise new standards for cellular telecommunications network technologies, including radio access, the core transport network, and service capabilities.[10] Technical reports focus on developments required to drive 5G New Radio architecture and interfaces: 3GPP TR 38.913 covers current views on scenarios and requirements and 3GPP TR 38.801 traces out architectural and interface aspects for 5G New Radio.

Table 4
Proposed characteristics for variants of edge computing.

Key Parameters	Cloudlets	Fog Computing	Mobile Edge Computing	Micro-Data Centre	Edge AI
Rapid Response	No	Yes	Yes	Yes	Yes
Latency	Low	Low	Low	Low	Ultra-low
Mobility	Yes	Yes	Yes	Yes	Yes
User	Local and Smart City	Security industry and network providers	Telecom and Software Providers	Hardware	Local and Smart City
Security Provider	None	Service Provider	Service Provider	Service Provider	Service Provider
SLA	None	Essential	Essential	Essential	Essential
Academic research input	High	High	High	High	Low

4.3. Fog computing

In 2015, The OpenFog Consortium was established by Cisco Systems, Intel, Microsoft, Princeton University, Dell and ARM Holdings. The reason behind the establishment was to create a common architecture of FC and advancement in various fields such as the transportation, health and education sectors.

The OpenFog Consortium has to invest in reference architectures, developer guides, samples, and SDKs to articulate the value of FC to developers and IT teams. They are critical for the success of this new initiative. FC, along with the cloud, will accelerate the adoption of IoT in the. The Industrial Internet Consortium.[11] emerged from various industry stakeholders: multinational corporations, small and large research development innovators, academic institutes and government organizations. Twenty "testbeds" have been created where technologies, applications, products, services, and processes can be investigated; these testbeds include ones for smart water, energy and precision crop management where the architectures and functions of Edge Computing are important components [13].

4.4. Micro-data centre

Standardization for micro-data centres is important to ensure consistency, reliability, and interoperability between different components and systems. It can also help reduce costs, simplify deployment, and improve overall efficiency [30]. Some of the key areas that should be standardized for micro-data centres.

- Power and Cooling: Standardizing the power and cooling infrastructure can help ensure compatibility and interoperability between different components and reduce the risk of equipment failure [31]. This includes standardizing power distribution, uninterruptible power supply (UPS), and cooling systems.
- Networking: Standardizing the networking infrastructure can help ensure that different components can communicate with each other seamlessly. This includes standardizing protocols, cabling, and network topology [26].
- Management and Monitoring: Standardizing the management and monitoring tools can help simplify deployment and reduce the risk of human error. This includes standardizing remote management protocols, monitoring and reporting tools, and alerting mechanisms [24].

[8] Nokia Centralized RAN: https://www.cambridgenetwork.co.uk/node/558949
[9] Open Edge Computing: https://openedgecomputing.org/
[10] 3GPP: https://www.3gpp.org/about-us/introducing-3gpp
[11] Industry IoT Consortium: https://www.iiconsortium.org/

Table 5
Comparison for variants of edge computing, application-driven view.

Key Parameters	Cloudlets	Fog Computing	Mobile Edge Computing	Micro-Data Centre	Edge AI
Application Domain	Mobile	IoT and Mobile	Mobile	IIoT	IoT
Real-time interaction	Yes	Yes	Yes	Yes	Yes
5G	No	Yes	Yes	Linked to/with	Yes
Tactile Internet	Not Possible	Possible	Possible	Possible	Possible
Smart City	No	Yes	Yes	Yes	Yes
Working Environment	"Data centre in a box" at the business premises	Indoor and Outdoor	Indoor and Outdoor	Indoor and Outdoor	Indoor and Outdoor
Service type	Mobile Device	Fixed Device	Mobile Device	Movable Device	IoT

Table 6
Comparison for variants of edge computing, functionality-driven view.

Key Parameters	Cloudlets	Fog Computing	Mobile Edge Computing	Micro-Data Centre	Edge AI
Location Awareness	Limited	Unlimited	Unlimited	Limited	Unlimited
Number of users	Few users at a time 25–50	Many Users at a time 100–2000	Many Users at a time 100–1000	Many Users at a time 100–4000	Many Users at a time 100–4000
Content Consumption	Fixed locale	Anywhere	Anywhere	Fixed but mobile	Anywhere
Mobile Management	Yes	No	Yes	No	Yes

Table 7
Comparison of variants of edge computing, technology-driven view.

Key Parameters	Cloudlets	Fog Computing	Mobile Edge Computing	Micro-Data Centre	Edge AI
Connectivity	WLAN and Wired	WLAN, Wired,Cellular	WLAN, Wired, Cellular	WLAN and Wired	WLAN, Wired, Cellular
Numbers of servers	One	Numerous	Numerous	Numerous	Numerous
Client Hardware requirements	Local mobile device	Distributed/Hierachical	Distributed/Hierachical	Local-mobile	Distributed/Hierachical

4.5. Edge AI

Edge AI is the newest development in the field of AI and Computing [83]. The algorithms that make up AI are finding their way into a growing number of the goods and services that we use. Edge AI refers to AI that is implemented directly into a smart device [148]. The way in which we make use of it and the technical potential presented by the study raises a number of issues to understand the problems, the technologies that are utilized, the use cases, the stakeholders, and the chances for the future of computing [149].

5. Comparison of edge computing paradigms and edge AI

Edge Computing paradigms are clearly related but can be distinguished on the bases of applications, functionalities, technology and implementation (commercialization and business models) to establish a taxonomy. The significant features of this proposed taxonomy are summarized in Tables 4–7.

The majority of the criteria used were evolved from tables, data and text in published sources [10,11,52,105,107,124,150]. The principal differentiators for applications are mobile devices versus fixed devices, means of access, application domains and drivers (Table 5).

Functionalities differ markedly across the four paradigms, especially in content awareness and the number of users (Table 6). Technologies also differ greatly when the four paradigms of Edge Computing are compared, in particular the number of server nodes and the use of standardized open environments (Table 7).

The financial investments required are key items in decision making but, equally, opportunities are major factors in implementation routes for Edge Computing (Table 4). Business models must be considered to accurately delineate commercialization opportunities. If, for example, cloudlets are essentially free at the point of use, this eliminates many applications using fixed sensors [150]. In the following sections, possible applications, functionalities, technologies and implementation strategies will be discussed in depth to illustrate how these differentiators impact the research challenges faced by Edge Computing.

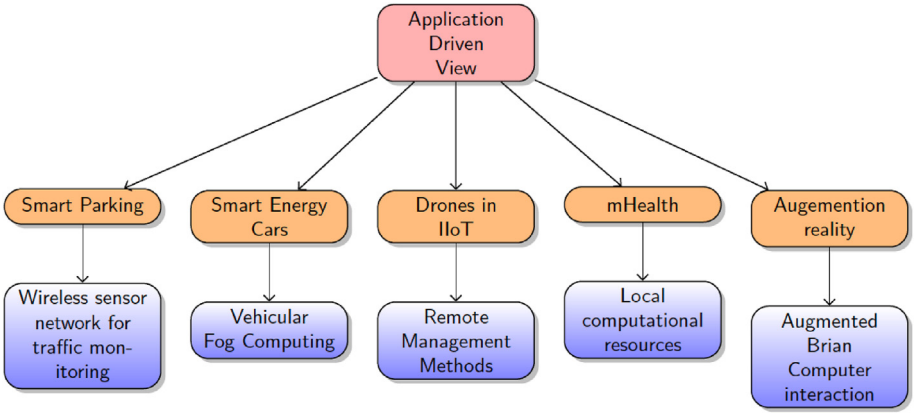

Fig. 7. Taxonomy of edge computing applications.

There are three main drivers for Edge Computing development: the application-driven view, the functionality-driven view and the technology-driven view. These are not mutually exclusive but focus attention on how novel service offerings can be initiated at the conceptual level.

Edge Computing is an essential part of the application discussed below. These applications provide benefits where sensing, pre-processing and segmentation can be done at the end devices or the edge network [33]. To illustrate and analyse this view, six examples of Edge Computing and Edge AI applications are being investigated as shown in Fig. 7.

5.1. Application-driven view

This section describes different modern applications which are utilizing Edge Computing.

5.1.1. Smart parking

The main aim of this smart parking application is to show the availability of different car parking spaces to help a car driver identify vacant spaces through ultrasound sensors deployed in car parking areas to identify occupations [151]. In 2001, the first smart parking experiment was initiated at the Baltimore/Washington International Airport [152]; car parking was terminated only at 99% occupancy. Currently, various companies are working on a solution to provide a service to car drivers requiring parking spaces in city centres by providing an up-to-date ”heat map” of availability [153] However, the application inevitably collects large amounts of data from the various sensors causing a serious problem on the cloud side of data processing. An Approach inspired by Edge Computing principles can be used by an algorithm to do processing on the user device and make decisions on whether or not to send out information to the server in the cloud.

5.1.2. Smart energy cars

This application considers a “Vehicle Edge Computing“ where a gateway to the intra-vehicle network as the edge device and hence placing the edge functionality on a mobile node. Where a gateway to an intra-vehicle network is the edge device and hence places the edge functionality on a mobile node [35]. This solution tries to implement a delay-tolerant and cost-effective solution leveraged on opportunistic wireless access. For example, when a hydrogen-fuelled or electric car is being driven on the motorway then the car can send data to the central server to find out how much hydrogen or battery charge is left before the driver needs to refill or recharge at a hydrogen filling station [154]. Respectively, In this scenario, processing can either be fully completed in the car with an embedded device or data is sent to the central server for processing, the choice is being made on the basis of bandwidth and latency factors in the communications available to the vehicle [44].

5.1.3. Drones in industrial IoT

In 2013, Amazon expressed interest in drones to deliver customer parcels in shorter times [155]. TV and film industries increasingly use inexpensive drones to shoot scenes without any requirement to hire expensive helicopters or small planes [147]. Architecturally this will be similar to Vehicular Edge Computing but using drones [46]. Another example of the non-military use of drones is in the oil and gas industries to replace old management methods such as using helicopters for acquiring a video of difficult-to-access terrain [156]. To achieve these aims, Edge Computing can interface with automatic drones collecting data and transmitting the data in real-time so that onsite managers can visualize how resources (physical and computational) are allocated [69]. In this scenario, drones themselves can make smarter decisions, for example, whether it sends the whole data to the cloud or completes the data processing on board [155].

5.1.4. Daily activities (mobile health)

“Smart” watches provide information over Wi-Fi or cellular networks on daily activities, for example, the number of steps walked, heart rate and quality of sleep. However, this technology imposes a high risk to privacy because single user data is transferred to a central cloud for storage and processing [157]. Edge Computing eliminates the need for such long-reach data transmission if more local computational resources are available [48]. Pre-processing can be performed at the edge node in this case as with the smart parking scenario discussed above.

5.1.5. Augmented reality

The key concept of this technology is to increase the user's perception of and contact with the real world [158]. As evolving mobile AR games become more complex and offer interactivity, key issues such as high latency, and low QoE for mobile users become central points for developers [49]. According to one survey, delays should not exceed 20 ms, otherwise virtual scenes become unworkable, and this demotivates AR users [45].

MEC can shift mobile networks into real-time location-based gaming, bringing cloud computing capabilities inside a radio access network and this allows delivery of computationally demanding AR applications running close to the end user by avoiding high latency, high real-time responses and backhaul network bottlenecks [61].

The authors of [158] proposed six potential AR applications to be explored with Edge Computing paradigms: medical visualization, maintenance and repair, annotation, robot path planning, entertainment, and military aircraft navigation/targeting while [159] proposed the augmented brain-computer interface to detect user's brain states in real-life situations.

5.1.6. Healthcare

Edge Computing applications have the potential for different industries, but social uses, specifically the field of healthcare has attracted attention [116]. [160] Introduced the idea of FAST (a FC, distributed analytics-based Fall Monitoring System for Stroke Mitigation) to monitor strokes and reduce the severity of strokes. The design of real-world pervasive health applications for FC is still an open question [117]. presented and implemented real-time processing algorithms that executed in a fog node such as personal laptops in order to achieve the smallest processing time for health-related data [114]. [118] proposed the eWALL system to offer solutions for chronic obstructive pulmonary disease and mild dementia. This scheme works on a computational distributed approach with FC to process sensitive information from the patients; the authors argued that this mechanism could be reduced the communications overload between the patients and their physicians [131].

5.1.7. Edge AI

Edge AI has a wide range of applications in various industries [54]. Some examples of how Edge AI is being used today.

Smart Home: Edge AI can be used to power smart home devices such as voice assistants, thermostats, and security cameras [161]. These devices use machine learning algorithms to learn user preferences and behaviour and make real-time decisions based on sensor data;
Driverless Cars: Edge AI is being used in autonomous vehicles to process sensor data and make decisions in real-time, allowing the vehicle to react quickly to changes in its environment [162],
Healthcare: Edge AI can be used to analyse medical data from wearable devices and sensors to monitor patients in real-time, detect abnormalities, and provide early warnings for potential health issues [162],
Business Marketing: Edge AI can be used to analyse customer data, including purchase history and behaviour, to provide personalized recommendations and improve customer experiences [163]. Edge AI has the potential to revolutionize industries by enabling intelligent and autonomous devices that can make real-time decisions based on sensor data [164]. Its applications are diverse, and the technology is

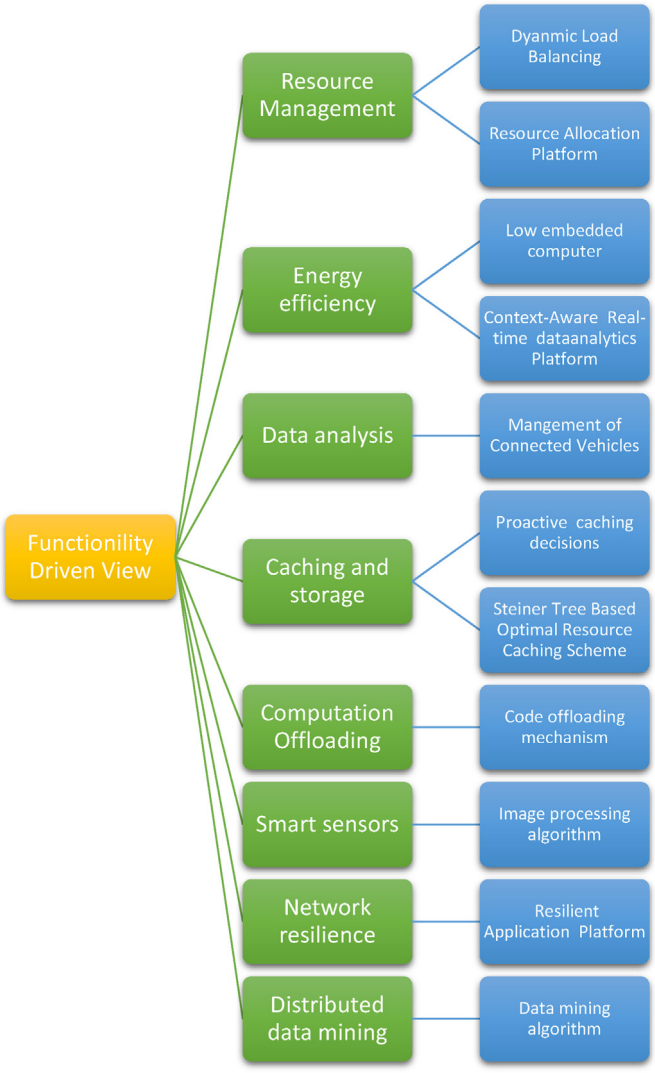

Fig. 8. Taxonomy of edge computing functionality.

still in its early stages, so we can expect to see even more creative and innovative applications of Edge AI in the future [165].

5.2. Functionality-driven view

This section describes different functionalities for application scenarios in Edge Computing. All of these features can be incorporated into the overall aim of performance enhancement but a number can be highlighted and discussed separately as shown in Fig. 8.

5.2.1. Resource management

Resource management means ensuring that sufficient resources are available at the edge network. For example, in smart parking scenarios where data is coming from sensors and transferred to edge devices efficiently and reliably [66]. Central to resource management are dynamic load balancing [166] and developing resource allocation platforms [16]. On-demand resources, variable workloads and integrating data streams from a number of heterogeneous devices across a geographical area are important considerations and but trade-offs between computing power and speed of communication may be inevitable [167].

5.2.2. Energy efficiency

Heterogeneous data flow has been increasing continuously the energy efficiency of end devices declining; therefore, edge networks require intelligent data reduction methods. Authors in work [168] proposed low-embedded computers to perform data mining techniques on raw data collected from wearable sensors for telehealth applications.

The work documented in Ref. [169] proposed the idea of a Context-Aware Real-time data analytics Platform (CARDAP) which can be deployed in complicated mobile analytics applications. Its authors gave the example of sensing activity in smart cities where CARDAP integrates energy data and delivers benefits in energy efficiency.

The research published in Ref. [170] compared the idea of fog computing versus traditional cloud regarding reducing CO_2 emission and evaluating the performance of an environment with a high number of internet-connected devices. A survey of several Indian cities showed that real-time FC applications outperformed cloud computing by reducing environment CO_2 [171].

5.2.3. Data analysis

Data handling is an obvious concern in the Big Data era; FC architecture is well situated to process machine-to-machine data streams with, for example, connected vehicles [172,173]. One real reason is mainly to improve the overall efficiency and performance by reducing the communications overhead (before avoiding the need for sending granular/raw data on limited bandwidth wireless access links) and reducing latency [58].

5.2.4. Caching and storage

General caching and storage processing frameworks for analysing users' data traffic provide platforms to store users' data traffic and extract useful information for proactive caching decisions. FC has been claimed to be well suited for this purpose [174,175]. Such capability improves the user experience and overall application performance by reducing latency experienced by the user.

In order to make proactive caching decisions efficiently, a data processing platform inside the network infrastructure must read and combine inputs from multiple sources and arrive at intelligent insights [59]. After network analysing tools have been utilized, raw data must be exported into big data storage platforms such as the Hadoop Distributed File System.[12]

5.2.5. Computation offloading

With mobile application demands growing at high rates globally, mobile devices are increasingly constrained by limited resources and poor battery longevity [10]. Mobile fog architectures have been discussed with mobile cloud computing and code offloading mechanisms [135,176–180].

5.2.6. Network resilience

A resilient network continues to function and maintain operations despite transient faults or compromises [88]. Edge Computing architectures and systems run an increased risk of failure because computing resources are not contained in a controlled environment with a single set of emergency and backup protocols [181]. The authors propose a resilient application platform to provide core services (for example, distributed data management) and mechanisms to monitor and optimize functionality in relatively long-lived distributed applications [90]. An alternative argument is to improve the resilience by moving the functionality closer to the end devices by relying less on any less reliable links (wireless in particular) [89].

5.2.7. Smart sensors

In the broadest terms, smart sensors are developed over geographical locations connected over a public or private network to either cloud or edge sensors. For example; At the edge network smart sensors can offload the data and a computation offloading algorithm can filter the

[12] Hadoop: https://hadoop.apache.org/

meaningful data for further processing and sending to the remote cloud data centres [182].

5.2.8. Distributed data mining

Distributed data mining is a very useful approach in terms of edge computing paradigms where multiple nodes are connected over fast-speed networks. From time to time, conveying large amounts of data to a data centre is costly and unrealistic and to solve this problem data mining algorithms have been developed [183].

The functionality of Edge AI is to enable AI algorithms to run on edge devices, such as smartphones, IoT devices, and robots, without relying on cloud computing [184]. This means that data is processed and analysed locally, in real-time, at the edge of the network, rather than being transmitted to a remote server for processing [185].

5.3. Technology-driven view

One strategy of Edge Computing may be to focus on virtualization and industrialised applications at the edge network. Such an evolution to Edge Computing would be enabled by currently emerging technologies such as Software Defined Networking (SDN), Network Functions Virtualization (NFV) and Information-Centric Networking (ICN) [186]. These new technologies provide novel tools that increase flexibility in designing networks [103]. Complementary technologies will enable the programmability of control and network functions and eventual migration of these key constituents of the network to the cloud.

5.3.1. Virtualization

Virtualization is part of any new emerging technology in the IT industry, whether running multiple operating systems or several applications on the single server [128]. The key ability of this technology is to reduce the number of physical servers, which can greatly decrease power consumption, air conditioning costs etc. With the growth of IoT, mobile devices and sensors place more pressure on remote data centres [187, 188], so there is a chance to transfer the applications and intelligence from the cloud to the edge network.This will provide a new way of virtualization at the edge a network where a physical server can provide flexible and dedicated storage and cache [129].

5.3.2. Edge computing simulation tools

Simulation software tools provide insight results with a set of mathematical formulas through simulation without having to implement and deploy a real system which is often complex and expensive [189]. There are two main types of simulation packages available: continuous simulation and discrete event simulation [55]. Discrete simulations are utilized to simulate statistical models while the continuous simulation is deployed for a physical processes such as human respiration radio frequency data communication etc. In this section, we identify some important simulation tools which can be used for the Edge Computing paradigms [145].

- CloudSim: This framework for the modelling and simulation of cloud computing infrastructures and services was developed by the Cloud Computing and Distributed Systems (CLOUDS) Laboratory, University of Melbourne [190]. CloudSim is a development toolkit for developers to formulate the cloud and describe necessary input parameters and evaluate outputs before new cloud-based services are made available [191]. Rather than relying on theoretical or empirical approaches, CloudSim can act as a testbed for important commercial factors, for example, resource leasing to cope with varying load and pricing schemes, and supports the study of virtualized server hosts with different policies for providing resources to virtual machines [151].
- Open Cirrus: The simulator is an open cloud and fog computing test bed supported by multiple commercial entities, including Intel, Hewlett-Packard (HP) and Yahoo! Open Cirrus aspires to obtain the below objectives: Foster systems-level research in cloud computing; promote new cloud computing applications and application-level research; provide a set of experimental datasets; formulate APIs and open-source stacks for edge computing [105].
- iFogSim [192]: proposed a simulation tool for modelling and simulating resource management in Edge Computing but particularly for IoT and FC [69]. Essential targets for this modelling and simulation are latency, throughput, network congestion, energy consumption and financial costs. Two case studies were presented and analysed in quantitative detail: a latency-sensitive online game and intelligent surveillance through distributed camera networks. This work overlaps with commercial or near-commercial systems [193–196].

Future readers can access other simulators from our another paper [5].

5.3.3. Software Defined Networking

With various types of cloud applications and services, the performance of data centre networks becomes highly important [60]. The data centre network has a dynamic and often vast amount of traffic due to various cloud services, which makes it more challenging to control congestion in an efficient and user-friendly way. Therefore, SDN is an evolving approach to maintaining and managing network topology and controlling congestion efficiently [9]. SDN offers higher programmability, automation, and network control; highly scalable and flexible networks with fast adaptation to changing business needs [124]. In SDN, the controller acts as a centralized unit that provides all networking functions, such as path selection and policy deployment, which possess several advantages. Using SDN to manage the interaction between cloud and edge resources, a network can remain dynamic, agile and efficient while providing a better experience for the end user [151]. First, the SDN provides real-time knowledge of available resources that is flexible and reliable [55]. A centralized controller allows for optimal decision-making for each unit within the system, and a dedicated control channel allows translation of high-level policies to low-level configuration instructions to give the system fine-grained control [197]. Defined a software framework to increase the efficiency of MCC services by integrating different software-defined system components into MEC architectures [145].

5.3.4. Network function virtualization (NFV)

NFV aims at improved capital efficiencies compared to dedicated hardware-based implementation solutions by deploying commercial off-the-shelf computing hardware to provide virtualized networks, sharing hardware and reducing the number of different hardware architectures [124]. This virtualization proposes better scalability and improved resilience and accelerated service innovation through software-based service deployment [129]. Standard automation and facilitated standardization are other vital features of the proposed NFV.

Open Edge Computing discussed potential benefits to Edge Computing arising from the use of NFV, including the reduction of hardware costs with NFV infrastructures [137].

5.3.5. Information centric networking (ICN)

As complex networks, mobile traffic expand and rapidly increase demand for IoT devices [128]. Thus, high bandwidth and the need for low latency are required at the edge network. 5G Americas discussed MEC and ICN as both these emerging technologies intensively researched in the context of future wireless networks [103]. The aim is to provide a network infrastructure service for contemporary uses of and demands placed on the Internet focusing on user mobility and content distribution [198]. This approach values resilience to disruptions and failures driven by the massive demand for internet applications and services. ICN gives network-layer functions content awareness to enable routing, forwarding, caching and data-transfer operations to be achieved independently of IP addresses [199]. Data “chunks” are defined so that the network

interprets and treats content on a semantic basis and with no necessity for either deep packet inspection or use of the application layer [88].

MEC and ICN are complementary concepts but independent: both are solutions with, for example, 5G mobile networks, which could utilize ICN for content distribution, transparent mobility among multiple technologies, or to retransmit lost packets over an unreliable radio link, but without a commitment to developing a full MEC scenario [90]. Conversely, MEC could be used to reduce latency for AR applications without resorting to ICN. Synergies are, however, possible when MEC and ICN are structured to work cooperatively [89].

5.3.6. Key technologies for edge AI

The technology behind Edge AI is a combination of hardware and software that enables AI algorithms to run on edge devices such as smartphones, drones, robots, and other IoT devices [200]. Here are some of the key technologies that are used in Edge AI.

- Edge devices: These devices are hardware components that are equipped with processing power, sensors, and storage, which enable them to process data and run AI algorithms locally [201].
- Machine learning frameworks: ML frameworks, such as TensorFlow Lite and PyTorch, are used to develop and run machine learning models on edge devices [202].
- Neural network architectures: Neural network architectures, such as convolutional neural networks (CNNs) and recurrent neural networks (RNNs), are used to train and run deep learning models on edge devices [22].
- Model compression: It is a technique that reduces the size of machine learning models by removing redundant information, making it possible to run models with limited resources [20].
- Federated learning: It is a technique that allows machine learning models to be trained on distributed data sources without sharing raw data, improving privacy and reducing network traffic [25].
- Edge-to-cloud integration: It allows for the development of hybrid solutions that leverage the strengths of both edge and cloud computing, enabling real-time processing on edge devices and training models on the cloud [27].

With the right combination of hardware and software, Edge AI technology can enable AI algorithms to run on edge devices, bringing intelligence to the network's edge and enabling new applications and use cases [32]. The technology is rapidly evolving, with new techniques and frameworks being developed to improve the performance, efficiency, and security of Edge AI solutions [62].

6. Edge AI for limited resource computing environment

Edge computing and other resource-constrained locations provide problems and possibilities for AI deployment [50]. Instead of a data centre, edge computing processes data at the network's edge. This provides real-time processing and lower latency, but AI implementation is complicated by constrained CPU cores and network access [63]. We will examine these problems and potential and highlight crucial considerations to consider when implementing AI systems in edge computing and other resource-constrained situations in this section [64].

6.1. Open challenges

The following is a list of the primary issues that must be overcome when implementing AI systems in contexts with limited resources and computational capabilities [65,67,68,70–72].

- **Computing Resources:** Edge computing devices, such as IoT sensors and edge gateways, often have fewer processing resources than central clouds [148]. As a result, it might be difficult to run sophisticated AI models and techniques on such devices because they could demand more resources than are really accessible [149].
- **Internetworking:** It is typically difficult to transport significant volumes of data and models to and from smart edge devices because these systems frequently have restricted or inconsistent network access [161]. Because of this, the quality and dependability of AI algorithms that rely on real-time data inputs might also be negatively impacted [162].
- **Data confidentiality and security:** Edge computing devices frequently gather and handle confidential material, such as records related to an individual's health or banking transactions [145]. While implementing AI systems in these kinds of situations, one of the most difficult challenges is going to be guaranteeing the data's security and confidentiality [114].

6.2. Possible opportunities

Here are several advantages of implementing AI systems in edge computing as well as other low-resource settings [73,75,76,80,81,83].

- **Real-time data analysis:** Edge computing offers real-time data analysis by executing computations at the network's periphery, hence fueling apps like driverless cars, smart buildings, and advanced robotics [162].
- **Minimize delay:** Compared to transmitting data to a centralised data centre, the computation for data handled via edge computing is much shorter. This is useful for time-sensitive software, such as video games and augmented reality [163].
- **Better Privacy/Security:** Edge computing can assist in increasing confidentiality and security by lowering the quantity of information that has to be transferred over the Internet and by offering more command over the processes and storage of confidential documents [164]. This is accomplished by analyzing information immediately at the edge, which reduces the quantity of information that must be transported over the network [165].

7. Discussion

The evolving discussion on Edge Computing has identified four distinct paradigms differentiated on the basis of intended markets and business models [59]. FC may turn out to be the preferred solution for IoT applications to minimize the time required for time-critical processing, for example, facial recognition by surveillance cameras in public spaces, related machine-to-machine operations and Big Data analytics [203]. MEC, in contrast, is centred on extending the applications available to and intended for genuinely mobile devices. Industrial applications in remote geographical locations or at temporary sites are ideal candidates for mDCs [58]. Although cloudlets initiated these developments in edge computing, they have proved marginal in commercial practice but could find implementations in municipality-centred schemes based on free access and open-source software, in contrast to the subscription or fee-based services in other edge computing paradigms [66]. We propose that these markers, based on markets, technologies and business models, be respected to maximize productive dialogue and accelerate technology development in edge computing [54]. Inevitably, some blurring of boundaries will be caused as, for example, vehicles become smarter and more connected, with directional, informational and entertainment streams occurring simultaneously and mobile and wearable devices communicating more with fixed IoT locations [116].

Different network architectures will also be required for these different scenarios [125,133,144,204]. In all likelihood, these architectures will emerge logically from multiple successful applications emerging from demonstration projects and defining their own characteristics and functional requirements [44]. For example, two new architectures - client and vehicular - can be recognized. Vehicular Edge Computing introduces a computation offloading mechanism with

embedded devices in cars offloading data to MEC servers [46]. Implementing offloading and computation resource mechanisms will improve the utilization of MEC servers [48]. Client Edge Computing, including CCTV cameras, mobile phones and smart sensors, can perform some processing rather than sending all data to the edge network; therefore, this model can be helpful to service providers [49].

These diverse factors, comprising technical features, architectures, industrial sectors and business models, are summarized in Table IV–V This proposed taxonomy aims to eliminate the overlapping uses of terminology in the field of edge computing while highlighting the different uses of the various technological developments for future applications in 5G, Smart Cities and Big Data [45]. This analysis can be usefully subdivided into sectors, applications and industry players [61].

Sectors: Moving resources closer to the "fixed edge" represented by surveillance cameras, traffic flow enhancement systems, smart meters, etc., is a feature of FC [13]. Associated long-term data storage can therefore be allocated to cloud computing resources [30]. Mobile MEC users form a transient population with varying requirements [96]. This is a more challenging sector with requirements for hardware and software and functional architectures that are only emerging from demonstration projects [33]. In contrast, mDCs have an already defined segment in the Industrial IoT but could readily be deployed to support FC, MEC and even Cloudlets (although CL would require some form of public funding if open-access platforms are to be implemented) [35].

Applications: Big Data applications would dominate FC, but the focus would be analytics and storage. MEC could generate large amounts of data from users [144], but user requirements for storage and computational resources still need accurate definitions. Far more likely is the use of MEC in informational contexts [52]. A multiplicity of applications would evolve over time, with many s contributing to the widening spectrum of offerings hosted by MEC platforms [15].

Industry players: Edge computing has already demonstrated the active participation of different industrial interests, namely hardware and IT (mDCs and FC) and telecommunications (MEC) [31]. Nascent commercial opportunities have been identified for mDCS and in proof of concept studies for MEC [49]. Edge Computing will be an enabling technology for future wireless systems when trying to meet challenging latency and data rate requirements of applications [34].

7.1. The benefits of edge AI

AI systems are very helpful in settings with real challenges for end customers because they can interpret analogue types of metadata such as language, images, audio, scents, temperatures, emotions, and more [148]. Due to delay, speed, and confidentiality issues, it would be difficult, if not inconceivable, to install these AI applications in a centralised cloud or business data centre [149]. The advantages of AI at the edge include the following.

- **Intelligence:** AI solutions are much more effective and versatile than traditional programmes, which can respond exclusively to signals that the programmer had expected. When it comes to AI [184]. However, neural networks are not taught to answer individual questions but rather to respond to a broad class of questions, regardless of how novel the questions themselves may be. Without AI, it would be impossible for programmes to handle the vast variety of text, speech, and video data that they are asked to process [161].
- **Cost:** By moving to compute resources towards the edge, applications may make do with less bandwidth via the Internet, resulting in significant savings in communication expenses [162].
- **Privacy:** Because AI could examine actual information without revealing it to humans, it is a safer bet for people to entrust their look, tone, medical image, and other sensitive data to be studied in a clinical setting [162]. By keeping such data on-premises, Edge AI additionally protects users' confidentiality, sending just the results of their analyses and observations to the cloud [185]. Users' privacy can be preserved while still benefiting from the training data, even if a portion of it is released. By keeping sensitive information secret, edge AI makes meeting data regulations much easier [163].
- **Availability:** Given that edge AI is both decentralised and can function without an active internet connection, it is inherently more secure [131]. The upshot of this is improved uptime and dependability for AI applications that are essential to business operations [164].

The widespread availability of IoT gadgets, developments in parallel processing, and the advent of 5G have all contributed to the establishment of a solid foundation upon which generalised machine learning may be built [165]. This is enabling businesses to take advantage of the enormous opportunities provided by the incorporation of AI into their operations and the subsequent implementation of real-time intelligence, all while reducing costs and enhancing confidentiality [200]. In spite of the fact that we are currently in the very early stages of edge AI, the potential uses for this technology are imagination [184].

8. Future research directions

Edge AI is still in the early stages of implementation and there are inevitable challenges to process [18]. In the broadest view, there is the requirement to develop experimental test beds for a wide variety of purposes, including deployment scenarios and economic modelling [57, 205]. Similarly, generic processing functionalities would aid developers to mix and match when devising edge solutions [24]. New offerings such as SDN and NFV could be used to achieve dynamic configurations for edge processing [26]. In this section, we focus on open research challenges that are each important for the broad implementation of Edge Computing [147,155].

8.1. Computing resources at the "edge"

Much available hardware for Edge Computing is not designed for intensive computational activities but software solutions have begun to be proposed for example, cloud servers with radio applications to facilitate the functioning of VMs [127].

8.2. Offloading decision making

Assuming that appropriate computational resources are available in the "edge", quantitative decisions on computational offloading must be seamless and automatically made if mobile devices are to function effectively in such an architecture using computation offloading algorithms to find effective and (for mobile devices) energy-efficient solutions [206].

Computational offloading has been a topic explored in depth for MCC scenarios [207,208]. Computational offloading is a means of transferring computing-intensive tasks over to cloud resources to overcome technical limitations in mobile devices, in particular, to improve the battery life of the mobile devices and increase computational performance but also to reduce the total energy consumed [209,210]. Direct testing under defined experimental conditions has confirmed the benefits (shorter processing times and reduced energy consumption) of computational offloading from smartphones [211–213].

While the fine details of computational offloading in Edge Computing may not always precisely mirror those in MCC, the benefits are anticipated to be broadly similar [213–215]. For optimal offloading in MEC, both radio/wireless and computation resources must be considered for multiple users; a study with a single MEC server considered users being able to offload its various proportions of tasks and being allocated only some of the total computation power available but with the aim of minimizing the tie required for each user task [216,217].

8.3. Discovering appropriate edge nodes and resources

Finding and accessing the appropriate nodes in an Edge Computing network and discovery and trust are features that must be considered together, especially for mobile devices [14].

8.4. Ensuring Quality of Experience

Especially for mobile users, overloading edge computational resources is inconsistent with a guaranteed QoE [127,218]. The major challenge may be that of scalability, i.e. designing architectures that rapidly scale to multiple users, especially in MEC [14]. FC and mDCs face much smaller challenges because their users are defined (or more predictable) while Cloudlets may struggle with sudden user surges in city centre environments until predictive mechanisms are developed [140].

Central to MEC were strategies for avoiding overloading and "disaster management" procedures for recovering non-functional base stations [96]. Two different scenarios for MEC overload recovery have been considered and detailed simulation data and results are only presented for a scenario in which a second (recovery) MEC server communicates with disconnected mobile users originally serviced by a MEC that is no longer functional [204]. The envisaged pathway between the recovery MEC server and the disconnected mobile users goes via mobile devices directly connected to this second server. It is assumed that the disconnected mobile users are out of range of the recovery MEC server. The mobile devices directly connected to the recovery MEC server act as relay nodes and communicate with the disconnected mobile users. As the number of relay devices increases and as their rate of data transfer increases, data throughput and the number of disconnected mobile users capable of being serviced increase greater the number of relay devices per disconnected user, the faster the recovery system works [16]. The authors point out, however, that the time taken for the recovery MEC server to calculate the data allocation to be used in the recovery network increases greatly as the number of relay devices and disconnected users increases [138].

8.5. Security

The major challenge to Edge Computing is that of security for users who may have confidence in hardware from manufacturers they trust but have no access to the identities of other components: switches, routers and base stations in publicly accessible modes [114,131]. At the conceptual level, FC and MEC may develop different security protocols and policies because they face different threats and have different weaknesses [78], but advances in one paradigm will be valuable for others [219].

8.6. Edge AI: roadmap for the future

The deployment of AI systems in edge computing as well as other resource-constrained situations, brings problems and possibilities [27]. It is crucial to take into account aspects like computing resources, network connection, data protection, and real-time processing needs when implementing AI systems in these settings [32,217]. By giving specific attention to these considerations, businesses may successfully deploy Intelligent systems in edge computing as well as other resource-constrained situations, gaining the advantages of increased confidentiality, safety, and real-time computation [62].

Moreover, there are a number of challenges that must be overcome in order to successfully implement AI systems in edge computing as well as other resource-limited situations [50,63–65].

- Is it possible for the edge computing devices you're utilizing to handle the processing demands of the AI models you intend to implement?
- Answering questions like "how would you manage the transmission of data and modelling between both the edge computing devices as well as other infrastructures?" and which are the network infrastructure for this transmission?"
- What precautions will you take to protect the confidentiality of personal information that may be handled by AI systems? What hazards may arise, and what steps will you counteract them?

AI systems used in edge computing as well as other low-resource settings [67,68,70,71].

- Self-driving cars employ edge computing to analyse sensor data in real-time and generate controlling and navigation choices.
- Edge computing is used by intelligent home systems to analyse sensing data collected from devices, including webcams and accelerometers and then make adjustments to the house's climate control, illumination, and surveillance [171].
- Edge computing is used by industrial robots to interpret sensor data from production machinery instantaneously, allowing for more precise process control and servicing.

8.6.1. Open issues and trends

Here are some potential open issues and trends for the future of Edge AI.

- Increased adoption: Edge AI is expected to gain popularity and see widespread adoption as more and more devices become equipped with computing power and can process data locally [185].
- Faster processing: Edge devices will be equipped with faster processors, which will enable them to process data more quickly and improve the performance of AI algorithms [200].
- Improved accuracy: Edge AI will see improvements in accuracy due to better data quality, more efficient algorithms, and the ability to incorporate data from multiple sources [201].
- Privacy and security: Edge AI will need to address privacy and security concerns related to processing data on local devices [202]. Techniques such as federated learning, which allows models to be trained locally and without exchanging raw data, will become more important [202].
- Energy efficiency: Edge devices will be designed to be more energy-efficient, as they will rely on batteries for power [201]. Techniques such as model compression and quantization will be used to reduce the energy consumption of AI algorithms [22].
- Edge-to-cloud integration: Edge AI and cloud computing will be integrated to create hybrid solutions that provide the benefits of both approaches [17]. For example, the cloud can be used for training models, while the edge is used for real-time inference [20]. Overall, the future of Edge AI looks promising, as it has the potential to transform industries and enable new applications that were not possible before [25].

9. Summary and conclusions

The similarity of Edge Computing concepts discussed in the literature arises from the main objective of moving computational resources away from centralized remote data centres and from the functionality derived from this transition [9]. The boundaries among those competing paradigms for Edge Computing have been blurred because of the different means of expressing this common aim. In addition, bodies such as ETSI have extended the use of the word "mobile" to include devices such as surveillance cameras which are immobile (but rotatable) [13].

The lack of directly demonstrated benefits of Edge Computing compared to Cloud Computing has been addressed in Section III, where the small number of experimental (hardware) studies on Cloudlets versus Cloud Computing and MEC versus Cloud Computing were discussed.

Differences among Edge Computing paradigms have related to different deployment architectures; for example, where the "edge"

process is placed and which network (industry segment) is the focus of the edge process are major differentiating factors [171]. "Connected devices" - of which there could be 50 billion globally by 2020 [220] - is an ambiguous phrase, covering both personal mobile devices (smartphones, tablets, etc.) and static devices in the IoT, all of which are suitable for various Edge Computing applications. These grey areas have allowed overlapping views that over-claim for particular paradigms; for example [221], defines Edge Computing too narrowly: "connected embedded computing for the Industrial Internet of Things (IIoT)" while [52] brings all mobile users into FC.

In contrast, our analysis of the technical, application and functionality features of Edge Computing, in combination with a discussion of potential applications and business models, points to four major sectors inside Edge Computing [191]: FC is emerging as the optimal solution for IoT applications to minimize the time required for time-critical processing and Big Data analytics; MEC focused on applications intended for genuinely mobile devices, mDCs already available for temporary deployments of industrial applications in specific locations or at temporary sites, while the related but different models presented for "cloudlets" can still be leveraged for publicly funded schemes with free access and open-source software for mobile users in urban environments [6]. With these sectors and technologies differentiated, investment decisions and deployment options can more easily be brought into focus and developers for applications can more confidently tailor solutions to different markets and challenges [7].

New Edge Computing concepts emerged from demonstrable improvements in wireless performance parameters and reduced energy usage (Table 3). The common functions in Edge Computing relate to the three-tier architectures (Fig. 2); these include low latencies and mobility (either human or location-to-location for the IoT) for connected devices (Table 4) and real-time interactions (Table 5); for mobile users in particular, these functions will contribute to an improved QoE. The key features and parameters are summarized in Tables 4–7) and Figs. 1, 2–8.

New technologies are emerging which will be valuable for Edge Computing; these include SDN, NFV and ICN, combining and complementing each other to provide greater network flexibility and reliability [5]. Enhanced functionalities include network resilience, resource management and optimized caching and storing (Fig. 8), while many applications will benefit in a wide spectrum of urban and individualized scenarios (Fig. 7).

With Multi-Access Edge Computing, the seeming conflict between FC and MEC approaches can be removed and an improved focus on the business and commercial opportunities of Edge Computing can be gained [222]. In particular, the implementation of Multi Access Edge Computing has a clear business case for service providers to develop services and products for "vertical" markets, i.e. where an entire sector such as healthcare is considered as a whole; hardware enterprises must also be flexible on precise mechanisms for implementing Multi-Access Edge Computing because a wide range of applications will be recognized with different requirements and market sizes [223,224].

For implementing AI systems in edge computing as well as other locations with limited resources, it is essential to select the appropriate technologies and techniques [72]. For instance, edge computing devices can need specific AI libraries and frameworks that are tailored to function with a constrained set of computer resources [73]. Edge computing devices may be dispersed along a vast geographical region, which calls for remote administration and monitoring. This means that businesses might also have to address the deployment and management of these AI systems [75]. In summary, implementing AI systems in situations with limited resources, such as edge computing and other resource-constrained settings, brings a mix of benefits and constraints [76]. Companies can effectively deploy AI systems in these situations by giving careful consideration to aspects such as computer resources, internetworking, data confidentiality and security, real-time operational needs and selecting the appropriate tools and technologies [80]. This unlocks new possibilities for real-time making decisions as well as enhanced confidentiality and safety [81].

Edge Computing is primarily driven by the wireless networking industry because the plethora of individuals and users will rely on wireless technologies in the imminent 5G world (Tables 5–7). Novel business applications and commercial opportunities are arising from the connected population of users (Table 7). However, major research and industry challenges remain in ensuring the security of personal data and the problems of accommodating very large numbers of heterogeneous user devices (Tables 4–5). These open research areas will require innovative solutions across the IT and telecommunications industries.

Table 8
Acronym glossary.

Acronym	Definition
3G	Third Generation.
5G	Fifth Generation.
LTE	Long Term Evolution
CCTV	Closed-circuit television.
ETSI	European Telecommunications Standards.
FC	Fog Computing.
IIoT	Industrial Internet of Things.
IoT	Internet of Things.
MCC	Mobile Cloud Computing.
mDCs	Micro Data Centres.
MEC	Multi-access Edge Computing.
NFV	Network Function Virtualization.
QoE	Quality of Experience.
RAN	Radio Access Network.
SDN	Software-defined networking.
SLA	Service Level Agreements.
AI	Artificial Intelligence
GPUs	Graphics processing units.
ML	Machine learning.
AR	Augmented reality.
VR	Virtual reality.
M2M	Machine-to-machine
IoE	Internet of Everything
IP	Internet protocol
ICN	Information-Centric Networking
CNNs	convolutional neural networks
RNNs	recurrent neural networks
CPU	Central Processing unit

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Appendix A. List of acronyms

Table 8 shows the list of acronyms.

References

[1] Cisco, Cisco Global Cloud Index: Forecast and Methodology 2015 – 2020, 2015 (white paper). [Online]. Available: http://www.cisco.com/c/dam/en/us/solutions/collateral/service-provider/global-cloud-index-gci/white-paper-c11-738085.pdf.
[2] M. Carroll, A. Van Der Merwe, P. Kotze, Secure cloud computing: benefits, risks and controls, in: Information Security South Africa (ISSA), 2011, IEEE, 2011, pp. 1–9.
[3] X. Chen, L. Jiao, W. Li, X. Fu, Efficient multi-user computation offloading for mobile-edge cloud computing, IEEE/ACM Trans. Netw. 24 (5) (2016) 2795–2808.
[4] R. Singh, J. Kovacs, T. Kiss, To offload or not? an analysis of big data offloading strategies from edge to cloud, in: 2022 IEEE World AI IoT Congress, AIIoT), 2022, pp. 46–52.
[5] S. Iftikhar, et al., Ai-based Fog and Edge Computing: A Systematic Review, Taxonomy and Future Directions, Internet of Things, 2022, 100674.
[6] M.S. Aslanpour, A.N. Toosi, C. Cicconetti, B. Javadi, et al., Serverless Edge Computing: Vision and Challenges, in: 2021 Australasian Computer Science Week Multiconference, 2021, pp. 1–10.

[7] M.S. Aslanpour, et al., Performance Evaluation Metrics for Cloud, Fog and Edge Computing: A Review, Taxonomy, Benchmarks and Standards for Future Research, vol. 12, Internet of Things, 2020, 100273.
[8] K. Church, A.G. Greenberg, J.R. Hamilton, On Delivering Embarrassingly Distributed Cloud Services, HotNets, 2008, pp. 55–60.
[9] S.S. Gill, S. Tuli, M. Xu, I. Singh, K.V. Singh, D. Lindsay, S. Tuli, D. Smirnova, M. Singh, U. Jain, et al., Transformative Effects of Iot, Blockchain and Artificial Intelligence on Cloud Computing: Evolution, Vision, Trends and Open Challenges, vol. 8, Internet of Things, 2019, 100118.
[10] V. Bahl, Emergence of micro data center (cloudlets/edges) for mobile computing [Online]. Available: https://www.microsoft.com/en-us/research/wp-content/uploads/2016/11/Micro-Data-Centers-mDCs-for-Mobile-Computing-1.pdf, 2015.
[11] M. Satyanarayanan, P. Bahl, R. Caceres, N. Davies, The case for vm-based cloudlets in mobile computing, IEEE pervasive Comput. 8 (4) (2009).
[12] F. Bonomi, R. Milito, J. Zhu, S. Addepalli, Fog computing and its role in the internet of things, in: Proceedings of the First Edition of the MCC Workshop on Mobile Cloud Computing, ACM, 2012, pp. 13–16.
[13] J. Singh, et al., Fog computing: a taxonomy, systematic review, current trends and research challenges, J. Parallel Distr. Comput. 157 (2021) 56–85.
[14] P. Garcia Lopez, A. Montresor, D. Epema, A. Datta, T. Higashino, A. Iamnitchi, M. Barcellos, P. Felber, E. Riviere, Edge-centric computing: vision and challenges, Comput. Commun. Rev. 45 (5) (2015) 37–42.
[15] M. Patel, Mobile-edge computing – introductory technical white paper [Online]. Available: https://portal.etsi.org/portals/0/tbpages/mec/docs/mobile-edge_computing_-_introductory_technical_white_paper_v1, 2014.
[16] S. Iftikhar, M.M.M. Ahmad, et al., Hunterplus: ai based energy-efficient task scheduling for cloud–fog computing environments, Internet Things 21 (2023), 100667.
[17] A. Chakraborty, et al., Journey from cloud of things to fog of things: survey, new trends, and research directions, Software Pract. Ex. 53 (2) (2023) 496–551.
[18] Y. Teoh, et al., Iot and Fog Computing Based Predictive Maintenance Model for Effective Asset Management in Industry 4.0 Using Machine Learning, IEEE Internet of Things Journal, 2021.
[19] Y. Shi, K. Yang, T. Jiang, J. Zhang, K.B. Letaief, Communication-efficient edge ai: algorithms and systems, IEEE Commun. Surv. Tutor. 22 (4) (2020) 2167–2191.
[20] M. Kamruzzaman, New opportunities, challenges, and applications of edge-ai for connected healthcare in smart cities, in: 2021 IEEE Globecom Workshops (GC Wkshps), IEEE, 2021, pp. 1–6.
[21] X. Wang, Y. Han, V.C. Leung, D. Niyato, X. Yan, X. Chen, Edge AI: Convergence of Edge Computing and Artificial Intelligence, Springer, 2020.
[22] S. Kalapothas, G. Flamis, P. Kitsos, Efficient edge-ai application deployment for fpgas, Information 13 (6) (2022) 279.
[23] A.Y. Ding, E. Peltonen, T. Meuser, A. Aral, C. Becker, S. Dustdar, T. Hiessl, D. Kranzlmüller, M. Liyanage, S. Maghsudi, et al., Roadmap for Edge Ai: A Dagstuhl Perspective, 2022, pp. 28–33.
[24] J. Yang, T. Baker, et al., A federated learning attack method based on edge collaboration via cloud, Software Pract. Ex. (2022) 1–18, https://doi.org/10.1002/spe.3180.
[25] T. Rausch, W. Hummer, V. Muthusamy, A. Rashed, S. Dustdar, Towards a Serverless Platform for Edge Ai, HotEdge, 2019.
[26] M. H. Anwar et al., "Recommender system for optimal distributed deep learning in cloud datacenters," Wireless Pers. Commun., pp. 1–25.
[27] P. Porambage, T. Kumar, M. Liyanage, J. Partala, L. Lovén, M. Ylianttila, T. Seppänen, Sec-edgeai: Ai for Edge Security vs Security for Edge Ai," *the 1st 6G Wireless Summit*, 2019 (Levi, Finland).
[28] E. Li, L. Zeng, Z. Zhou, X. Chen, Edge ai: on-demand accelerating deep neural network inference via edge computing, IEEE Trans. Wireless Commun. 19 (1) (2019) 447–457.
[29] X. Wang, Y. Han, C. Wang, Q. Zhao, X. Chen, M. Chen, In-edge ai: intelligentizing mobile edge computing, caching and communication by federated learning, Ieee Netw. 33 (5) (2019) 156–165.
[30] M. Xu, C. Song, et al., Coscal: Multi-Faceted Scaling of Microservices with Reinforcement Learning, IEEE Transactions on Network and Service Management, 2022.
[31] S.A. Moqurrab, N. Tariq, et al., A deep learning-based privacy-preserving model for smart healthcare in internet of medical things using fog computing, Wireless Pers. Commun. 126 (3) (2022) 2379–2401.
[32] K.B. Letaief, Y. Shi, J. Lu, J. Lu, Edge artificial intelligence for 6g: vision, enabling technologies, and applications, IEEE J. Sel. Area. Commun. 40 (1) (2021) 5–36.
[33] M. Xu, C. Song, et al., esdnn: deep neural network based multivariate workload prediction in cloud computing environments, ACM Trans. Internet Technol. 22 (3) (2022) 1–24.
[34] A. Raychaudhuri, et al., Green internet of things using mobile cloud computing: architecture, applications, and future directions, in: Green Mobile Cloud Computing, Springer, 2022, pp. 213–229.
[35] S. Ghafouri, A. Karami, et al., Mobile-kube: mobility-aware and energy-efficient service orchestration on kubernetes edge servers, in: 15th IEEE/ACM International Conference on Utility and Cloud Computing (UCC 2022), Washington State University, Portland, OR, United States, 2022. December 6-9, 2022.
[36] M. Adhikari, A. Hazra, V.G. Menon, B.K. Chaurasia, S. Mumtaz, A roadmap of next-generation wireless technology for 6g-enabled vehicular networks, IEEE Internet Things Magaz. 4 (4) (2021) 79–85.
[37] B. Varghese, N. Wang, S. Barbhuiya, P. Kilpatrick, D.S. Nikolopoulos, Challenges and opportunities in edge computing, in: Smart Cloud (SmartCloud), IEEE International Conference on, IEEE, 2016, pp. 20–26.
[38] S. Iftikhar, et al., Tesco: multiple simulations based ai-augmented fog computing for qos optimization, in: The 22nd IEEE International Conference on Scalable Computing and Communications, ScalCom, Hainan, China, 2022, 15-18 December 2022, 2022.
[39] W. Yu, F. Liang, X. He, W.G. Hatcher, C. Lu, J. Lin, X. Yang, A survey on the edge computing for the internet of things, IEEE Access 6 (2017) 6900–6919.
[40] W.Z. Khan, E. Ahmed, S. Hakak, I. Yaqoob, A. Ahmed, Edge computing: a survey, Future Generat. Comput. Syst. 97 (2019) 219–235.
[41] C.-H. Hong, B. Varghese, Resource management in fog/edge computing: a survey on architectures, infrastructure, and algorithms, ACM Comput. Surv. 52 (5) (2019) 1–37.
[42] Q. Luo, S. Hu, C. Li, G. Li, W. Shi, Resource scheduling in edge computing: a survey, IEEE Commun. Surv. Tutor. 23 (4) (2021) 2131–2165.
[43] G. Carvalho, B. Cabral, V. Pereira, J. Bernardino, Edge computing: current trends, research challenges and future directions, Computing 103 (2021) 993–1023.
[44] A. Dhillon, et al., Iotpulse: machine learning-based enterprise health information system to predict alcohol addiction in Punjab (India) using iot and fog computing, Enterprise Inf. Syst. 16 (7) (2022), 1820583.
[45] T. Shao, et al., Iot-pi: a machine learning-based lightweight framework for cost-effective distributed computing using iot, Internet Technol. Lett. 5 (3) (2022), e355.
[46] S. Iftikhar, et al., Fog computing based router-distributor application for sustainable smart home, in: 2022 IEEE 95th Vehicular Technology Conference: (VTC2022-Spring), IEEE, 2022, pp. 1–5.
[47] S.S. Nabavi, L. Wen, et al., Seagull optimization algorithm based multi-objective vm placement in edge-cloud data centers, Internet Things and Cyber-Phys. Syst. 3 (2023) 28–36.
[48] M. Golec, et al., Aiblock: blockchain based lightweight framework for serverless computing using ai, in: 2022 22nd IEEE International Symposium on Cluster, Cloud and Internet Computing (CCGrid), IEEE, 2022, pp. 886–892.
[49] S.U. Malik, et al., Effort: energy efficient framework for offload communication in mobile cloud computing, Software Pract. Ex. 51 (9) (2021) 1896–1909.
[50] L. Lovén, T. Leppänen, E. Peltonen, J. Partala, E. Harjula, P. Porambage, M. Ylianttila, J. Riekki, Edgeai: a vision for distributed, edge-native artificial intelligence in future 6g networks, 1st 6G wireless summit. (2019) 1–2.
[51] A. Souri, Y. Zhao, M. Gao, A. Mohammadian, J. Shen, E. Al-Masri, A trust-aware and authentication-based collaborative method for resource management of cloud-edge computing in social internet of things, IEEE Trans. Comput. Soc. Syst. (2023) 1–10, https://doi.org/10.1109/TCSS.2023.3241020.
[52] T.H. Luan, L. Gao, Z. Li, Y. Xiang, G. Wei, L. Sun, Fog computing: focusing on mobile users at the edge, arXiv preprint arXiv:1502.01815 (2015), 1-11.
[53] R. Singh, S. Armour, A. Khan, M. Sooriyabandara, G. Oikonomou, Identification of the Key Parameters for Computational Offloading in Multi-Access Edge Computing, in: 2020 IEEE Cloud Summit, 2020, pp. 131–136.
[54] S. Iftikhar, et al., Fogdlearner: a deep learning-based cardiac health diagnosis framework using fog computing, Austral. Comput. Sci. Week 2022 (2022) 136–144.
[55] S.S. Gill, P. Garraghan, R. Buyya, Router: fog enabled cloud based intelligent resource management approach for smart home iot devices, J. Syst. Software 154 (2019) 125–138.
[56] A. Kaur, et al., The future of cloud computing: opportunities, challenges and research trends, in: 2018 2nd International Conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud)(I-SMAC) I-SMAC (IoT in Social, Mobile, Analytics and Cloud)(I-SMAC), 2018 2nd International Conference on, IEEE, 2018, pp. 213–219.
[57] NSF Edge Computing Workshop Committee, Grand Challenges in Edge Computing 2016, 2016.
[58] M. Sriraghavendra, et al., Dosp: A Deadline-Aware Dynamic Service Placement Algorithm for Workflow-Oriented Iot Applications in Fog-Cloud Computing Environments, Energy Conservation Solutions for Fog-Edge Computing Paradigms, 2022, pp. 21–47.
[59] S.S. Nabavi, et al., Tractor: traffic-aware and power-efficient virtual machine placement in edge-cloud data centers using artificial bee colony optimization, Int. J. Commun. Syst. 35 (1) (2022), e4747.
[60] S.S. Gill, M. Xu, C. Ottaviani, P. Patros, R. Bahsoon, A. Shaghaghi, M. Golec, V. Stankovski, H. Wu, A. Abraham, et al., Ai for Next Generation Computing: Emerging Trends and Future Directions, vol. 19, Internet of Things, 2022, 100514.
[61] M. Golec, et al., Ifaasbus: a security-and privacy-based lightweight framework for serverless computing using iot and machine learning, IEEE Trans. Ind. Inf. 18 (5) (2021) 3522–3529.
[62] V. Mazzia, A. Khaliq, F. Salvetti, M. Chiaberge, Real-time apple detection system using embedded systems with hardware accelerators: an edge ai application, IEEE Access 8 (2020) 9102–9114.
[63] T. Sipola, J. Alatalo, T. Kokkonen, M. Rantonen, Artificial Intelligence in the Iot Era: A Review of Edge Ai Hardware and Software," in *2022 31st Conference of Open Innovations Association (FRUCT)*, IEEE, 2022, pp. 320–331.
[64] S. Soro, Tinyml for ubiquitous edge ai, arXiv preprint arXiv:2102.01255 (2021), MITRE Technical Report MTR200519, 1-26.
[65] W. Zhang, Z. Zhang, S. Zeadally, H.-C. Chao, V.C. Leung, Masm: a multiple-algorithm service model for energy-delay optimization in edge artificial intelligence, IEEE Trans. Ind. Inf. 15 (7) (2019) 4216–4224.
[66] P. Singh, et al., Machine learning for cloud, fog, edge and serverless computing environments: comparisons, performance evaluation benchmark and future directions, Int. J. Grid Util. Comput. 13 (4) (2022) 447–457.

[67] L. Chang, Z. Zhang, P. Li, S. Xi, W. Guo, Y. Shen, Z. Xiong, J. Kang, D. Niyato, X. Qiao, et al., 6g-enabled edge ai for metaverse: challenges, methods, and future research directions, J. Commun. Inf. Netw. 7 (2) (2022) 107–121.
[68] D. Liu, X. Chen, Z. Zhou, Q. Ling, Hiertrain: fast hierarchical edge ai learning with hybrid parallelism in mobile-edge-cloud computing, IEEE Open J. Commun. Soc. 1 (2020) 634–645.
[69] S.S. Gill, I. Chana, R. Buyya, Iot based agriculture as a cloud and big data service: the beginning of digital India, J. Organ. End User Comput. 29 (4) (2017) 1–23.
[70] R. Bibi, Y. Saeed, A. Zeb, T.M. Ghazal, T. Rahman, R.A. Said, S. Abbas, M. Ahmad, M.A. Khan, Edge ai-based automated detection and classification of road anomalies in vanet using deep learning, Comput. Intell. Neurosci. 2021 (2021) 1–16.
[71] X. Lin, J. Li, J. Wu, H. Liang, W. Yang, Making knowledge tradable in edge-ai enabled iot: a consortium blockchain-based efficient and incentive approach, IEEE Trans. Ind. Inf. 15 (12) (2019) 6367–6378.
[72] P. McEnroe, S. Wang, M. Liyanage, A Survey on the Convergence of Edge Computing and Ai for Uavs: Opportunities and Challenges, IEEE Internet of Things Journal, 2022.
[73] F. Foukalas, A. Tziouvaras, Edge artificial intelligence for industrial internet of things applications: an industrial edge intelligence solution, IEEE Indu. Electron. Magaz. 15 (2) (2021) 28–36.
[74] Y. Hu, W. Pang, X. Liu, R. Ghosh, B. Ko, W.-H. Lee, R. Govindan, Rim: offloading inference to the edge, in: Proceedings of the International Conference on Internet-Of-Things Design and Implementation, 2021, pp. 80–92.
[75] L. Lv, Z. Wu, L. Zhang, B.B. Gupta, Z. Tian, An edge-ai based forecasting approach for improving smart microgrid efficiency, IEEE Trans. Ind. Inf. 18 (11) (2022) 7946–7954.
[76] T.N. Gia, L. Qingqing, J.P. Queralta, Z. Zou, H. Tenhunen, T. Westerlund, Edge ai in smart farming iot: cnns at the edge and fog computing with lora, in: 2019 *IEEE AFRICON*, IEEE, 2019, pp. 1–6.
[77] Y.-L. Lee, P.-K. Tsung, M. Wu, Techology trend of edge ai, in: 2018 International Symposium on VLSI Design, Automation and Test (VLSI-DAT), 2018, pp. 1–2.
[78] J. Doyle, et al., Blockchainbus: a lightweight framework for secure virtual machine migration in cloud federations using blockchain, Secur. Priv. 5 (2) (2022), e197.
[79] S. Singh, I. Chana, M. Singh, The journey of qos-aware autonomic cloud computing, It Profession. 19 (2) (2017) 42–49.
[80] A. Nawaz, T.N. Gia, J.P. Queralta, T. Westerlund, Edge ai and blockchain for privacy-critical and data-sensitive applications, in: In 2019 Twelfth International Conference on Mobile Computing and Ubiquitous Network (ICMU), IEEE, 2019, pp. 1–2.
[81] H. Yang, K.-Y. Lam, L. Xiao, Z. Xiong, H. Hu, D. Niyato, H. Vincent Poor, Lead federated neuromorphic learning for wireless edge artificial intelligence, Nat. Commun. 13 (1) (2022) 4269.
[82] A.C. Chen Liu, O.M.K. Law, J. Liao, J.Y. Chen, A.J. En Hsieh, C.H. Hsieh, Traffic safety system edge ai computing, in: 2021 IEEE/ACM Symposium on Edge Computing, SEC), 2021, 01–02.
[83] G.K. Agarwal, M. Magnusson, A. Johanson, Edge ai driven technology advancements paving way towards new capabilities, Int. J. Innovat. Technol. Manag. 18 (1) (2021), 2040005.
[84] M. Li, F.R. Yu, P. Si, H. Yao, E. Sun, Y. Zhang, Energy-efficient m2m communications with mobile edge computing in virtualized cellular networks, in: 2017 IEEE International Conference on Communications, ICC), 2017, pp. 1–6.
[85] R. Singh, S. Armour, A. Khan, M. Sooriyabandara, G. Oikonomou, The advantage of computation offloading in multi-access edge computing, in: 2019 Fourth International Conference on Fog and Mobile Edge Computing, FMEC), 2019, pp. 289–294.
[86] S. Nunna, A. Kousaridas, M. Ibrahim, M. Dillinger, C. Thuemmler, H. Feussner, A. Schneider, Enabling real-time context-aware collaboration through 5g and mobile edge computing, in: Information Technology-New Generations (ITNG), 2015 12th International Conference on, IEEE, 2015, pp. 601–605.
[87] P. Corcoran, S.K. Datta, Mobile-edge computing and the internet of things for consumers: extending cloud computing and services to the edge of the network, IEEE Consumer Electron. Magaz. 5 (4) (2016) 73–74.
[88] S.S. Gill, A manifesto for modern fog and edge computing: vision, new paradigms, opportunities, and future directions, in: Operationalizing Multi-Cloud Environments: Technologies, Tools and Use Cases, Springer, 2021, pp. 237–253.
[89] M. Sri Raghavendra, et al., Deedsp: deadline-aware and energy-efficient dynamic service placement in integrated internet of things and fog computing environments, Trans. Emerg. Telecommun. Technol. 32 (12) (2021), e4368.
[90] A. Sengupta, et al., Mobile Edge Computing Based Internet of Agricultural Things: a Systematic Review and Future Directions, Mobile Edge Computing, 2021, pp. 415–441.
[91] S. Tuli, et al., Hunter: ai based holistic resource management for sustainable cloud computing, J. Syst. Software 184 (2022), 111124.
[92] M. Kumar, et al., Experimental performance analysis of cloud resource allocation framework using spider monkey optimization algorithm, Concurrency Comput. Pract. Ex. 35 (2) (2023), e7469.
[93] A.K. Bhardwaj, et al., Heart: unrelated parallel machines problem with precedence constraints for task scheduling in cloud computing using heuristic and meta-heuristic algorithms, Software Pract. Ex. 50 (12) (2020) 2231–2251.
[94] M. Satyanarayanan, Cloudlets: at the leading edge of cloud-mobile convergence, in: Proceedings of the 9th International ACM Sigsoft Conference on Quality of Software Architectures, ACM, 2013, pp. 1–2.
[95] M. Satyanarayanan, Z. Chen, K. Ha, W. Hu, W. Richter, P. Pillai, Cloudlets: at the leading edge of mobile-cloud convergence, in: Mobile Computing, Applications and Services (MobiCASE), 2014 6th International Conference on, IEEE, 2014, pp. 1–9.
[96] D. Lindsay, et al., The evolution of distributed computing systems: from fundamental to new frontiers, Computing 103 (8) (2021) 1859–1878.
[97] M. Satyanarayanan, P. Simoens, Y. Xiao, P. Pillai, Z. Chen, K. Ha, W. Hu, B. Amos, Edge analytics in the internet of things, IEEE Pervasive Comput. 14 (2) (2015) 24–31.
[98] K. Ha, P. Pillai, W. Richter, Y. Abe, M. Satyanarayanan, Just-in-time provisioning for cyber foraging, in: Proceeding of the 11th Annual International Conference on Mobile Systems, Applications, and Services, ACM, 2013, pp. 153–166.
[99] Y. Li, W. Wang, Can mobile cloudlets support mobile applications?, in: Infocom, 2014 Proceedings Ieee IEEE, 2014, pp. 1060–1068.
[100] S. Clinch, J. Harkes, A. Friday, N. Davies, M. Satyanarayanan, How close is close enough? understanding the role of cloudlets in supporting display appropriation by mobile users, in: Pervasive Computing and Communications (PerCom), 2012 IEEE International Conference on, IEEE, 2012, pp. 122–127.
[101] T. Verbelen, P. Simoens, F. De Turck, B. Dhoedt, Cloudlets: bringing the cloud to the mobile user, in: Proceedings of the Third ACM Workshop on Mobile Cloud Computing and Services, ACM, 2012, pp. 29–36.
[102] K. Habak, M. Ammar, K.A. Harras, E. Zegura, Femto clouds: leveraging mobile devices to provide cloud service at the edge, in: 2015 IEEE 8th International Conference on Cloud Computing (CLOUD), 2015, pp. 9–16.
[103] S. Tuli, et al., ithermofog: iot-fog based automatic thermal profile creation for cloud data centers using artificial intelligence techniques, Internet Technol. Lett. 3 (5) (2020), e198.
[104] Cisco, Cisco Global Cloud Index: Forecast and Methodology, 2015–2020, 2015 [Online]. Available: http://www.cisco.com/c/dam/en/us/solutions/collateral/service-provider/global-cloud-index-gci/white-paper-c11-738085.pdf.
[105] K. Saharan, A. Kumar, Fog in comparison to cloud: a survey, Int. J. Comput. Appl. 122 (3) (2015).
[106] I. Stojmenovic, S. Wen, The fog computing paradigm: scenarios and security issues, in: Proc 2014 Federated Conference Computer Science and Information Systems, IEEE, 2014, pp. 1–8.
[107] L.M. Vaquero, L. Rodero-Merino, Finding your way in the fog: towards a comprehensive definition of fog computing, Comput. Commun. Rev. 44 (5) (2014) 27–32.
[108] F. Bonomi, R. Milito, P. Natarajan, J. Zhu, Fog computing: a platform for internet of things and analytics, in: Big Data and Internet of Things: A Roadmap for Smart Environments, Springer, 2014, pp. 169–186.
[109] G.I. Klas, Fog Computing and Mobile Edge Cloud Gain Momentum, yucianga.info, 2015.
[110] M. Chiang, Fog networking: an overview on research opportunities, arXiv preprint arXiv:1601.00835 (2015), 1-11.
[111] A.V. Dastjerdi, H. Gupta, R.N. Calheiros, S.K. Ghosh, R. Buyya, Fog computing: principles, architectures, and applications, Internet of things. Morgan Kaufmann, (2016) 61–75.
[112] A.V. Natraj, Fog computing" focusing on users at the edge of internet of things, Int. J. Eng. Res. Special 5 (2016) 992–1128.
[113] R. Suryawanshi, G. Mandlik, Focusing on mobile users at edge and internet of things using fog computing, Int. J. Sci. Eng. Technol. Res. 4 (17) (2015) 3225–3231.
[114] M. Golec, et al., Biosec: a biometric authentication framework for secure and private communication among edge devices in iot and industry 4.0, IEEE Consumer Electron. Magaz. 11 (2) (2020) 51–56.
[115] A.T. Tran, R.C. Palacios, A systematic literature review of fog computing, Norsk konferanse for organisasjoners bruk av IT 24 (1) (2016).
[116] F. Desai, D. Chowdhury, et al., Healthcloud: a system for monitoring health status of heart patients using machine learning and cloud computing, Internet Things 17 (2022), 100485.
[117] R. Craciunescu, A. Mihovska, M. Mihaylov, S. Kyriazakos, R. Prasad, S. Halunga, Implementation of fog computing for reliable e-health applications, in: 49th Asilomar Conference on Signals, Systems and Computers, IEEE, 2015, pp. 459–463, 2015.
[118] O. Fratu, C. Pena, R. Craciunescu, S. Halunga, Fog computing system for monitoring mild dementia and copd patients-Romanian case study, in: 12th International Conference on Telecommunication in Modern Satellite, Cable and Broadcasting Services (TELSIKS), IEEE, 2015, pp. 123–128, 2015.
[119] Y. Shi, G. Ding, H. Wang, H.E. Roman, S. Lu, The fog computing service for healthcare, in: Proc 2nd International Symposium onFuture Information and Communication Technologies for Ubiquitous HealthCare, IEEE, 2015, pp. 1–5, 2015.
[120] M. Ahmad, M.B. Amin, S. Hussain, B.H. Kang, T. Cheong, S. Lee, Health fog: a novel framework for health and wellness applications, J. Supercomput. 72 (10) (2016) 3677–3695.
[121] F.Y. Okay, S. Ozdemir, A fog computing based smart grid model, in: 2016 International Symposium on Networks, Computers and Communications (ISNCC), 2016, pp. 1–6.
[122] X. Hou, Y. Li, M. Chen, D. Wu, D. Jin, S. Chen, Vehicular fog computing: a viewpoint of vehicles as the infrastructures, IEEE Trans. Veh. Technol. 65 (6) (2016) 3860–3873.
[123] S. Yi, Z. Hao, Z. Qin, Q. Li, Fog computing: platform and applications, in: Hot Topics in Web Systems and Technologies (HotWeb), 2015 Third IEEE Workshop on, IEEE, 2015, pp. 73–78.
[124] E. Borcoc, Fog computing, mobile edge computing, cloudlets - which one? [Online]. Available: https://www.iaria.org/conferences2016/filesICSNC16/Softnet2016_Tutorial_Fog-MEC-Cloudlets-E.Borcoci-v1.1.pdf, 2016.

[125] H. Li, G. Shou, Y. Hu, Z. Guo, Mobile edge computing: progress and challenges, in: Proc 4th IEEE International Conference on Mobile Cloud Computing, Services, and Engineering (MobileCloud), IEEE, 2016, pp. 83–84, 2016.
[126] X. Chen, L. Jiao, W. Li, X. Fu, Efficient multi-user computation offloading for mobile-edge cloud computing, IEEE/ACM Trans. Netw. 24 (5) (2016) 2795–2808.
[127] M.T. Beck, M. Werner, S. Feld, S. Schimper, Mobile edge computing: a taxonomy, in: Proc. Of the Sixth International Conference on Advances in Future Internet, Citeseer, 2014.
[128] K. Bansal, et al., Deepbus: machine learning based real time pothole detection system for smart transportation using iot, Internet Technol. Lett. 3 (3) (2020), e156.
[129] S. Tuli, et al., Next generation technologies for smart healthcare: challenges, vision, model, trends and future directions, Internet Technol. Lett. 3 (2) (2020), e145.
[130] ETSI, Mec proofs of concept [Online]. Available: https://www.etsi.org/technologies/multi-access-edge-computing/mec-poc.
[131] A. Kumar, et al., Securing the future internet of things with post-quantum cryptography, Secur. Priv. 5 (2) (2022), e200.
[132] G. Brown, Mobile Edge Computing Use Cases and Deployment Options, 2016.
[133] D. Sabella, A. Vaillant, P. Kuure, U. Rauschenbach, F. Giust, Mobile-edge computing architecture: the role of mec in the internet of things, IEEE Consumer Electron. Magaz. 5 (4) (2016) 84–91.
[134] Y.C. Hu, M. Patel, D. Sabella, N. Sprecher, V. Young, Mobile Edge Computing—A Key Technology towards 5g, vol. 11, ETSI White Paper, 2015.
[135] G. Orsini, D. Bade, W. Lamersdorf, Computing at the mobile edge: designing elastic android applications for computation offloading, in: IFIP Wireless and Mobile Networking Conference (WMNC), IEEE, 2015, pp. 112–119, 2015 8th.
[136] C. Vallati, A. Virdis, E. Mingozzi, G. Stea, Mobile-edge computing come home connecting things in future smart homes using lte device-to-device communications, IEEE Consumer Electron. Magaz. 5 (4) (2016) 77–83.
[137] R. Schuster, P. Ramchandran, Open edge computing-from vision to reality-[Online]. Available: https://wiki.opnfv.org/display/EVNT/Berlin+Design+Summit?preview=, 2016.
[138] S. S. Gill, "Quantum and Blockchain Based Serverless Edge Computing: A Vision, Model, New Trends and Future Directions," Internet Technology Letters, p. e275.
[139] W. Gao, Opportunistic peer-to-peer mobile cloud computing at the tactical edge, in: Military Communications Conference (MILCOM), IEEE, 2014, pp. 1614–1620, 2014 *IEEE*.
[140] S. Singh et al., "An Iiot Based Secure and Sustainable Smart Supply Chain System Using Sensor Networks," Transactions on Emerging Telecommunications Technologies, p. e4681.
[141] Y. Jararweh, L. Tawalbeh, F. Ababneh, F. Dosari, Resource efficient mobile computing using cloudlet infrastructure, in: Mobile Ad-Hoc and Sensor Networks (MSN), 2013 IEEE Ninth International Conference on, IEEE, 2013, pp. 373–377.
[142] D. Fesehaye, Y. Gao, K. Nahrstedt, G. Wang, Impact of cloudlets on interactive mobile cloud applications, in: Enterprise Distributed Object Computing Conference (EDOC), 2012 IEEE 16th International, IEEE, 2012, pp. 123–132.
[143] W. Hu, Y. Gao, K. Ha, J. Wang, B. Amos, Z. Chen, P. Pillai, M. Satyanarayanan, Quantifying the impact of edge computing on mobile applications, in: Proceedings of the 7th ACM SIGOPS Asia-Pacific Workshop on Systems, ACM, 2016, p. 5.
[144] A. Ceselli, M. Premoli, S. Secci, Cloudlet network design optimization, in: Proc IFIP Networking Conference, 2015, pp. 1–9.
[145] S. Tuli, et al., Healthfog: an ensemble deep learning based smart healthcare system for automatic diagnosis of heart diseases in integrated iot and fog computing environments, Future Generat. Comput. Syst. 104 (2020) 187–200.
[146] T. Taleb, K. Samdanis, B. Mada, H. Flinck, S. Dutta, D. Sabella, On multi-access edge computing: a survey of the emerging 5g network edge architecture orchestration, IEEE Commun. Surv. Tutor. 99 (2017), 1–1.
[147] A. Kumar, A.S. Yadav, et al., A secure drone-to-drone communication and software defined drone network-enabled traffic monitoring system, Simulat. Model. Pract. Theor. 120 (2022), 102621.
[148] M.M.H. Shuvo, Edge ai: leveraging the full potential of deep learning, in: Recent Innovations in Artificial Intelligence and Smart Applications, Springer, 2022, pp. 27–46.
[149] V.K. Rathi, N.K. Rajput, S. Mishra, B.A. Grover, P. Tiwari, A.K. Jaiswal, M.S. Hossain, An edge ai-enabled iot healthcare monitoring system for smart cities, Comput. Electr. Eng. 96 (2021), 107524.
[150] U. Shaukat, E. Ahmed, Z. Anwar, F. Xia, Cloudlet deployment in local wireless networks: motivation, architectures, applications, and open challenges, J. Netw. Comput. Appl. 62 (2016) 18–40.
[151] S.S. Gill, R.C. Arya, G.S. Wander, R. Buyya, Fog-based smart healthcare as a big data and cloud service for heart patients using iot, in: International Conference on Intelligent Data Communication Technologies and Internet of Things (ICICI) 2018, Springer, 2019, pp. 1376–1383.
[152] Baltimore–Washington International Airport, Bwi marshall's smart park technology: it's all about the lights [Online]. Available: https://parking.bwiairport.com/smart-park-technology/, 2015.
[153] P. Carnelli, J. Yeh, M. Sooriyabandara, A. Khan, Parkus: a novel vehicle parking detection system, in: Innovative Applications of Artificial Intelligence, 2017.
[154] R.A. Najdi, T.G. Shaban, M.J. Mourad, S.H. Karaki, Hydrogen production and filling of fuel cell cars, in: Advances in Computational Tools for Engineering Applications (ACTEA), 2016 3rd International Conference on, IEEE, 2016, pp. 43–48.
[155] A. Kumar, et al., A drone-based networked system and methods for combating coronavirus disease (covid-19) pandemic, Future Generat. Comput. Syst. 115 (2021) 1–19.
[156] S. Carlini, The Drivers and Benefits of Edge Computing, Schneider Electric–Data Center Science Center, 2016, p. 8.
[157] E.S. Udoh, A. Alkharashi, Privacy risk awareness and the behavior of smartwatch users: a case study of Indiana university students, in: Future Technologies Conference (FTC), IEEE, 2016, pp. 926–931.
[158] R.T. Azuma, Augmented Reality: Approaches and Technical Challenges," *Fundamentals of Wearable Computers and Augmented Reality*, 2001, pp. 27–63.
[159] J.K. Zao, T.T. Gan, C.K. You, S.J.R. Méndez, C.E. Chung, Y. Te Wang, T. Mullen, T.P. Jung, Augmented brain computer interaction based on fog computing and linked data, in: Intelligent Environments (IE), 2014 International Conference on, IEEE, 2014, pp. 374–377.
[160] Y. Cao, S. Chen, P. Hou, D. Brown, Fast: a fog computing assisted distributed analytics system to monitor fall for stroke mitigation, in: Networking, Architecture and Storage (NAS), 2015 IEEE International Conference on, IEEE, 2015, pp. 2–11.
[161] R. Ke, Y. Zhuang, Z. Pu, Y. Wang, A smart, efficient, and reliable parking surveillance system with edge artificial intelligence on iot devices, IEEE Trans. Intell. Transport. Syst. 22 (8) (2020) 4962–4974.
[162] R. Marculescu, D. Marculescu, U. Ogras, Edge ai: systems design and ml for iot data analytics, in: Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, 2020, pp. 3565–3566.
[163] R. Sachdev, Towards security and privacy for edge ai in iot/ioe based digital marketing environments, in: 2020 Fifth International Conference on Fog and Mobile Edge Computing (FMEC), IEEE, 2020, pp. 341–346.
[164] C. Surianarayanan, J.J. Lawrence, P.R. Chelliah, E. Prakash, C. Hewage, A survey on optimization techniques for edge artificial intelligence (ai), Sensors 23 (3) (2023) 1279.
[165] L. Stäcker, J. Fei, P. Heidenreich, F. Bonarens, J. Rambach, D. Stricker, C. Stiller, Deployment of deep neural networks for object detection on edge ai devices with runtime optimization, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 1015–1022.
[166] S. Ningning, G. Chao, A. Xingshuo, Z. Qiang, Fog computing dynamic load balancing mechanism based on graph repartitioning, China Commun. 13 (3) (2016) 156–164.
[167] K. Hong, D. Lillethun, U. Ramachandran, B. Ottenwälder, B. Koldehofe, Mobile fog: a programming model for large-scale applications on the internet of things, in: Proceedings of the Second ACM SIGCOMM Workshop on Mobile Cloud Computing, ACM, 2013, pp. 15–20.
[168] H. Dubey, J. Yang, N. Constant, A.M. Amiri, Q. Yang, K. Makodiya, Fog data: enhancing telehealth big data through fog computing, in: Proceedings of the ASE BigData & SocialInformatics 2015, ACM, 2015, p. 14.
[169] P.P. Jayaraman, J.B. Gomes, H.L. Nguyen, Z.S. Abdallah, S. Krishnaswamy, A. Zaslavsky, Cardap: a scalable energy-efficient context aware distributed mobile data analytics platform for the fog, in: East European Conference on Advances in Databases and Information Systems, Springer, 2014, pp. 192–206.
[170] S. Sarkar, S. Misra, Theoretical modelling of fog computing: a green computing paradigm to support iot applications, IET Netw. 5 (2) (2016) 23–29.
[171] M. Singh, S. Tuli, et al., Dynamic shift from cloud computing to industry 4.0: eco-friendly choice or climate change threat, in: IoT-based Intelligent Modelling for Environmental and Ecological Engineering: IoT Next Generation EcoAgro Systems, Springer, 2021, pp. 275–293.
[172] S.K. Datta, C. Bonnet, J. Haerri, Fog computing architecture to enable consumer centric internet of things services, in: Consumer Electronics (ISCE), 2015 IEEE International Symposium on, IEEE, 2015, pp. 1–2.
[173] N.K. Giang, M. Blackstock, R. Lea, V.C. Leung, Developing iot applications in the fog: a distributed dataflow approach, in: Internet of Things (IOT), 2015 5th International Conference on the, IEEE, 2015, pp. 155–162.
[174] I. Abdullahi, S. Arif, S. Hassan, Ubiquitous shift with information centric network caching using fog computing, in: Computational Intelligence in Information Systems, Springer, 2015, pp. 327–335.
[175] S. Jingtao, L. Fuhong, Z. Xianwei, L. Xing, Steiner tree based optimal resource caching scheme in fog computing, China Commun. 12 (8) (2015) 161–168.
[176] L.F. Bittencourt, M.M. Lopes, I. Petri, O.F. Rana, Towards virtual machine migration in fog computing, in: P2P, Parallel, Grid, Cloud and Internet Computing (3PGCIC), 2015 10th International Conference on, IEEE, 2015, pp. 1–8.
[177] M.A. Hassan, M. Xiao, Q. Wei, S. Chen, Help your mobile applications with fog computing, in: Sensing, Communication, and Networking-Workshops (SECON Workshops), 2015 12th Annual IEEE International Conference on, IEEE, 2015, pp. 1–6.
[178] J. Preden, J. Kaugerand, E. Suurjaak, S. Astapov, L. Motus, R. Pahtma, Data to decision: pushing situational information needs to the edge of the network, in: Cognitive Methods in Situation Awareness and Decision Support (CogSIMA), 2015 IEEE International Inter-disciplinary Conference on, IEEE, 2015, pp. 158–164.
[179] M.G.R. Alam, Y.K. Tun, C.S. Hong, Multi-agent and reinforcement learning based code offloading in mobile fog, in: Information Networking (ICOIN), 2016 International Conference on, IEEE, 2016, pp. 285–290.
[180] R. Singh, S. Armour, A. Khan, M. Sooriyabandara, G. Oikonomou, Heuristic approaches for computational offloading in multi-access edge computing networks, in: 2020 IEEE 31st Annual International Symposium on Personal, Indoor and Mobile Radio Communications, 2020, pp. 1–7.
[181] W. Emfinger, A. Dubey, P. Volgyesi, J. Sallai, G. Karsai, Demo abstract: riaps—a resilient information architecture platform for edge computing, in: 2016 IEEE/ACM Symposium on Edge Computing, *SEC)*. IEEE, 2016, pp. 119–120.
[182] R. Frank, Understanding Smart Sensors, Artech House, 2013.
[183] L. Zeng, L. Li, L. Duan, K. Lu, Z. Shi, M. Wang, W. Wu, P. Luo, Distributed data mining: a survey, Inf. Technol. Manag. 13 (4) (2012) 403–409.

[184] Y. Shi, K. Yang, Z. Yang, Y. Zhou, Mobile Edge Artificial Intelligence: Opportunities and Challenges, 2021.
[185] M. Shafique, A. Marchisio, R.V.W. Putra, M.A. Hanif, Towards energy-efficient and secure edge ai: a cross-layer framework iccad special session paper, in: 2021 IEEE/ACM International Conference on Computer Aided Design (ICCAD), IEEE, 2021, pp. 1–9.
[186] S. Mousavi, S.E. Mood, A. Souri, M.M. Javidi, Directed Search: A New Operator in Nsga-Ii for Task Scheduling in Iot Based on Cloud-Fog Computing, IEEE Transactions on Cloud Computing, 2022.
[187] R. Morabito, N. Beijar, Enabling data processing at the network edge through lightweight virtualization technologies, in: Sensing, Communication and Networking (SECON Workshops), 2016 IEEE International Conference on, IEEE, 2016, pp. 1–6.
[188] F. Ramalho, A. Neto, Virtualization at the network edge: a performance comparison, in: World of Wireless, Mobile and Multimedia Networks (WoWMoM), 2016 IEEE 17th International Symposium on A, IEEE, 2016, pp. 1–6.
[189] M. Ghobaei-Arani, A. Souri, A.A. Rahmanian, Resource management approaches in fog computing: a comprehensive review, J. Grid Comput. 18 (1) (2020) 1–42.
[190] T. Guérout, T. Monteil, G. Da Costa, R.N. Calheiros, R. Buyya, M. Alexandru, Energy-aware simulation with dvfs, Simulat. Model. Pract. Theor. 39 (2013) 76–91.
[191] H. Singh, S. Tyagi, et al., Metaheuristics for scheduling of heterogeneous tasks in cloud computing environments: analysis, performance evaluation, and future directions, Simulat. Model. Pract. Theor. 111 (2021), 102353.
[192] H. Gupta, A.V. Dastjerdi, S.K. Ghosh, R. Buyya, ifogsim: a toolkit for modeling and simulation of resource management techniques in internet of things, edge and fog computing environments, *Softw Pract Exper* 47 (2017) 1275–1296.
[193] Cisco, Cisco devnet [Online]. Available: https://developer.cisco.com/site/iox/documents/developer-guide/?ref=overview, 2016.
[194] C. Adjih, E. Baccelli, E. Fleury, G. Harter, N. Mitton, T. Noel, R. Pissard-Gibollet, F. Saint-Marcel, G. Schreiner, J. Vandaele, et al., Fit iot-lab: a large scale open experimental iot testbed, in: Internet of Things (WF-IoT), 2015 IEEE 2nd World Forum on, IEEE, 2015, pp. 459–464.
[195] L. Sanchez, L. Muñoz, J.A. Galache, P. Sotres, J.R. Santana, V. Gutierrez, R. Ramdhany, A. Gluhak, S. Krco, E. Theodoridis, et al., Smartsantander: iot experimentation over a smart city testbed, Comput. Network. 61 (2014) 217–238.
[196] G. Brambilla, M. Picone, S. Cirani, M. Amoretti, F. Zanichelli, A simulation platform for large-scale internet of things scenarios in urban environments, in: Proceedings of the First International Conference on IoT in Urban Space, ICST (Institute for Computer Sciences, Social-Informatics and Telecommunications Engineering), 2014, pp. 50–55.
[197] Y. Jararweh, A. Doulat, A. Darabseh, M. Alsmirat, M. Al-Ayyoub, E. Benkhelifa, Sdmec: software defined system for mobile edge computing, in: Cloud Engineering Workshop (IC2EW), 2016 IEEE International Conference on, IEEE, 2016, pp. 88–93.
[198] B. Ahlgren, C. Dannewitz, C. Imbrenda, D. Kutscher, B. Ohlman, A survey of information-centric networking, IEEE Commun. Mag. 50 (7) (2012).
[199] G. Americas, Understanding Information Centric Networking and Mobile Edge Computing (White Paper), 2016.
[200] C. Delacour, S. Carapezzi, M. Abernot, G. Boschetto, N. Azemard, J. Salles, T. Gil, A. Todri-Sanial, Oscillatory neural networks for edge ai computing, in: 2021 IEEE Computer Society Annual Symposium on VLSI (ISVLSI), IEEE, 2021, pp. 326–331.
[201] H. Yang, J. Wen, X. Wu, L. He, S. Mumtaz, An efficient edge artificial intelligence multipedestrian tracking method with rank constraint, IEEE Trans. Ind. Inf. 15 (7) (2019) 4178–4188.
[202] D. Liu, H. Kong, X. Luo, W. Liu, R. Subramaniam, Bringing ai to edge: from deep learning's perspective, Neurocomputing 485 (2022) 297–320.
[203] R. Singh, T. Kiss, Edge-cloud synergy: unleashing the potential of parallel processing for big data analytics, in: 2022 IEEE 13th Annual Information Technology, Electronics and Mobile Communication Conference, IEMCON), 2022, 0001–0006.
[204] D. Satria, D. Park, M. Jo, Recovery for Overloaded Mobile Edge Computing, Future Generation Computer Systems, 2016.
[205] S. Bhattacharyya, Research on edge computing: a detailed study, Int. J. Inf. Technol. 2 (6) (2016).
[206] Y. Mao, J. Zhang, K.B. Letaief, Dynamic computation offloading for mobile-edge computing with energy harvesting devices, IEEE J. Sel. Area. Commun. 34 (12) (2016) 3590–3605.
[207] H.T. Dinh, C. Lee, D. Niyato, P. Wang, A survey of mobile cloud computing: architecture, applications, and approaches, Wireless Commun. Mobile Comput. 13 (18) (2013) 1587–1611.
[208] W. Zhang, Y. Wen, H.-H. Chen, Toward transcoding as a service: energy-efficient offloading policy for green mobile cloud, IEEE Network 28 (6) (2014) 67–73.
[209] K. Kumar, Y.-H. Lu, Cloud computing for mobile users: can offloading computation save energy? Computer 43 (4) (2010) 51–56.
[210] B.-G. Chun, S. Ihm, P. Maniatis, M. Naik, A. Patti, Clonecloud: elastic execution between mobile device and cloud, in: Proceedings of the Sixth Conference on Computer Systems, ACM, 2011, pp. 301–314.
[211] E. Cuervo, A. Balasubramanian, D.-k. Cho, A. Wolman, S. Saroiu, R. Chandra, P. Bahl, Maui: making smartphones last longer with code offload, in: Proceedings of the 8th International Conference on Mobile Systems, Applications, and Services, ACM, 2010, pp. 49–62.
[212] M.S. Gordon, D.A. Jamshidi, S.A. Mahlke, Z.M. Mao, X. Chen, Comet: code offload by migrating execution transparently, OSDI 12 (2012) 93–106.
[213] K. Kumar, J. Liu, Y.-H. Lu, B. Bhargava, A survey of computation offloading for mobile systems, Mobile Network. Appl. 18 (1) (2013) 129–140.
[214] S. Yi, C. Li, Q. Li, A survey of fog computing: concepts, applications and issues, in: Proceedings of the 2015 Workshop on Mobile Big Data, ACM, 2015, pp. 37–42.
[215] Z. Pang, L. Sun, Z. Wang, E. Tian, S. Yang, A survey of cloudlet based mobile computing, in: Cloud Computing and Big Data (CCBD), 2015 International Conference on, IEEE, 2015, pp. 268–275.
[216] H.Q. Le, H. Al-Shatri, A. Klein, Efficient resource allocation in mobile-edge computation offloading: completion time minimization, in: Information Theory (ISIT), 2017 IEEE International Symposium on, IEEE, 2017, pp. 2513–2517.
[217] S. Singh, I. Chana, A survey on resource scheduling in cloud computing: issues and challenges, J. Grid Comput. 14 (2016) 217–264.
[218] P. Simoens, L. Van Herzeele, F. Vandeputte, L. Vermoesen, Challenges for orchestration and instance selection of composite services in distributed edge clouds, in: Integrated Network Management (IM), 2015 IFIP, IEEE International Symposium on, 2015, pp. 1196–1201.
[219] R. Roman, J. Lopez, M. Mambo, Mobile edge computing, fog et al.: A survey and analysis of security threats and challenges, Future Generation Computer Systems, 2016.
[220] Ericsson, The telecom cloud opportunity [Online]. Available: https://www.ericsson.com/res/site_AU/docs/2012/ericsson_telecom_cloud_discussion_paper.pdf, 2012.
[221] P. Steve Jennis, 2017 Iiot Prediction: Edge Computing Goes Mainstream, 2017.
[222] C. Wehner, Multi-access edge computing. solving tomorrow's problems today [Online]. Available: https://www.artesyn.com/computing/assets/mec_wp_1487813912.pdf, 2017.
[223] F. Rayal, A perspective on multi-access edge computing [Online]. Available: http://frankrayal.com/wp-content/uploads/2014/01/A-Perspective-on-Multi-Access-Edge-Computing.pdf, 2017.
[224] R. Singh, S. Armour, A. Khan, M. Sooriyabandara, G. Oikonomou, Towards multi-criteria heuristic optimization for computational offloading in multi-access edge computing, in: 2021 IEEE 22nd International Conference on High Performance Switching and Routing, HPSR), 2021, pp. 1–6.