Neural Computing and Applications (2024) 36:11461–11494
https://doi.org/10.1007/s00521-024-09723-w

ORIGINAL ARTICLE

Edge artificial intelligence for big data: a systematic review

Atefeh Hemmati[1] · Parisa Raoufi[2] · Amir Masoud Rahmani[3]

Received: 6 December 2022 / Accepted: 25 March 2024 / Published online: 16 April 2024

Abstract
Edge computing, artificial intelligence (AI), and machine learning (ML) concepts have become increasingly prevalent in Internet of Things (IoT) applications. As the number of IoT devices continues to grow, relying solely on cloud computing for real-time data processing and analysis is proving to be more challenging. The synergy between edge computing and AI is particularly intriguing due to AI's reliance on rapid data processing, a capability facilitated by edge computing. Edge AI represents a significant paradigm shift, leveraging AI within edge computing frameworks to reduce reliance on internet connections and mitigate data latency issues. This approach accelerates data processing, supporting use cases that demand real-time inference. Additionally, as cloud storage costs continue to rise, the feasibility of streaming and storing large volumes of data comes into question. Edge AI offers a compelling solution by performing big data analytics closer to the end device where edge computing is deployed. This paper presents a systematic literature review (SLR) of 85 articles published between 2018 and 2023 within Edge AI. The study provides a comprehensive examination of the analysis of measurement environments and assesses factors applied to Edge AI for big data. It offers taxonomies specific to Edge AI within the big data domain, presents case studies, and outlines the challenges and open issues inherent in Edge AI for big data.

Keywords Internet of Things · Edge computing · Artificial intelligence · Machine learning · Big data

1 Introduction

The advancement of technology has rendered nearly everything digital and prompted disruptive innovation. This digital revolution is fueled by cutting-edge technology, utilizing big data and artificial intelligence (AI), which extend from machine learning (ML) systems and the Internet of Things (IoT) to emerging edge computing technologies. To foster significant growth in the smart manufacturing sector, new manufacturing technologies and innovations such as big data, AI, the IoT, and edge computing must be collaboratively integrated. This integration is crucial for generating compelling, remarkable, pioneering advancements in smart manufacturing businesses [1].

Big data encompasses processes involving collecting, extracting, and analyzing complex datasets that defy conventional tools to achieve objectives. While big data refers solely to the data, a significant portion of this process falls under data science, data mining, data analysis, ML, and AI. Often mentioned alongside concepts like data mining or business intelligence, big data serves as the primary foundation for conducting advanced analyses. Big data comprises a growing volume of diverse datasets generated at an exceedingly high velocity [2]. This dataset is characterized by three primary features: volume, variety, and accelerated production speed. The sheer volume and rapid production pace of big data have surpassed the capabilities of traditional data processing hardware and software, rendering them inadequate for effective management [3].

Big data is increasingly pivotal in powering today's advanced analytics technologies, notably AI and ML. Big data analytics entails meticulously examining vast datasets

✉ Amir Masoud Rahmani
rahmania@yuntech.edu.tw

[1] Department of Computer Engineering, Science and Research Branch, Islamic Azad University, Tehran, Iran

[2] Department of Computer Engineering, Central Tehran Branch, Islamic Azad University, Tehran, Iran

[3] Future Technology Research Center, National Yunlin University of Science and Technology, Douliou, Yunlin 64002, Taiwan

to unearth hidden patterns, correlations, and other valuable insights. The reality is that the generation of data by humanity shows no signs of abating; it continues to increase unabated [3, 4].

Cloud computing manages a substantial volume of present-day data, and while its capabilities are commendable, it is not devoid of challenges. One such challenge is the persistent threat to cloud security that organizations constantly face. Addressing this data processing challenge goes beyond merely expanding cloud capacity. Instead, the solution lies in embracing edge computing and Edge AI technologies. Edge AI enhances the speed, security, and efficiency of data processing, providing a robust approach to mitigate the risks associated with cloud security [4]. Edge AI refers to the processing of ML algorithms locally [3]. Contrary to eliminating cloud computing, Edge AI complements and strengthens cloud computing capabilities [5].

Data storage and processing have become notably more user-friendly with the advent of edge computing. This is achieved by executing tasks on local devices, IoT devices, or specialized edge servers. Unlike cloud-based services, edge services are not susceptible to delays and bandwidth constraints, which often hinder efficiency [2, 3].

Edge AI, an integration of AI with edge computing, involves running AI algorithms on nearby devices equipped with edge computing capabilities. This allows users to instantaneously process data on their devices without internet connectivity. Currently, most AI operations are centralized in cloud-based data centers, requiring significant processing power. However, this centralized approach is vulnerable to interruptions or slowdowns due to connection or network issues. Edge AI addresses these challenges by integrating AI operations into edge computing hardware. By collecting data and serving users locally without interacting with remote physical locations, Edge AI saves users time and enhances overall efficiency [2, 3, 6].

Edge AI is rapidly transitioning from being a preference to becoming necessary for emerging products and services, such as self-driving vehicles and smart home devices. Leveraging edge computing, algorithms can execute and perform computationally intensive tasks like image segmentation locally on IoT devices rather than relying on cloud-based processing [3]. The adoption of Edge AI offers several advantages, including enhanced security, cost-effectiveness, and reliability within interconnected systems. Unlike cloud-based solutions, Edge AI provides a fail-safe computational approach that remains unaffected by network irregularities and data breaches, which are more prone to occur in cloud environments [4, 5].

So far, numerous articles have extensively covered the topics of AI and edge computing. However, it is noteworthy that comprehensive and systematic literature reviews focusing specifically on Edge AI and big data are currently lacking. While some review works in the Edge AI domain exist, they often have limitations. For instance, Chang et al. [3] explored a coordinated end-edge-cloud architecture for flexible AI of Things (AIoT) systems. They discussed the latest advancements in edge network training and inference for AI models. Nevertheless, this review lacks in-depth insights into the field and may overlook crucial aspects such as detailed study methodologies.

Furthermore, Hua et al. [4] provided a formal exposition on edge computing and elucidated its rising prominence as a computing architecture. Their paper offers an overview of extending AI into various domains and optimizing edge computing. It is a roadmap for exploring diverse research topics and leveraging the synergy between AI and edge computing. However, notable limitations include the absence of a comprehensive taxonomy and insufficient coverage of research literature. Similarly, Zhou et al. [5] reviewed research initiatives in edge intelligence, starting with discussing the background and objectives of AI operating at the network edge. While their work contributes to understanding this emerging field, it also suffers from shortcomings, such as lacking a taxonomy and inadequate coverage of relevant articles.

We conducted an SLR article on Edge AI for big data to address existing shortcomings, offer a thorough and detailed examination of Edge AI, and explore its intersection with big data. Our main contributions include:

- Analyzing measurement environments and evaluation factors usually applied to Edge AI for big data
- Presenting three new taxonomies in Edge AI for the big data domain
- Identifying challenges and open issues in Edge AI for big data

The subsequent sections of this paper consist of related work and an overview of the Edge AI concept discussed in Sect. 2. Section 3 outlines the research methodology and details the selection and review of relevant research articles. A comparison and categorization of Edge AI for big data are presented in Sect. 5, followed by a discussion and comparative analysis in Sect. 6. Lastly, Sect. 7 concludes the paper.

2 Related work and overview of the Edge AI concept

In this section, we review related work in Sect. 2.1. Additionally, in Sect. 2.2, we present an overview of the concepts associated with Edge AI.

2.1 Overview of Edge AI concepts

In this subsection, we will delve into the Edge AI concept, explore how Edge AI operates, highlight the benefits of Edge AI, and examine the intersection of Edge AI with big data.

2.1.1 Edge AI concept

The adoption of edge computing has experienced notable growth in recent years, partly propelled by advancements in AI technology. As IoT and corporate data applications gain popularity, there is a growing demand for devices capable of handling data swiftly and intelligently [2]. Edge AI, which merges edge computing with AI capabilities, empowers edge devices to employ AI techniques for processing information at the edge, thus reducing decision-making latency [4]. Consequently, Edge AI integrates AI into edge computing devices to enhance data processing speed and enable intelligent automation [3].

Some individuals mistakenly equate Edge AI with distributed AI, considering them identical concepts. However, Descriptive AI enables scalable data usage and AI deployment beyond centralized and cloud environments. Unlike distributed AI, Descriptive AI does not necessitate aggregating all data in one location. In a distributed AI system, processor cores may be geographically dispersed, and AI algorithms can be executed across multiple interconnected processors in diverse configurations. Processor nodes may vary in architecture, enabling the utilization of various clouds or Edge devices collectively for distributed AI operations. Distributed AI is typically applied to address intricate and multifaceted AI challenges [7].

In contrast, Edge AI complements cloud architecture and does not conflict with it. Edge AI entails edge computing, with the distinction being that computations involve AI algorithms. In the evolving landscape of edge functionality alongside the cloud, particularly in AI applications, Edge AI primarily handles inference tasks, while cloud AI is employed for training new models. Inference algorithms require significantly less processing power than training algorithms, making it feasible to implement inference tasks on Edge processors with lower computational capabilities than those required for cloud-based processing [2, 8, 9].

Edge AI involves processing and applying ML algorithms on local hardware. Local computing reduces network latency for data transfer and addresses security concerns by performing tasks on the device.

2.1.2 Edge AI operation

Edge AI integrates AI workflows utilizing data sources at the edge and centralized data centers, such as the cloud and devices. Unlike cloud AI, which typically conducts all operations in the cloud, Edge AI encompasses dedicated edge servers, IoT devices, and remote devices. By enabling data storage and computation locally, Edge AI enhances user accessibility [1, 3].

Edge AI leverages edge computing capabilities by integrating AI algorithms on local devices. This enables data processing and analysis even without connectivity, granting users access to data from diverse sources. By combining the strengths of edge computing and AI, Edge AI minimizes service downtime and latency. Furthermore, it seamlessly integrates AI processes as a crucial component in edge computing devices. Users benefit from time and cost savings, as data aggregation and meeting user needs can be accomplished without interacting with physical locations [1]. Historically, prevalent AI applications were developed using symbolic AI techniques, where rules were encoded into programs. For example, expert systems were trained to handle operations, while AI algorithms were trained to detect fraudulent activities. Optical character recognition (OCR) relies on non-symbolic AI techniques like neural networks to recognize numbers and text [6].

Over time, researchers have discovered new methods to enhance neural networks in the cloud. Additionally, inference has increased, which involves teaching AI models to respond to input data. Edge AI, therefore, enhances the creation and operation of AI processes outside the cloud. It is an excellent inferencing solution, while cloud AI focuses on developing new algorithms. Furthermore, inference algorithms require less processing power than training algorithms [3].

The fundamental elements of Edge AI include:

- **Edge computing:** Edge computing encompasses various processes for gathering, processing, and analyzing data at the edge of a network. Essentially, it enables computation and data storage at the location where data are collected.
- **AI:** AI refers to the capability of machines to mimic human cognitive abilities. This typically involves tasks such as language comprehension and problem-solving. AI entails sophisticated machinery that combines automation with advanced analytical capabilities.

2.1.3 Benefits of Edge AI

By utilizing Edge AI, edge computing leverages AI processes from the cloud to end-user devices. This approach

addresses issues commonly encountered in traditional cloud environments, such as long delays and security concerns. As a result, the advantages offered by Edge AI broaden the horizons for developing new processes and applications [4, 10].

- The primary benefit of Edge AI is the reduction in latency and the increase in processing speed. Carrying out inference locally minimizes delays associated with cloud communication, resulting in shorter response times and increased system availability.
- Furthermore, local processing consumes less bandwidth and is more cost-effective than cloud-based processing. This reduces bandwidth usage by sending less data over the Internet, while cloud-based AI services are more expensive. Edge AI enables using expensive cloud resources as post-processing storage systems, thereby gathering information for future analyses and reducing the cost of real-time updates.
- Edge AI also offers highly reliable autonomous technologies with enhanced data security. Since data are processed locally, there is a lower risk of compromising the sensitivity and confidentiality of the data.
- Additionally, Edge AI ensures the performance of critical operations for automation in scenarios where networks or clouds may be inactive. This reliability is crucial for systems like autonomous vehicles and industrial robots.
- Most importantly, Edge AI saves money on energy costs and promotes power efficiency by processing data locally. It consumes less power overall, ensuring high efficiency in AI operations at the edge.

2.1.4 Edge AI and big data

Today, most connected IoT devices collect data and immediately transmit it to cloud servers. However, storing and processing this influx of data in the cloud can create big data-related challenges. One significant solution to mitigate these challenges is adopting edge computing architecture. Edge computing refers to the outer periphery of the Internet network, where gateways and modems are situated [2, 9, 11]. In edge computing, IoT devices transmit their data to the network and servers in real time, thus alleviating the burden on centralized cloud infrastructure. The synergy between Edge AI and big data is depicted in Fig. 1, showcasing how these technologies work together to address data processing and analysis needs at the edge of the network.

Edge computing involves processing data near its source and origin, allowing for local analysis and processing rather than transmitting data over long distances to the cloud, which can result in latency. Utilizing edge

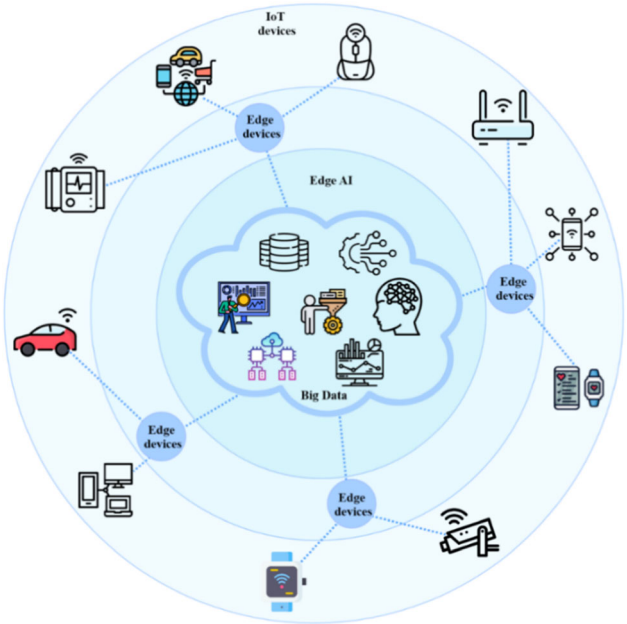

Fig. 1 Edge AI and big data

computing presents an efficient solution to mitigate latency issues, bringing computing and data storage closer to where it's needed. This approach conserves bandwidth and enhances response time. A notable example of edge devices is self-driving cars, where much processing occurs on the vehicle and edge side rather than solely relying on cloud-based processing. This enables sensitive devices requiring real-time decision-making to operate without delays, emphasizing the importance of edge computing.

In network communication, delay times are not guaranteed on the TCP platform. However, the advent of technologies like 5G network communication holds promise for reducing delays. Although 5G is still in its nascent stages and achieves remarkably low latencies, the telecommunications infrastructure must adapt to embrace the 5G platform fully. Despite these challenges, the integration of Edge architecture is crucial. By filtering and analyzing data at the edge, network bandwidth usage is minimized, thus reducing the risk of network overload. Moreover, apart from latency reduction, IoT's edge architecture enhances security and user experience. Like smartphone users who prioritize data privacy, ensuring that information and images remain secure and inaccessible to unauthorized parties is essential, a principle that extends to smart factories and IoT systems [1, 12].

ML and AI algorithms are pivotal in the operation of edge computing. ML aids edge devices in minimizing data storage requirements and predicting their behavior. AI algorithms anticipate future system inputs and allocate processing resources by analyzing input patterns and evaluating conditions. This optimization reduces

information processing time and enhances the server's response time, improving overall efficiency [2].

One application of Edge AI in the industry is predictive maintenance, where sensors' data are utilized to anticipate car breakdowns early on, preventing more significant risks and costly failures. However, data volumes have escalated dramatically with the proliferation of IoT devices. While sensitive IoT data are typically aggregated and stored in the cloud, the latency between data centers and end-users can render this arrangement unreliable. Organizations increasingly rely on processing data locally to address operational needs and ensure real-time decision-making. However, ensuring information security in cloud-based data remains a pressing concern for many businesses. Therefore, while faster processing may ideally occur in cloud data, local processing is essential to guarantee timely responses and effective decision-making at the edge [1, 4].

2.2 Related work

AI and edge computing research has explored a diverse array of topics. Yet, there remains a notable scarcity of comprehensive studies specifically focused on Edge AI about big data. This subsection provides a brief overview of select publications in the Edge AI domain.

Sanjay Misra [1] provided insightful details on current AI research, mainly focusing on cybersecurity challenges. The book aimed to analyze various tools and technologies, offering practical applications and discussions on their integration. It included quizzes to test readers' understanding and case studies demonstrating the transformative impact of AI, cloud computing, and edge computing on businesses. The book's strengths lie in its comprehensive coverage of emerging topics in AI, cloud computing, and edge computing, encompassing technologies such as ML, deep learning (DL), fog computing, and concepts related to cybersecurity and privacy. Furthermore, the focus on implementing and integrating different tools and technologies adds practical value to the discussions.

Saleh et al. [2] presented a literature review of recent ML-based and edge-computing mechanisms, particularly addressing challenges and potential future research directions in edge-computing servers. The authors offered clear explanations of mobile computation offloading (MCO) and its significance in mobile edge computing (MEC). Additionally, the paper reviewed ML-based and meta-learning-based MCO in MEC, highlighting issues, challenges, and future research directions. However, notable shortcomings include the field's lack of classification and taxonomy and inconsistent methodology usage, indicating areas for improvement in future research endeavors.

Chang et al. [3] comprehensively analyzed a coordinated end-edge-cloud architecture tailored for adaptable AIoT systems. The paper delves into fundamental concepts such as edge computing, AI, and IoT, providing a foundation for exploring the overarching architecture of AIoT. It offers a practical AIoT example to illustrate how AI can be effectively applied in real-world scenarios and an overview of promising AIoT applications. Furthermore, the study reviews the latest AI model inference and training advancements at the network's edge. It concludes by presenting challenges and potential directions in this captivating field. Including practical applications in the article, along with delineating open challenges and future research directions, is crucial for guiding further research and development efforts in this area.

Hua et al. [4] provided a formal definition of edge computing and elucidated the rationale behind its growing popularity as a computing model. They discussed the pertinent issues driving interest in edge computing, outlining traditional approaches and their limitations. The article serves as a roadmap for exploring novel edge computing and AI research concepts. A notable strength of the article is its reference to research results related to utilizing AI for edge computing optimization and its application across various domains within edge computing architecture. This highlights the practical insights and solutions offered by the article. However, a weakness lies in the lack of taxonomy and precise classification of the work, which may pose challenges for researchers seeking to navigate the literature in this area.

Zhou et al. [5] thoroughly analyzed current edge intelligence research projects. They commenced by reviewing the history and objectives of AI operations at the network edge, followed by an overview of broad architectures, frameworks, and cutting-edge key technologies for DL models aimed at training and inference at the network edge. Additionally, they discussed forthcoming edge intelligence research opportunities. The researchers adeptly contextualized their analysis by emphasizing recent advances in DL and the increasing significance of AI applications, particularly at the network's edge. However, their work exhibited some weaknesses, including a lack of specific details, such as concrete examples or case studies, and the absence of a taxonomy, which could have enhanced the clarity and organization of their research findings.

Deng et al. [6] categorized AI for Edge and AI on Edge. The former focused on utilizing popular and efficient AI technologies to offer more optimal solutions to significant edge computing challenges. At the same time, the latter explored the entire process of developing AI models. This paper provided comprehensive perspectives on this emerging interdisciplinary field, covering fundamental concepts and research strategies to provide future edge intelligence research initiatives with essential context.

However, a notable weakness of the work is the absence of a taxonomy, which could have contributed to a more transparent organization and classification of the research findings.

In Table 1, we present a summary of related work in the Edge AI domain.

3 Research methodology

This section introduces a systematic methodology for conducting a comprehensive and ethically sound investigation into Edge AI for big data. The systematic approach serves as a foundation for appropriately positioning novel research methodologies. These checks require more effort than standard reviews, and systematic evaluations may be crucial before undertaking a quantity-based superior analysis. Our search utilized reputable journals such as Springer, IEEE, and Science Direct. Section 4 outlines the process of formulating research questions, while Sect. 4.1 addresses the methodology for paper selection. The pivotal step in the systematic process is the meticulous review of the selected papers.

4 Research questions (RQs)

This paper aims to address five research questions (RQs) using the objectives and topic of the proposed study as a guide. The following questions are posed.

- RQ1: What are the use cases of Edge AI for big data?
- RQ2: How do big data drive Edge AI?
- RQ3: What evaluation factors usually apply to Edge AI for big data?
- RQ4: What measurement environments are used for evaluating the Edge AI for big data?
- RQ5: What are Edge AI's challenges and open issues for big data?

4.1 Procedure for selecting papers

The paper selection procedure involves the following steps:

- Stage 1: Keyword, abstract, and title-based automated search.
- Stage 2: Selection of papers based on the abstract and conclusion.
- Stage 3: Final selection based on the entire text.

In Stage 1, academic publishers such as IEEE, ACM, Wiley, Springer, and Science Direct are the basis for an electronic search using Google Scholar as the primary search engine. The search method chosen involved finding related studies by incorporating alternate spellings of the original elements. This process yielded 289 papers of various types relevant to the research topic.

- ("Edge computing") AND ("Artificial Intelligence" OR "AI") AND ("Big Data" OR "Data Analytics")
- ("Edge computing") AND ("Machine Learning" OR "ML") AND ("Big Data" OR "Data Analytics")
- ("Edge computing") AND ("Deep Learning" OR "DL") AND ("Big Data" OR "Data Analytics")
- ("Edge Artificial Intelligence" OR " Edge AI") AND ("Big Data" OR "Data Analytics")
- ("Task Offloading") AND ("Big Data" OR "Data Analytics")
- ("Edge computing") AND ("Deep Neural Network" OR "DNN") AND ("Big Data" OR "Data Analytics")

Table 1 Related work of Edge AI

Ref	Main topic	Review type	Type of publication	Year of Publication	Publisher	Covered Year	Taxonomy
Sanjay Misra [1]	AI for cloud and edge computing	–	Book	2022	Springer	Not limited-2022	Presented
Saleh et al. [2]	MEC using AI to offload mobile computation	Review	Journal	2022	Springer	2019–2021	Not presented
Chang et al. [3]	Recent developments in Edge-powered AI of Things	Survey	Journal	2021	IEEE	2008–2022	Presented
Hua et al. [4]	Edge computing and AI	Survey	Journal	2022	ACM	2004–2022	Not presented
Zhou et al. [5]	Edge computing and AI	Survey	Journal	2019	IEEE	2006–2019	Not presented
Deng et al. [6]	Edge Computing and AI Meet	Survey	Journal	2020	IEEE	2014–2019	Not presented

- ("Edge Devices") AND ("Big Data" OR "Data Analytics") AND ("Artificial Intelligence" OR "AI")
- ("Fog Computing") AND ("Big Data" OR "Data Analytics") AND ("Artificial Intelligence" OR "AI")

In Stage 2, utilizing our established criteria, we identified relevant papers published between 2018 and 2023, leading to the selection of 239 papers. Subsequently, in Stage 3, unsuitable papers were eliminated based on a full-text review, resulting in a final selection of 85 papers. The distribution of these 85 chosen research papers by publisher is illustrated in Fig. 2.

An overview of the method used in this study is shown in Fig. 3.

We have categorized Edge AI for big data into applications, requirements, and supporting technologies, presenting three conceptual taxonomies illustrated in Figs. 4, 5, and 6.

Figure 4 delineates the application category, encompassing smart cities, smart homes, the Internet of Vehicles (IoV), healthcare, and manufacturing. The second primary category, requirements, is depicted in Fig. 5 and includes orchestration, scalability, reliability, and security. The final primary group is supporting technology, illustrated in Fig. 6, encompassing hardware, software, data management, immersive technologies, and network components.

5 Comparison and categorization of Edge AI for big data

In this section, we conducted a comprehensive analysis and review of the 85 selected research articles within the domain of Edge AI for big data, as delineated in Sect. 3. Utilizing the taxonomies presented in Figs. 4, 5, and 6, we systematically categorized the examined research articles in the Edge AI for the big data field into three subsections: applications, requirements, and supporting technologies. This classification aims to facilitate a better understanding and address the five RQs mentioned in Sect. 3 in the next section. In other words, this section serves as the foundation for the upcoming section, where we present our findings from the analysis of these articles, preparing the groundwork for answering our RQs comprehensively.

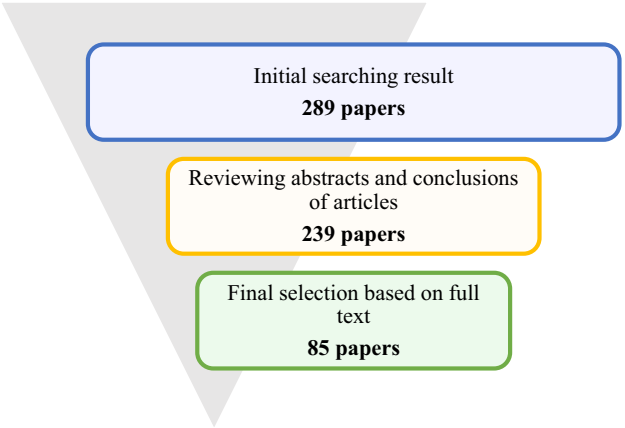

Fig. 3 Overview of the methodology

5.1 Applications

The inherently data-driven characteristics of IoT systems underscore the imperative for deploying Edge AI applications to facilitate on-site data processing. These applications enable data processing locally, ensure secure data transfer, and enhance privacy on Edge AI devices. Within this category of articles, we have classified them into five distinct subcategories: smart city, smart home, IoV, healthcare, and manufacturing, using the taxonomy proposed in Fig. 4 for Edge AI applications in big data contexts.

5.1.1 Smart city

Enabling real-time data processing at the network's edge, Edge AI significantly enhances the functionality of smart cities. AI-powered edge devices deployed in smart city infrastructures can efficiently monitor and regulate traffic

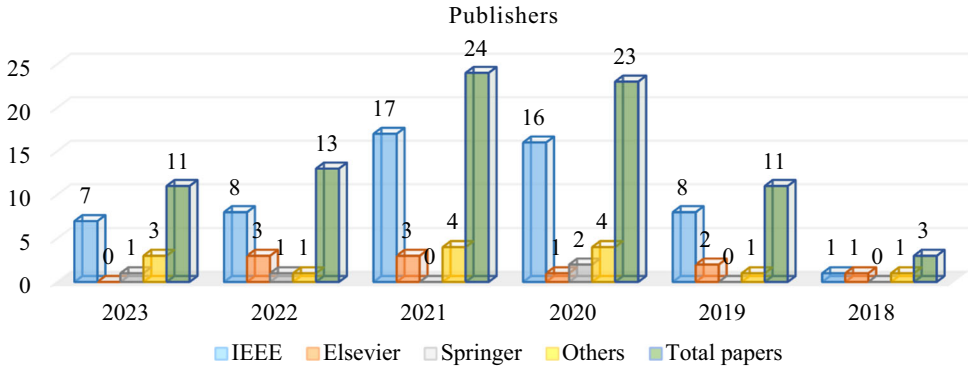

Fig. 2 Distribution of research papers by publisher

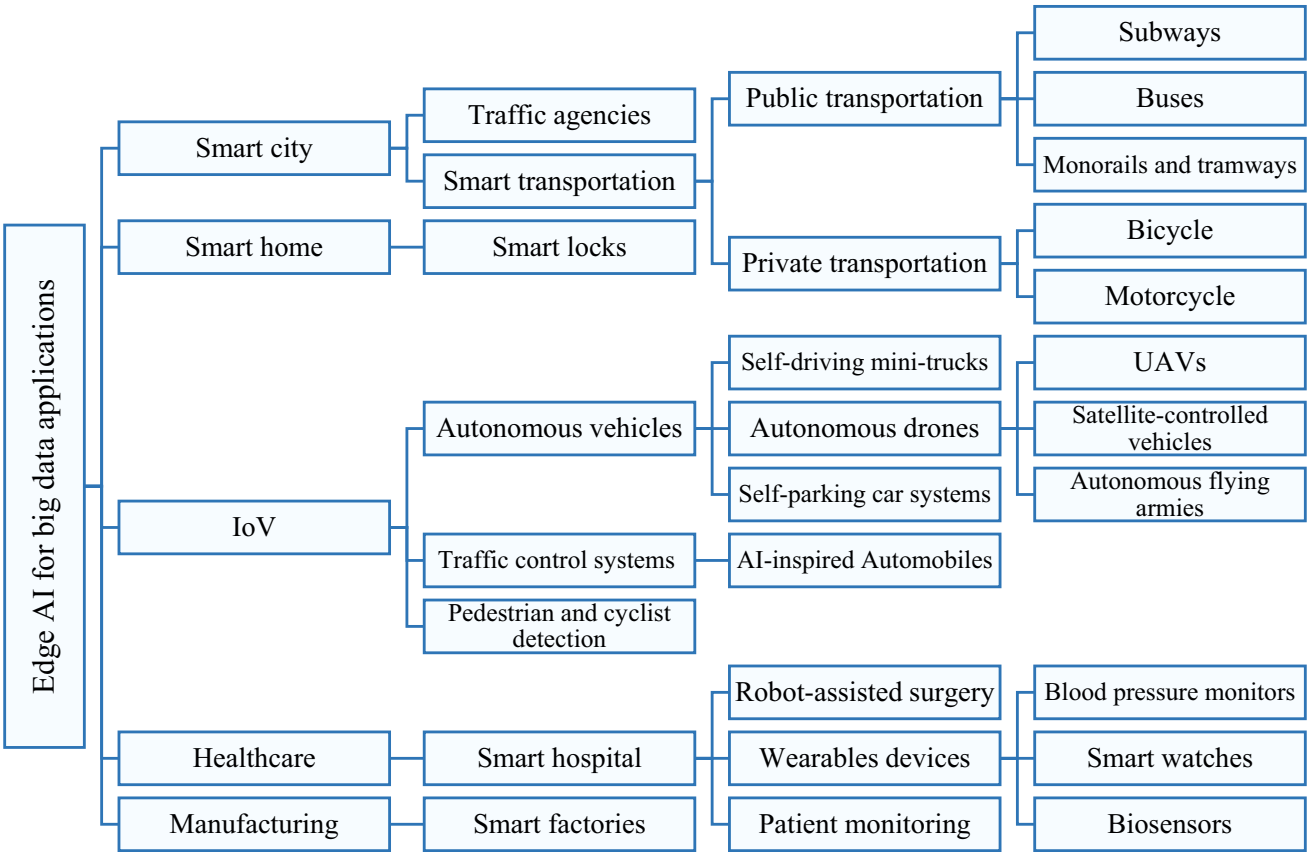

Fig. 4 Edge AI for big data applications taxonomy

flow, optimize energy consumption, improve public safety through surveillance, and proactively mitigate environmental disasters. The following compilation presents the articles within this category that underwent our review.

Lv et al. [13] introduced a collaborative calculation method. They described the alternate direction method of multipliers (ADMM), MEC theories, and methodologies combined with a subordinate game theory based on the Stackelberg principle. This approach ensures network stability in expansive settings.

Labba et al. [14] devised an Edge AI-based strategy for remote learning, incorporating a federated ML model for real-time student failure prediction and a generic operational architecture for an AI unit at the edge. They presented a real-world K–12 student scenario in online education to underscore the strategy's effectiveness, revolutionizing online education by providing students with real-time assessments while preserving their privacy.

Zhang et al. [15] proposed an edge-to-edge (E2E) cooperation AI technique involving training offloading upon request. They formally defined learning resource movement using learning capacity and its offloading and explored E2E and cloud-edge collaboration for mutual learning offloading. Additionally, they devised an E2E learning offloading allocation model based on different deep neural network (DNN) subtasks and their diverse learning resource requirements.

Rahman et al. [16] delineated several crucial aspects of Industrial Internet of Things (IIoT) data combined with AI for monitoring smart cities. They introduced an AI-enabled IIoT framework and a crowdsourcing application to harness human intelligence for real-time event and object collection from IIoT data. Moreover, the authors integrated multiple AI algorithms to promptly generate analytics, reports, and alerts from IIoT data. This process involved automated classification of collected events and objects, which were then executed on distributed edge and cloud nodes.

Bui et al. [17] devised a computational negotiation method tailored for IoT networks, empowering dispersed edge devices to make decisions regarding smart application utilization autonomously. They emphasized the significance of edge analytics in autonomously augmenting IoT system efficiency. Additionally, they provided two sample IoT application use cases within the smart city framework to evaluate the effectiveness of their proposed methodology.

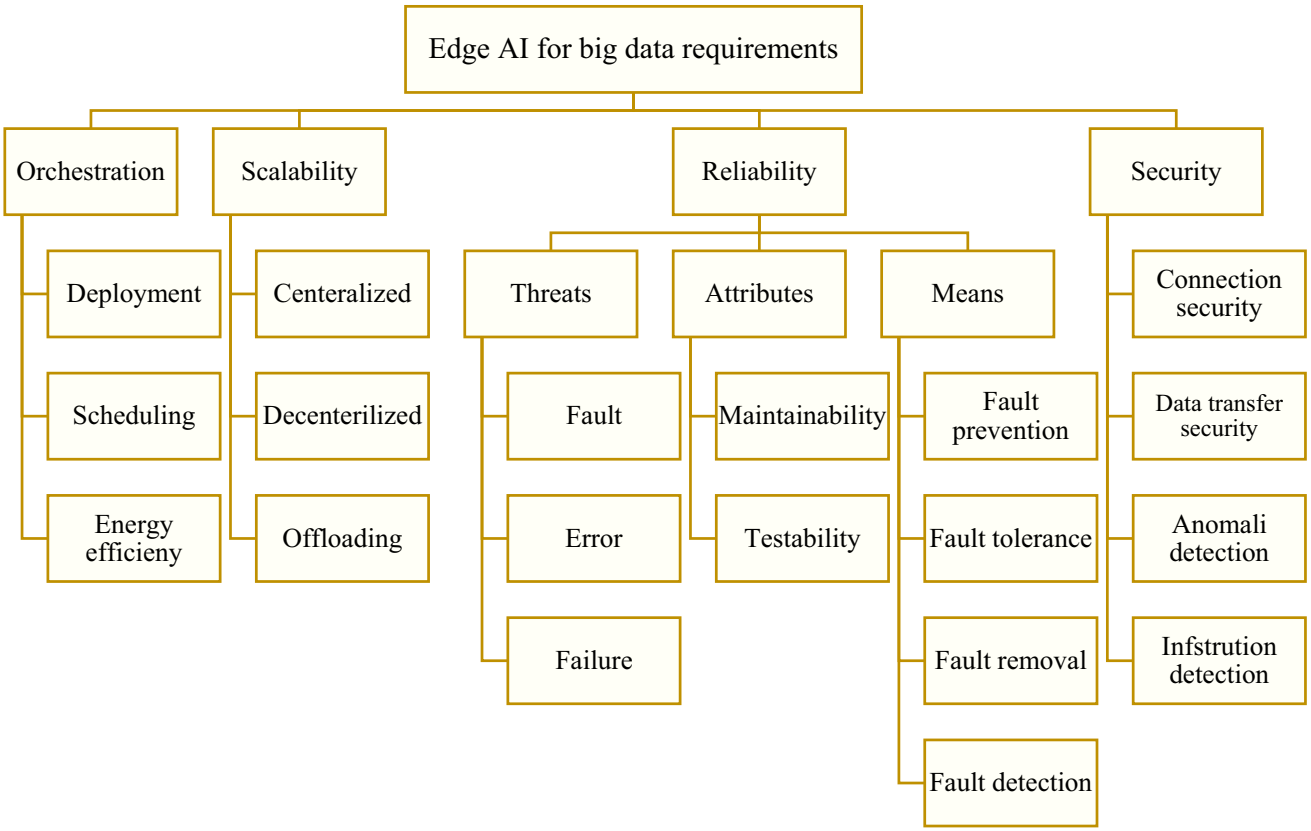

Fig. 5 Edge AI for big data requirements taxonomy

Rocha Neto et al. [18] addressed challenges arising from the extensive deployment of monitoring cameras in smart city environments, which strain network and processing infrastructure. The authors proposed a distributed system for video analytics that capitalizes on edge computing capabilities. Simulations using the YAFS simulator showcased the efficacy of the proposed algorithm in distributing the workload.

Chen et al. [19] aimed to bolster the resilience of IoT systems in smart cities, especially within the realm of 5G and beyond 5G networks. They developed a distributed learning framework utilizing edge intelligence to bolster the networking prowess of smart terminal nodes. Results showcased that this strategy significantly curtailed training time and costs by harnessing the power of a multicore CPU in conjunction with edge intelligence.

5.1.2 Smart home

In the age of Edge AI, contemporary residences are transitioning into smart homes. Edge computing empowers devices to process data locally, reducing latency and enhancing privacy. This capability enables rapid responses to tasks such as temperature adjustments and facial recognition at the door, resulting in a safer and more comfortable living environment. The following articles from this category have undergone analysis and are listed below.

Nasir et al. [20] leveraged Edge AI-enabled technology to propose a fully functional smart home system based on IoT and edge computing. The suggested solution employed edge devices as a computing platform to conserve energy, enhance security, and facilitate remote management of all appliances through a secure gateway. The authors employed a proprietary lightweight DNN architecture for a human fall detection case study validated using the Le2i dataset.

Lin [21] developed an innovative smart home system architecture with edge analytics capabilities. Addressing energy efficiency concerns in smart homes, the authors presented a case study demonstrating the automatic identification of tracked electrical appliances for smart home energy management. This study tested the viability and efficacy of the new smart home system architecture complemented by locally distributed and integrated adaptable edge-sensing devices.

To address challenges arising from non-independent and identically distributed (non-IID) data and disparate computing capabilities in Federated Learning (FL) within smart homes, Li et al. [22] proposed an innovative cluster FL

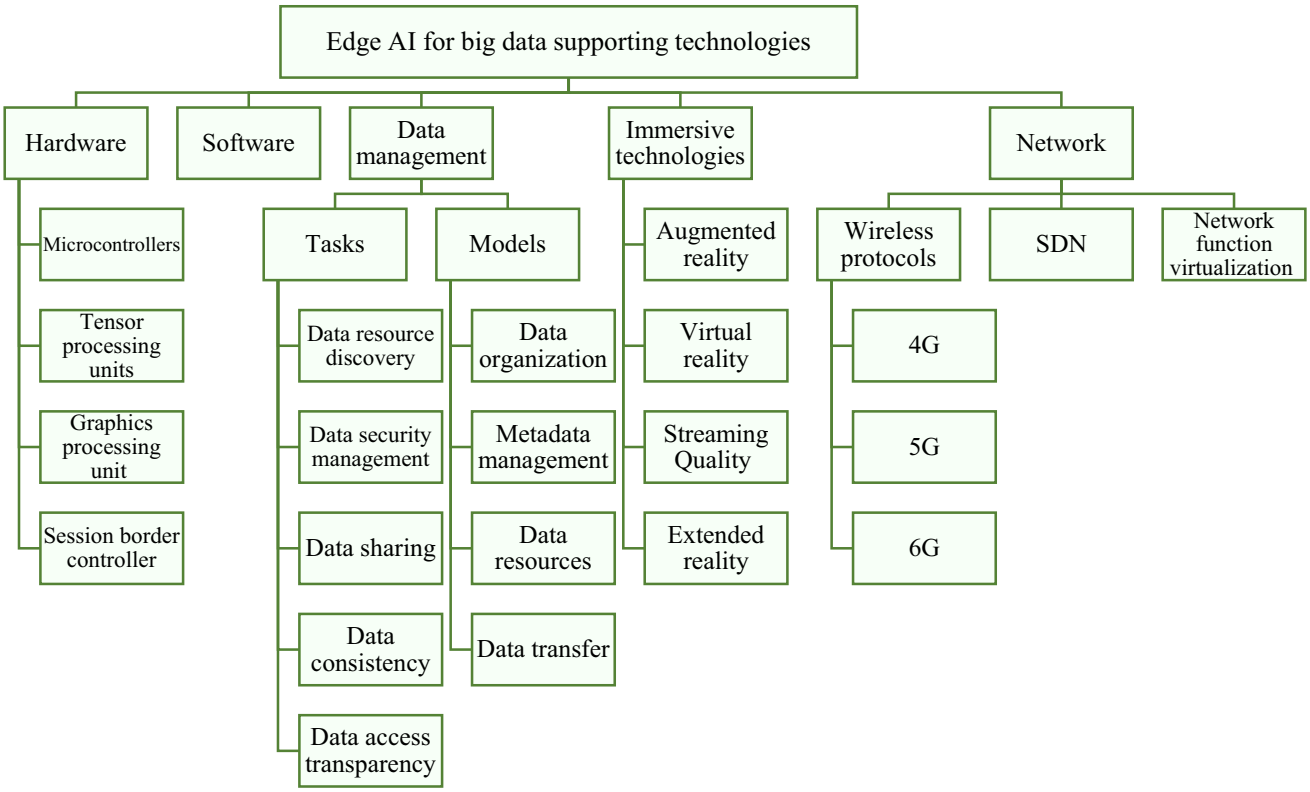

Fig. 6 Edge AI for big data supporting technologies taxonomy

architecture that capitalizes on edge-cloud collaboration. Their approach utilizes a Gaussian mixture model-based clustering technique within FL, aiming to refine model accuracy by grouping features in the FL dataset. Additionally, they introduced a model training strategy grounded in edge-cloud collaboration, showcasing enhancements in the accuracy of global models while preserving regular network service provision.

Zhang et al. [23] introduced an edge-based solution to tackle the activity recognition challenge in smart homes. They presented an edge-computing architecture to maintain consistency and scalability in executing tasks on edge devices. The researchers developed a convolutional neural network (CNN) model tailored explicitly for edge-device activity recognition tasks. Through experiments, they demonstrated promising performance outcomes compared to existing ML methods.

Benadda et al. [24] focused on the hardware design and integration of smart home automation systems, particularly concerning power consumption. They proposed an integrated approach that combines IoT, AI, and DL analysis to facilitate intelligent electricity usage. This approach leverages power grid measurements and user facility interactions. The results highlighted the feasibility and effectiveness of the proposed hardware design and integration in optimizing power consumption within smart home environments.

Chen et al. [25] developed a smart home energy management prototype, utilizing fog-cloud computing and featuring non-intrusive appliance load monitoring (NIALM) as an IoT application. This prototype employed an artificial neural network (ANN)-based NIALM approach to monitor electrical appliances without requiring intrusive plug-load power meters. The authors successfully demonstrated the feasibility and usability of the prototype.

5.1.3 IoV

The development of the IoV has significantly progressed with the integration of Edge AI. Current applications involve incorporating AI algorithms into infrastructure, vehicles, and mobile components, facilitating autonomous driving, real-time traffic management, and predictive vehicle maintenance. This integration empowers vehicles to make rapid decisions by analyzing data at the edge, enhancing road safety and efficiency and establishing a foundation for the future of transportation. Below is a compilation of articles within this category that underwent examination.

Su et al. [26] introduced an advanced driving assistance system (ADAS) for vehicle platforms, leveraging YOLO

v3-tiny's neural network and edge computing to detect pedestrians and obstacles. Their design tackled accuracy concerns by incorporating the SqueezeNet approach with quantization, effectively reducing operations while preserving accuracy. However, the model's weakness is reducing input size and network layer layers. Nevertheless, the research significantly enhances accuracy and detection rates in ADAS applications.

Sonmez et al. [27] introduced a two-stage ML-based vehicular edge orchestrator, focusing on service time and task completion success. They utilized EdgeCloudSim for realistic simulations, showcasing the model's performance compared to competing schemes. The research excels in demonstrating strong performance, although the ML-based workload orchestrator could benefit from enhancements in feature utilization for classification and regression stages.

Liu et al. [28] addressed ultralow network response latency in IoV applications by proposing a digital twin (DT) supported edge intelligent cooperation scheme. They formulated mathematical expressions for network response time and introduced a mathematical optimization model for latency minimization, demonstrating the superiority of the proposed algorithm in simulations.

Jiang et al. [29] proposed an Edge AI framework for object detection in the Internet of Vehicles (IoV) system, introducing an abductive learning algorithm for interpretability and robustness. The algorithm demonstrated high simulation accuracy by leveraging the YOLO classifier and blockchain technology for model sharing. This approach balances efficiency and computational complexity, making it a promising solution for object detection in IoV systems.

Alladi et al. [30] suggested an AI-based intrusion detection architecture for Internet of Vehicles (IoV) networks, deploying DL engines (DLEs) on multiaccess edge computing (MEC) servers. Experimental results on a MEC testbed highlighted the scheme's effectiveness in identifying and classifying vehicular traffic cyberattacks. This research capitalizes on AI and edge computing to enhance intrusion detection capabilities in the dynamic IoV environment, offering promising advancements in network security.

5.1.4 Healthcare

Edge AI is transforming patient care and diagnostics in healthcare. Portable medical devices enhanced by Edge AI can analyze patient data at the point of service, offering swift and precise diagnoses. This technology empowers healthcare professionals to monitor patients and promptly respond remotely when needed. Below is a compilation of articles within this category that have undergone our review.

Aazam et al. [31] delved into utilizing ML-based edge computing for smart and opportunistic healthcare services. By considering various healthcare and safety scenarios, they applied k-nearest neighbors (KNN), naïve Bayes (NB), and support vector classification (SVC) algorithms on real data traces. The empirical findings shed light on the efficacy of edge computing work offloading based on ML, with the paper standing out for its effectiveness in the proposed model.

Hayyolalam et al. [32] underscored the advantages of intelligent edge technologies and AI in healthcare systems. They introduced a groundbreaking smart healthcare paradigm to enhance the integration of AI and edge technologies. The paper also discusses potential challenges and future research prospects arising from the convergence of these technologies.

Li et al. [33] devised a secure framework for SDN-based Edge computing in an IoT-enabled healthcare system. Their structure incorporates a lightweight authentication technique for IoT devices, ensuring secure data gathering, processing, and analysis on Edge servers. Computer-based simulations validated the framework's efficacy, offering superior solutions for IoT-enabled healthcare systems.

Guo et al. [34] introduced the FEEL (FEderated Edge Learning) system for privacy-preserving mobile healthcare. Their approach includes an edge-based training task offloading strategy to enhance training efficiency and utilizes FL for improved inference performance. The paper introduces a differential privacy approach to bolster privacy protection during model training, demonstrating efficient and privacy-preserving model training.

Mansour et al. [35] presented an AI and IoT convergence-based illness diagnostic paradigm for smart healthcare systems, focusing on heart disease and diabetes. Their model encompasses data collection, preprocessing, classification, and parameter adjustment. The Crow Search Optimization (CSO) algorithm-based Cascaded Long Short-Term Memory (CSO-CLSTM) model significantly enhances diagnostic accuracy.

Nguyen et al. [36] tackled security challenges in healthcare data by proposing a decentralized approach with blockchain technology. They integrated blockchain at the network's edge in an intelligent homecare system to enhance healthcare quality and overcome security limitations. Their approach emphasizes blockchain's immutability and its potential to improve healthcare services.

Yadav et al. [37] introduced a Computation Offloading using Reinforcement Learning (CORL) scheme to minimize latency and energy consumption in edge-enabled sensor networks. The algorithm optimally offloads tasks, balancing energy and latency, with experimental results

demonstrating energy savings, latency minimization, and efficient node resource utilization.

Rangarajan et al. [38] addressed rapid COVID-19 diagnosis by integrating AI with smartphone chest X-ray images. They utilized five pre-trained CNN models and augmentation techniques to increase the dataset using generative adversarial network (GAN) techniques. The best-performing models, VGG16 and Xception, trained with synthetic images, outperformed augmented images when deployed on smartphones, highlighting their effectiveness.

5.1.5 Manufacturing

Edge AI is revolutionizing manufacturing by streamlining processes and elevating quality assurance. In intelligent factories, Edge AI-powered devices monitor equipment health, detect flaws in real time, and predict maintenance needs, reducing downtime and enhancing productivity. Below is a compilation of items in this category that underwent examination.

Liao et al. [39] tackled channel selection optimization, which is crucial for timely task completion. Through ML, Lyapunov optimization, and matching theory, they developed a learning-based selection scheme demonstrating assured effectiveness with restricted deviation from ideal effectiveness, leading to improved energy consumption. However, limited computing and battery capacity pose drawbacks.

Ren et al. [40] focused on developing a big data platform for an intelligence-based industrial IoT sensor monitoring system, leveraging edge computing and AI. The architecture effectively monitored conditions in a production facility, especially in monitoring a production line's concrete-making state. However, security weaknesses prompt future research attention.

Foukalas et al. [41] proposed a federated active transfer learning (FATL) architecture to address Edge AI challenges through training and testing. The FATL model demonstrated increased accuracy, tackling unresolved Edge AI difficulties for future IIoT applications and offering insightful information regarding the suggested FATL model.

Hu et al. [42] contributed insights into intelligent manufacturing and integrating AI and IoT in factory settings. Their iRobot-Factory implementation showcased significant improvements in chip assembly and production efficiency, accompanied by decreased system instructions.

Ghahramani et al. [43] focused on semiconductor manufacturing, utilizing evolutionary computing and neural network algorithms for comprehensive analysis. Their intelligent feature selection algorithm contributes to smart manufacturing advancements in semiconductor production.

Lin et al. [44] proposed a smart manufacturing factory framework based on edge computing, exploring the implications of job shop scheduling problems (JSP). Utilizing deep Q network (DQN), their research demonstrated the effectiveness of integrating edge computing with DQN in addressing JSP in smart manufacturing.

Ammayappan [45] conducted AI-based research for a smart framework enabling dynamic system optimization and decreased communication load. While exhibiting improved adaptability, optimization of computational task learning remains a challenge, urging future research to concentrate on collaborating with mobile devices and edge nodes for optimization.

Hao et al. [46] proposed the Smart-Edge-CoCaCo algorithm to minimize total delay and optimize computation offloading decisions. Experimental results demonstrated lower computation delay than traditional cloud computing models, especially with increased computing task data and concurrent users.

Table 2 provides an overview of the diverse application areas and use cases of Edge AI. It highlights how Edge AI is applied across various domains, including healthcare, manufacturing, transportation, and smart homes. Each application area presents unique use cases where Edge AI technologies are leveraged to improve efficiency, enhance decision-making processes, and enable real-time data analysis at the network's edge.

Table 3 offers insights into the development phases, tools/methods, and datasets utilized in the Edge AI application category. It provides a comprehensive breakdown of the methodologies and resources employed during developing Edge AI solutions, shedding light on the processes involved in building and implementing such systems.

Table 4 summarizes the strong points, limitations, and proposed solutions in the Edge AI application category. This table encapsulates the essential findings and observations from the reviewed literature, highlighting the strengths and weaknesses of Edge AI solutions.

5.2 Requirements

With Edge AI, ML algorithms can be executed directly at the edge, enabling data analysis within IoT devices rather than relying solely on centralized cloud computing facilities or private data centers. Utilizing our proposed taxonomy of Edge AI for big data requirements, as illustrated in Fig. 5, we have categorized articles in this group into four subcategories: orchestration, scalability, reliability, and security. A detailed analysis of articles within this category is provided below, aiming to offer insights into the various aspects of Edge AI implementation and its impact on addressing big data requirements.

Table 2 Application areas and use cases of Edge AI

Application area	Scenario/use case
Smart city	AI-enhanced infrastructure optimization [13–18]
	5G technology and beyond 5G[19]
Smart home	IoT and edge computing in smart homes[20]
	Edge analytics for home energy management[21]
	Non-IID data with edge-cloud collaboration[22, 25]
	Edge-based solution for activity recognition[23]
	Hardware design for smart home automation[24]
IoV	Pedestrian and knight identification[26]
	Service time and task completion[27]
	Edge intelligent cooperation[28]
	Abductive learning[29]
	Vehicular traffic classification[30]
Healthcare	Smart and opportunistic healthcare[31, 32]
	Patient data processing[33, 36, 38]
	Edge-based training[34]
	Heart disease and diabetes diagnosis[35]
	Latency and energy cost minimization[37]
Manufacturing	Edge server resource allocation[39]
	Condition monitoring for vibration[40]
	Edge AI training[41, 43–45]
	iRobot-factory, cognitive manufacturing[42]
	Computation offloading decisions[46]

Table 3 Development phases, tools/methods, and datasets utilized in the Edge AI application category

Development phase	Used tool or method	Used dataset
Implementation	Stackelberg game[13]	Le2i dataset[20]
	A real-world scenario of K-12 learners and FL[14]	PAMAP2 dataset[31]
	AI techniques(DNN FL, CNN, DL, ANN, KNN, NB, SVC, LSTM) [16, 17, 20–25, 30–32, 34, 35, 40, 43]	Generated a dataset [33]
		Heart disease and diabetes datasets [35]
	Distributed learning[19]	
	Squeeze Net and Quantization method[26]	
	YOLO classifier[29]	
	MATLAB [33]	
	Blockchain[36]	
Simulation	DNN method[15]	–
	YAFS simulator[18]	
	EdgeCloudSim [27]	
	AI techniques[41, 44–46]	
Both simulation and implementation	Digital Twin, Queuing Model[28]	VGG16, MobileNetV2, Xception, NASNetMobile, InceptionResNetV2 datasets[38]
	iFogSim simulator[37]	
	Matching theory and Lyapunov optimization[39]	
	AI techniques [37, 38, 42]	

Table 4 Strong points, limitations, and solutions in Edge AI application category

Ref	Strong point	Limitation and solutions
[13]	Good network stability in a large-scale setting	For less flexibility, explore applications in future big data, IoT, and AI research
[14]	Improving online education	Explore enhancements and scalability in future implementations
[15]	Enhancing learning efficiency and resource utilization	Provide real-world implementation challenges
[16]	Capturing and processing IoT data in real time, Better event management and decision-making	Address latency and accuracy challenges in distributed processing
[17]	Enabling efficient decision-making in IoT systems	Address privacy concerns
[18]	Real-time responses	Address scalability issues
[19]	Robust networking capability, Decreased training time and cost	Limited to multicore CPU for edge intelligence
[20]	Reducing energy costs, Improving security in smart homes, Achieving high accuracy	The adaptability of the proposed system should improve to different types of smart home scenarios
[21]	Providing real-time responsiveness and local actionable, Improving energy consumption, Enhancing the monitoring and control of electrical home appliances	The need for compatible IoT devices
[22]	Improved model accuracy and maintained regular network service	Optimization for various edge and cloud configurations
[23]	Low-cost hardware, Better consistency, Better scalability	Generalizability to diverse home environments needs validation
[24]	Intelligent electricity usage	Explore additional use cases beyond monitored settings needed
[25]	Non-intrusive appliance monitoring	Limited to specific NIALM approaches, so exploring alternative NIALM approaches is needed
[26]	Improved accuracy, Improved detection rate	Explore optimization for input size and assess performance with varying network layers
[27]	Good performance	The ML-based workload orchestrator must improve the traits utilized in the classification and regression stages
[28]	Superiority in network response latency	The complexity of the Markov decision process should address
[29]	High accuracy, Enhanced interpretability	Explore computational complexity optimizations
[30]	Efficient handling of cybersecurity challenges in IoV networks	Explore optimizations for diverse cyberattack types
[31]	Effectiveness of the proposed model	Scalability in diverse healthcare scenarios needs validation
[32]	Reducing latency and energy consumption	Security and privacy concerns in healthcare data need to address
[33]	Enhancing security, Improving communication speed, Reducing latency, Optimizing resource utilization	Improving the suggested framework by preserving patients' and their data's privacy
[34]	Improving model inference performance and enhancing privacy protection during model training	The need for real-world deployment and validation
[35]	Achieving high accuracy in diagnosing, Enhancing the performance	Employing feature selection approaches that alleviate the curse of dimensionality
[36]	Overcoming security limitations, Enhancing healthcare services	Explore additional blockchain optimizations
[37]	Energy savings, Latency minimization, Efficient resource utilization	Specific latency and energy metrics need to be detailed
[38]	Swift diagnosis of COVID-19 cases, Improved performance with synthetic images	Optimize dataset augmentation techniques
[39]	Better energy consumption	concentrate on resource allocation in the edge server under information uncertainty
[40]	Effectiveness in monitoring the state of a production line for making concrete	The security of platforms for the industrial IoT based on edge computing and AI should receive attention
[41]	Increasing accuracy	Security concerns should be addressed
[42]	Significant improvements in production efficiency	Enhance dataset details and Explore additional efficiency metrics
[43]	Optimizing semiconductor manufacturing processes	Explore additional optimization strategies
[44]	Superiority in addressing Job Shop Scheduling	Validate in real-world smart manufacturing scenarios

Table 4 (continued)

Ref	Strong point	Limitation and solutions
[45]	Improving the system's capacity for adaptability	concentrate on collaborating with mobile devices and edge nodes for computational task-learning optimization
[46]	Minimized total delay, Optimized computation offloading	Define specific metrics for delay optimization

5.2.1 Orchestration

In the context of Edge AI, orchestration involves effectively managing and coordinating numerous edge devices and their AI workloads. Orchestration technologies optimally disperse AI workloads across the edge network, balancing computational loads and reducing latency. This seamless cooperation across components enables the smooth operation of complex applications such as driverless vehicles or industrial automation. The following list includes papers from this category that have been examined. Table 7 details each article's challenge, evaluation metric, development phase, advantage, limitation, and solutions.

Zhang et al. [47] devised a heuristic technique to optimize workload allocation and virtual machine processing power to minimize energy and delay costs while preserving output quality. They introduced a versatile service model tailored to address various computing difficulties, demonstrating the Pareto optimality of the optimization model. Their approach significantly reduced energy consumption and delay expenses, enhancing overall efficiency. However, refinement is needed to extend the method's applicability beyond Edge AI scenarios.

Zhu et al. [48] conducted a detailed analysis of energy utilization patterns in intelligent edge devices and cloud services within an intelligent edge computing testbed. Their findings revealed notable improvements, particularly in reduced energy usage, as demonstrated through extensive simulations. This strategy offers the distinct advantage of promoting lower energy consumption, contributing to sustainable and efficient edge computing environments.

Dong et al. [49] proposed a learning-based decision-making paradigm for efficient energy dispatch in an islanding microgrid, leveraging a cloud-edge computing architecture. While effective, their algorithmic approach requires the development of more sophisticated and interpretable ML models. Nevertheless, their solution addresses energy management challenges within microgrid settings, promising operational efficiency and resource utilization improvements.

Zeb et al. [50] explored the significance of AI in next-generation networks (NGNs), emphasizing its role in achieving zero-touch service management and self-optimizing networks. They highlighted the convergence of AI, network softwarization, and edge-native computing architecture, showcasing the potential for enhanced service-oriented architecture. Notably, their use case study demonstrated the accuracy of predictions using a DL-based forecaster model in a multisite cloud/edge-native NGN testbed, highlighting the transformative potential of edge intelligence frameworks.

5.2.2 Scalability

Scalability, denoting the capacity of edge computing systems to accommodate a growing number of devices and data as demands escalate, stands as a critical aspect of Edge AI. It ensures that edge networks can adapt to shifting needs without compromising performance. This is particularly crucial for applications like Smart Cities and IoT, where the number of connected devices may vary considerably. The papers investigated within this category are outlined below.

Fragkos et al. [51] integrated game theory and reinforcement learning (RL) concepts to optimize the data offloading process of UAVs in a multiserver MEC environment. Utilizing stochastic learning automata theory, UAVs autonomously select MEC servers for partial data offloading, with simulations evaluating the framework's performance across diverse operational scenarios. The research demonstrates efficacy in different operating strategies and scenarios, though future expansion based on contract theory concepts is warranted.

Debauche et al. [52] proposed scalable poultry monitoring utilizing open hardware, a wireless sensor network, and software powered by a Gated Recurrent Unit AI algorithm. Their approach showcases good performance in validating and predicting environmental parameters. They plan to integrate video and animal chicken analysis in future iterations to enhance anomaly detection in poultry further.

Chen et al. [53] tackled challenges in low-latency object identification, categorization, computer efficiency, and augmented reality (AR) applications by introducing a practical mathematical model of augmented reality.

Combining decentralized MEC with federated learning (FL), their architecture significantly reduces execution latency, as evidenced by numerical results utilizing the CIFAR-10 dataset. Notably, their framework requires fewer training iterations than centralized learning, thus streamlining the training process.

Guo et al. [54] explored a distributed ML strategy for a multiuser MEC network in a cognitive eavesdropping setting, employing an FL framework to optimize offloading, bandwidth, and computational capability allocation ratios. Simulation results indicate reduced system latency and energy consumption costs, enhanced bandwidth, and computational power allocation for users with higher job priorities.

Paissan et al. [55] introduced PhiNets, a scalable backbone optimized for deep-learning-based image processing on resource-constrained IoT platforms. Leveraging inverted residual blocks, PhiNets efficiently distributes computational cost, working memory, and parameter memory, achieving state-of-the-art results in detection and tracking tasks while significantly reducing parameter count compared to previous models.

Fleischer et al. [56] unveiled a multi-TOPS (tera operations per second) AI core designed to accelerate DL training and inference across diverse systems, from edge devices to data centers. The AI core's programmable architecture with a custom instruction set architecture (ISA) achieves high sustained utilization for various neural network topologies, demonstrating enhanced compute precision for training and inference accuracy through a dataflow architecture and on-chip scratchpad hierarchy optimization.

5.2.3 Reliability

Ensuring the reliability of Edge AI applications is paramount to guarantee utility and safety. From remote sensors in challenging environments to autonomous drones navigating unpredictable weather conditions, edge devices must consistently perform across diverse scenarios. Achieving minimal downtime and ensuring uninterrupted operation necessitates incorporating redundant systems, failover procedures, and robust hardware within Edge AI reliability measures. The following is a compilation of papers from this category that have undergone examination.

Sankar et al. [9] employed ML techniques to extract optimal features from the training evaluation dataset. These selected features were fed into the CNN and other fully connected layers for further processing. The proposed method underwent evaluation using three datasets—polarity, Rotten Tomatoes, and IMDb—based on standard assessment criteria such as precision, accuracy, recall, and f-measure. Notably, this study utilized a pre-trained sentiment analysis model within an Android application framework to categorize reviews on a smartphone, omitting the use of cloud or server-side APIs.

Zhang et al. [57] introduced the distributed storage and calculation-based KNN (D-KNN) method for distributed storage and processing. Various tests were conducted to validate the efficacy of the D-KNN algorithm, with experimental results demonstrating a significant increase in the operational efficiency of KNN search. The algorithm can be swiftly and adaptably implemented in a cloud-edge computing environment to handle large datasets in Cyber-Physical Systems of Security (CPSS). This paper's advantages lie in enhancing efficiency and flexibility while reducing time complexity.

Liu et al. [58]proposed MEC for offloading computation tasks, emphasizing improved service reliability. Unlike previous approaches, they showed that AI-powered time-critical services can tolerate minor image distortions while maintaining high inference accuracy. Evaluating User Datagram Protocol (UDP)-based offloading in MEC, the results demonstrated improved normalized service reliability compared to Transmission Control Protocol (TCP)-based offloading. An Early Termination of the Image Reception (ETR) offloading scheme enhanced reliability.

Wu et al. [59] introduced HybridFL, a multilayer FL protocol for MEC. Addressing unreliable end devices, the protocol employs regional slack factors in client selection, mitigating impact without explicitly identifying states. Experimental results showed significant improvements in FL training—shortened round lengths, accelerated global model, and reduced end device energy consumption.

Qiu et al. [60] tackled reliable AI Service Chains (AISCs) in the edge intelligence cloud (EIC). Optimizing VNF and backup VNF deployment, secure link selection, and throughput maximization, they presented NP-hard integer linear programming. Two online algorithms with competitive ratios address on-site and off-site scenarios, with theoretical analyses and experiments demonstrating effectiveness.

Mutalemwa et al. [61] provided an overview of ultra-reliable low-latency communications (URLLC) in 5G and beyond. Focusing on AI-enabled edge computing and caching solutions, the article categorizes techniques, discusses mechanisms, and analyzes state-of-the-art edge caching schemes. It also highlights the advantages of FL frameworks and discusses IEEE 802.1 time-sensitive networking and emerging IETF deterministic networking standards. The conclusion presents open issues and research opportunities.

5.2.4 Security

Due to the dispersion of edge networks, security is significant in Edge AI. Threats to cybersecurity and physical security frequently affect edge devices. Therefore, strong security measures are necessary to protect data, AI models, and the integrity of the entire edge ecosystem. Only a few elements of a thorough security plan in Edge AI include encryption, access limits, and frequent security updates. Top attention is also given to ensuring data privacy and confidentiality, particularly in smart homes and healthcare applications that include sensitive data. Below is a list of the papers from this category that have been examined.

Zhang et al. [62] developed an approach for splitting models and integrating differential privacy in the FedMEC FL framework for MEC. By dividing a DNN model into two sections, FedMEC allows the edge servers to tackle the trickiest calculations. Additionally, they used the differentially private data perturbation approach to reduce privacy issues caused by local design variables in situations when updates from an edge device to the edge server are impacted by Laplace noise. The outcomes illustrated the usefulness and efficacy of their FedMEC plan. In their upcoming study, they want to examine the optimal perturbation intensity for the selectively private data perturbation method.

Shahbazi et al. [63] created a system using an integration of edge computing, blockchain technology, and ML to assist in the design of the manufacturing system. Based on the optimization model, the system's assignment problem was developed. The uses of blockchain technology may be measured and further investigated. Other technologies can be included To advance the development of the production system. The benefit of this research is speeding up the processing. To mention weakness, we can say the proposed system should be analyzed more thoroughly. The proposed system should be studied in more detail.

Makkar et al. [64] suggested the applicability of Internet attack detection by FML. Their main contributions include using FL to fulfill client search requests by identifying dangerous spam pictures that might cause these AI systems to get useless data. According to their testing findings, FML is appropriate in real-world situations where fluctuating picture sizes and animation ratios to authentic samples of images featured in adverts may divert customers from obtaining pertinent results. With the assessed outcomes, the cutting-edge FedLearnSP demonstrated considerable picture spam detection.

Zhou et al. [10] created three cutting-edge defensive tactics to repel damaging assaults in three ways. To develop a trustworthy continuous upgrading of AI, they initially introduced a cloud-edge collaborative anti-attack technique by assuring the security of the data generated during the training phase. They also suggested an edge-enhanced protection method based on evolutionary track and trace and punishment techniques to solve the security issue quickly and efficiently at the inference stage of the AI model. Low cost, high confidentiality, and availability are the benefits of this paper.

Kozik et al. [65] suggested a distributed Extreme Learning Machine (ELM)-based attack detection method that uses HPC cluster capabilities for labor- and cost-intensive classifier training. The suggested method uses specialized edge nodes close to the target data sources to gather aggregated traffic using NetFlow and classify and analyze it live. By conducting several experiments on a dataset of actual attacks, the researchers have also demonstrated the effectiveness of the suggested detection scheme. Better flexibility is its advantage.

Manoharan et al. [66] presented a unique system with poisoning attacks on the ML training dataset, incorporating real data pieces that lower the classifier's training accuracy. The proposed system contains three parts: a target classifier, a discriminator, and a generative adversarial networks (GAN) generator. The proposed model successfully undermines the classifiers, which is the benefit of this paper.

Table 5 in the Edge AI requirements category outlines various challenges in implementing Edge AI solutions. These challenges may include scalability, reliability, security, and efficiency issues. By identifying and categorizing these challenges, researchers and practitioners gain insights into the key areas that need attention and further development in Edge AI.

Table 6 analyzes the strong points, limitations, and proposed solutions within the Edge AI requirements category. This table provides a comprehensive overview of the advantages offered by different Edge AI approaches and the potential drawbacks and strategies to mitigate them.

Table 7 in the Edge AI requirements category details the development phases, tools, methods, and datasets utilized in various Edge AI projects. This table offers valuable insights into the methodologies and resources researchers and developers employ to design, implement, and evaluate Edge AI solutions.

Table 8 highlights the evaluation metrics utilized in the Edge AI requirements category. By standardizing the evaluation metrics used in Edge AI research, researchers can compare and assess the effectiveness of different approaches and identify areas for improvement.

5.3 Supporting technologies

Businesses across diverse sectors increasingly leverage Edge AI technologies to improve real-time monitoring and streamline processes. This technology offers enhanced data

Table 5 Challenges in the Edge AI requirements category

Category	Challenge
Compute offloading and edge computing	Delivering low-latency compute offloading services [47]
	Edge intelligence framework deployment [50]
	Scalable poultry monitoring using open hardware [52]
	Low-latency object identification and categorization [53]
	UDP-based offloading for time-critical services [58]
	HybridFL for MEC [59]
	Reliable AI service chains in edge intelligence cloud [60]
Energy efficiency	A lack of resources to maximize AIoT's energy efficiency [48]
Data and decision-making	Data-driven solution for Learning-based decision-making paradigm [49]
	Deep-learning-based image processing in IoT [55]
	Distributed ML strategy in multiuser MEC network [54]
	Incorporating game theory and RL in data offloading for UAVs [51]
	Accelerating DL training and inference [56]
Network and communication	Low-resolution reference layer [9]
	Network bandwidth [10]
	Sharing information between different resources [64]
Security and Privacy	Privacy-preserving [62]
	Cyber-security [65]
	Hackers attack the training dataset to corrupt the entire performance system [66]
	Cyber-physical-social systems [57]
Real-time response	Fast and real-time response [10]
	Timely response for the devices [63]
5G and beyond	Enabling techniques for URLLC in 5G and beyond [61]

security, reduced latency, and cost savings, necessitating a culture of innovation and a profound understanding of business dynamics. AI's contribution to scalability through drones, robots, and sensors is critical. Our taxonomy categorizes relevant articles into hardware, software, data management, immersive technologies, and networks.

5.3.1 Hardware and software

In Edge AI, hardware and software are crucial components and systems facilitating AI processing at the network's edge. The choice of hardware significantly influences the efficiency and performance of Edge AI applications. The following compilation of scrutinized papers within this category sheds light on various aspects:

Chang et al. [67] presented an AI fall detection system employing the PEFDM edge-computing architecture. Integrating AI computing capabilities, this system seamlessly operates on common edge computing systems, effectively alleviating computational burdens. Experiments with real subjects demonstrated PEFDM's efficacy in identifying falls in seniors, with accuracy being a significant benefit despite complexity limitations.

Karras et al. [68] proposed a framework based on an FPGA system-on-chip for rapid ML execution in an edge context. This structure facilitates the dynamic installation of ML processes, offering control locally or remotely, thereby showcasing exceptional adaptability. The design implemented a variant of the YOLO classifier, demonstrating its effectiveness, achieving outstanding results while managing resource usage effectively.

Vivek Parmar [69] delved into hardware design, focusing on XR applications such as hand recognition and eye separation. The study evaluated diastolic inference accelerators and a CPU through simulations to assess the impact of compression and hardware-specific limitations on performance. Various hardware options and cutting-edge technology nodes were compared to understand their implications on XR application performance.

Kulkarni et al. [70] investigated hardware-accelerated simulation-based reasoning over statistical models, integrating the distributed ABC inference method with AI chip technologies. Using a COVID-19 spread prediction model, both systems surpassed Tesla V100 GPUs, highlighting the potential of inference application in epidemiological models.

Table 6 Strong points, limitations, and solutions in Edge AI requirements category

Ref	Advantage	Limitation and solutions
[47]	Improved energy and delay cost reduction Provide flexibility with heterogeneous algorithms	Developing more generic MASM techniques can be employed in edge computing situations other than Edge AI scenarios
[48]	Lower energy consumption	Need for continuous monitoring for long-term sustainability
[49]	Effectiveness of the proposed algorithmic solution	The suggested learning-based framework must be further evaluated through trial research in the multienergy microgrids comprising additional energy sources and demands
[50]	Achieving advanced service management and network optimization	Address through iterative development and Ensure compatibility and interoperability
[51]	Efficacy in various operational strategies and circumstances	Need to expand this model according to the concepts of contract theory
[52]	Good performance	To find anomalies in the poultry, they should use video therapy and animal chicken analysis in future tasks
[53]	Fewer training iterations	Optimize resource allocation; Ensure efficient hardware/software integration
[54]	More bandwidth and computational capability, Minimizing energy consumption	Continuous validation and refinement based on simulation results
[55]	Efficient distribution of computational cost, working memory, and parameter memory. Designing scalable backbone for resource-constrained IoT platforms	Optimize for low memory and energy constraints; Address computational limitations
[56]	Achieving high sustained utilization for a range of neural network topologies	Address challenges related to precision optimization and on-chip hierarchy
[9]	Good accuracy	The work that has been presented is still in the experimental stage and has to be scaled up and modified to meet the requirements of the operational mode
[57]	Improving the operation efficiency, Improving the flexibility, Reducing time complexity	Optimize for scalability and diverse data sets; Address potential data privacy concerns
[58]	Demonstrating tolerance for minor image distortion while maintaining high accuracy, Improving normalized service reliability with ETR offloading scheme	Optimize for minimizing transmission latency
[59]	Achieving improvements in FL efficiency and energy consumption	Address challenges with unreliable end devices; Optimize for diverse scales of MEC systems
[60]	Optimizing deployment of VNFs and BVNFs on trusted edge servers, Achieving improvements in AISC reliability, throughput, and deployment cost	Address NP-hard problem challenges
[61]	Categorizing and exploring enabling techniques for URLLC in 5G and beyond	Address potential limitations in AI-enabled edge caching solutions
[10]	Low cost, High confidentiality and availability	Optimize for real-time problem solving; Address potential challenges in track and trace techniques
[62]	Effectiveness and practicality	Examine the optimal perturbation intensity for the selectively private data perturbation method
[63]	Speeding up the processing	The proposed system should be analyzed in more detail, and the production system should be enhanced
[64]	Significant image spam detection	Not enough details on their experiments
[65]	Better flexibility	Address potential challenges in distributed computing and classifier training; Optimize for scalability and diverse attack scenarios
[66]	The proposed model successfully undermines the classifiers	Optimize for diverse poisoning attack scenarios; Address potential classifier vulnerabilities

Wang et al. [71] proposed a hardware-level countermeasure for securing edge devices against cyber-attacks, utilizing ML based on Hardware Performance Counter (HPC) features. The study identified significant HPC events for accurate attack detection, demonstrating good detection accuracy with minimal latency, power consumption, and hardware overhead.

Table 7 Development phases, tools/methods, and datasets utilized in Edge AI requirements category

Development phase	Used tool or method	Used dataset
Implementation	AI methods [9, 48, 50, 52, 56, 57, 60–63, 65, 66]	Internet Movie Database, Rotten Tomatoes, Polarity datasets[9]
	CIFAR-10[53]	Real-world datasets[57]
	Python [9]	MNIST dataset[62]
		Manufacturing dataset[63]
		CTU dataset[65]
Simulation	Kernel density estimation method[47]	–
	AI methods [51, 54]	
Both simulation and implementation	Python [49, 58, 59]	A realistic islanding microgrid dataset[49]
	AI methods[49, 55, 64]	EfficientNetv1 and MobileNetv2[55] Dogs vs. Cats dataset [10]
	Monte Carlo [58]	
	Raspberry Pi 3 [10]	Real-time image dataset[64]

Table 8 Evaluation metrics utilized in Edge AI requirements category

Ref	Energy consumption	System cost	Delay and Latency	Throughput	Accuracy	Bandwidth	F-measure, Recall, and Precision
[47]	✔	✔	✔	✗	✗	✗	✗
[48]	✔	✗	✗	✗	✗	✗	✗
[49]	✔	✔	✔	✗	✗	✗	✗
[50]	✗	✗	✔	✔	✗	✗	✗
[51]	✗	✗	✔	✗	✗	✗	✗
[52]	✗	✗	✗	✗	✔	✗	✗
[53]	✗	✗	✔	✗	✔	✗	✗
[54]	✔	✔	✔	✗	✗	✔	✗
[55]	✔	✔	✗	✗	✗	✗	✗
[56]	✗	✗	✗	✗	✔	✗	✗
[9]	✗	✔	✔	✗	✗	✗	✗
[57]	✗	✔	✗	✗	✗	✗	✗
[58]	✗	✗	✔	✗	✔	✗	✗
[59]	✔	✗	✗	✗	✔	✗	✗
[60]	✗	✔	✗	✔	✗	✗	✗
[61]	✗	✗	✔	✔	✗	✗	✗
[10]	✗	✔	✔	✗	✗	✗	✗
[62]	✗	✗	✗	✗	✔	✗	✗
[63]	✗	✔	✔	✗	✗	✗	✗
[64]	✗	✗	✔	✗	✗	✗	✗
[65]	✗	✗	✗	✗	✗	✗	✗
[66]	✗	✗	✔	✗	✔	✗	✔

Mazzia et al. [72] proposed a real-time embedded solution for apple detection in orchards using the YOLOv3-tiny algorithm on various embedded platforms. The study demonstrated the feasibility of deploying the solution on inexpensive and power-efficient embedded hardware, showcasing mean average detection accuracy.

Hielscher et al. [73] introduced an approach for rapidly creating, testing, and deploying entire System-on-Chip (SoC) platforms with application-specific neural network (NN) hardware accelerators. Focusing on addressing the computational demands of NNs for Industrial IoT Systems, the authors demonstrated the feasibility of their approach

by generating a condition monitoring system for high-speed valves.

Hu et al. [74] suggested an SD-EIC infrastructure layout for quality control monitoring. The proposed system efficiently increases the use of hardware resources, enhances the effectiveness of product quality inspection, and decreases overall deployment costs. The system can adapt nimbly to various industrial circumstances.

Ayala-Romero et al. [75] presented a Bayesian learning framework for simultaneously setting the service and the Radio Access Network (RAN) to reduce overall energy usage while meeting service criteria for accuracy and latency. Their technique exceeded cutting-edge benchmarks based on neural networks and adapted to multiple hardware platforms and service needs.

Jayakodi et al. [76]introduced a comprehensive hardware and software co-design framework for energy-efficient Edge AI. The framework adaptively selects a deep neural network (DNN) based on input examples' difficulty levels, optimizing energy efficiency without compromising prediction accuracy.

Deng et al. [77] proposed an edge computing framework combining software orchestration and hardware acceleration to enhance edge intelligence performance. This framework aims to address resource constraints and latency considerations challenges, demonstrating improved system performance in low-latency edge intelligence deployment.

5.3.2 Data management

Efficient data management is paramount for the effectiveness of Edge AI, given the substantial volumes of data produced by edge devices. This data must be meticulously gathered, processed, and stored to derive valuable real-time insights. Data compression, filtering, and aggregation represent pivotal methods employed in data management solutions at the edge, aiming to alleviate the strain on constrained edge resources while ensuring the derivation of actionable insights. The following publications in this category have been subject to analysis:

Wang et al. [78] delved into the challenge of minimizing processing and communication time among users submitting various computing task requests. They formulated the resource and task allocation problem as an optimization task to fulfill users' delay requirements. The advantages include minimizing the maximum delay across all users and reducing the number of iterations required for convergence.

Wang et al. [79] proposed edge node-filling gated recurrent units. A movable edge node can collect data from nearby nodes and the history information of the present missing data node. The experimental results demonstrated that the edge computing-based missing value filling performs better in quality than other filling methods and significantly lowers energy consumption in the AIoT. The benefit of this paper is reducing energy consumption.

Munir et al. [80] presented a novel approach for AI-as-a-Service (AIaaS)-enabled edge computing, focusing on service aggregation for mobile agents. They integrated flow control techniques with density-based spatial clustering of applications with noise (DBSCAN), offering a low computational cost algorithm for mobile agent AI service aggregation. The advantages include improved server utilization and enhanced complexity analysis, while the drawback highlights the need for further experimentation.

Vita et al. [81] introduced a deep RL approach to oversee data migration in MEC setups, adapting continually as the system evolves. They utilized the Keras ML framework alongside the OMNeT + + /SimuLTE simulator for simulation. Initial findings suggest the strategy's feasibility, showing promising performance.

Zhaofeng et al. [82] introduced BlockTDM, a blockchain-based solution for trustworthy data management in edge computing environments. BlockTDM supports matrix-based multichannel data segmentation and encryption of sensitive information. Extensive experiments validate its security, reliability, and efficiency, making it suitable for edge computing scenarios requiring heightened security and data integrity.

Li et al. [83] introduced EdgeCare, a decentralized and collaborative data management system tailored for mobile healthcare systems, leveraging edge computing for enhanced performance. Numerical results and security analyses underscore the effectiveness of EdgeCare in protecting healthcare data and supporting efficient data trading.

Li et al. [84] suggested Edgent, a framework for on-demand, collaborative device-edge DNN co-inference. Edgent offers two architectural options: (1) DNN partitioning, which promotes device-edge collaboration, and (2) DNN right-sizing, which uses an early-exit mechanism to reduce the DNN inference latency and enable low-latency edge intelligence. The experimental evaluation of their prototype on a Raspberry Pi showed that Edgent is practical and effective for low-latency edge intelligence. The gain was efficiency in low-latency edge intelligence.

Lv et al. [85] proposed a technique that effectively solves the task delivery challenge in multiedge node computing centers without data duplication. Employing a heterogeneous edge collaborative storage approach that integrates computation and data resolves the inherent conflict between smart devices' limited computing and storage capabilities, thereby enhancing the performance of data processing applications.

5.3.3 Immersive technologies

To improve user experiences, the entertainment sector is progressively implementing Edge AI. For instance, content delivery networks (CDNs) use edge servers to lower streaming latency, enabling customers to enjoy more fluid video playing. Edge devices analyze sophisticated graphics and sensor data locally in augmented reality (AR) and virtual reality (VR) applications to produce immersive experiences with little lag. The papers from this category that have been examined are listed below.

Ishii et al. [86] developed a fighting game AI utilizing Monte-Carlo tree search (MCTS) enhanced with three highlight cues in the evaluation function to optimize action selection for engaging gameplay. The AI targets gaming for live-streaming platforms, enhancing audience entertainment. User research conducted using the FightingICE platform, employed in an international AI competition, revealed the suggested AI's superior entertainment value compared to a standard MCTS AI, with detailed assessments for each strategy considered in the study.

Yang et al. [87] proposed a game-based strategy to improve the age of information (AoI) by leveraging edge-based mobile devices and monitoring machines with AI. They designed an architecture for the Epidemic Prevention and Control Center (EPCC)-based medical condition surveillance system, employing edge servers to relay AI bots' biosensing data. Through experiments, their method demonstrated reduced AoI values for transmitted biosensing data under various parameter settings.

Li et al. [88] introduced an Edge AI-based protection system providing guidelines and resource allocation strategies for contemporary threat detection. Based on the edge Bayesian Stackelberg game and CTI, their approach includes a DRL-based resource allocation plan for rapid response at the edges. Experimental results indicated enhanced edge protection and contemporary threat defense capacity.

Ning et al. [89] proposed a cost-efficient in-home health monitoring system for the Internet of Medical Things (IoMT), focusing on intra-WBANs and beyond-WBANs communication. Using cooperative and non-cooperative games, their algorithm effectively reduced system-wide costs and expanded the reach of MEC-enabled patient benefits.

Using game theory, Long et al. [90] addressed the task assignment problem in collaborative edge and cloud environments within the IoT. They formulated a non-cooperative game among multiple edge data centers and introduced the greedy energy-aware algorithm (GEA) and best response algorithm (BRA). Results demonstrated the BRA algorithm's efficiency in reaching a solution close to Nash equilibrium for task allocation in this context.

5.3.4 Network

Edge AI relies heavily on robust and low-latency networks to facilitate seamless communication between edge devices and centralized cloud resources. Innovations such as 5G and edge computing are vital in ensuring dependable and swift data transmission, enabling efficient operations of remote monitoring, IoT, and autonomous systems. The papers examined in this category are summarized below:

Yang et al. [91] proposed the Edge-Based IoT Platform for AI (EBI-PAI), leveraging software defined-network (SDN) and serverless technology. EBI-PAI aims to enhance the Quality of Experience (QoE) by offering a unified service calling interface and automatic scheduling of services. Simulation results demonstrated significant improvements in QoE while maintaining budget constraints, although reliability considerations were lacking.

Wu et al. [92] introduced scalable frequency pooling (SFP) and arbitrary-oriented spatial pooling (ASP) for efficient feature extraction to improve classification precision. They also proposed an edge-cloud joint inference architecture for Facial Expression Recognition (FER), effectively balancing classification accuracy and inference latency.

Zhang et al. [93] presented RTSD-Net, a real-time strawberry detection system based on a modified lightweight YOLOv4-tiny architecture. Their modifications aimed to streamline the structure for efficient strawberry detection in field conditions, showing improved detection speed with a simplified model structure.

Baghban et al. [94] highlighted the importance of edge computing in IoT environments, emphasizing its support for time-critical applications. They proposed the DRL-Dispatcher method, integrating edge nodes into an edge federation, which demonstrated superior performance in optimizing request service provisioning, resulting in improved profit and lower response latency.

Mwase et al. [95] addressed the significant impact of AI in various industries and proposed an architecture for AI deployment in fully edge-based scenarios. Their strategies aimed to mitigate communication inefficiencies in distributed edge environments, showcasing performance improvements and encouraging multidisciplinary contributions to address deployment challenges.

Table 9 provides insights into the main scope and utilized technology within the Edge AI-supporting technologies category. This category encompasses a range of technologies to enhance Edge AI systems' capabilities, including hardware, software frameworks, communication protocols, and security mechanisms. Table 9 offers a comprehensive overview of the diverse approaches and tools employed to bolster Edge AI systems by categorizing the main scope and utilized technology. This facilitates a

Table 9 Main scope and used technology in Edge AI supporting technologies category

Category	The main scope and used technology
Game theory and optimization	Fighting game, Monte-Carlo tree search (MCTS) with highlight cues[86]
	Game-based optimization, Edge-based mobile devices, monitoring machines with AI[87]
	Intelligence defense approach, Edge Bayesian Stackelberg game, CTI, DRL-based resource allocation plan[88]
	Cost-efficient in-home health monitoring system, Cooperative game for intra-WBANs, decentralized non-cooperative game for beyond-WBANs[89]
	Task assignment problem in collaborative edge and cloud environments, Game theory perspective, M/M/1 and M/M/C queueing models[90]
	Task and Resource Allocation, Several stack RL [78]
	Incomplete Value Filling, Edge computing-based method, experimental validation[79] Stackelberg game-based optimization algorithm[83]
Service aggregation and management	Aggregation of Services for Mobile Agents, Data-driven strategy, DBSCAN [80]
	Deep RL, Keras ML framework[81]
	Blockchain-based trusted data management, user-defined encryption[82]
Edge Computing Technologies	DNN Inference On-Demand Acceleration Using Edge Computing, DNN partitioning, DNN right-sizing, configurators for static and dynamic bandwidth[84]
	Heterogeneous edge collaborative storage method[85]
	AI-Based Edge Computing for Fall Detection, PEFDM with AI computing capabilities[67]
	AI-Based Inference at the Edge, FPGA system-on-chip[68]
	Hardware design space exploration, DNN for hand recognition and eye separation[69]
	AI chip technologies, ABC inference method[70]
	ML based on Hardware Performance Counter (HPC) features[71]
	YOLOv3-tiny algorithm on embedded platforms[72]
	Rapid creation of SoC platforms with NN accelerators, Application-specific NN hardware accelerators[73]
	Image processing, virtual devices[74]
	Bayesian learning framework, The software-defined base station (vBS), GPU-enabled server[75]
	Hardware and software co-design for Edge AI, Adaptive DNN selection, resource management policy[76]
	Software orchestration, hardware acceleration[77]
QoE	Software-defined network (SDN), serverless technology[91]
Edge AI and inference	Edge AI-driven framework, Edge-cloud joint inference architecture[92]
	Real-time strawberry detection using DNN (RTSD-Net), Lightweight YOLOv4-tiny architecture[93]
IoT optimization and edge computing	Integration of edge nodes into an edge federation for optimizing IoT request service provisioning, DRL-Dispatcher, RL[94]
	Enabling AI in fully edge-based scenarios, Resource-constrained edge[95]

better understanding and comparison of the various methodologies and technologies used in this domain.

Table 10 offers a detailed analysis of the strong points, limitations, and proposed solutions within the Edge AI-supporting technologies category. This table provides valuable insights into the challenges and opportunities associated with enhancing Edge AI systems by identifying the strengths and weaknesses of different approaches and technologies. Moreover, the proposed solutions outlined in Table 10 offer potential pathways for overcoming existing limitations and improving the effectiveness and efficiency of Edge AI-supporting technologies, thereby contributing to the advancement of Edge AI systems.

Table 11 presents an overview of the development phases, tools/methods, and datasets utilized in the Edge AI-supporting technologies category. By cataloging the development phases and tools/methods employed in research and development efforts within this category, Table 11 offers valuable guidance for researchers and practitioners seeking to explore and implement Edge AI-supporting technologies. Additionally, including datasets utilized in these endeavors provides essential context for understanding the experimental settings and evaluation methodologies employed in studies focused on enhancing Edge AI systems.

Table 10 Strong points, limitations, and solutions in Edge AI supporting technologies category

Ref	Advantage	Limitation and solutions
[67]	Good accuracy in fall detection, Reduced computational burden through edge computing and addressed Privacy and bandwidth issues	Complexity is a limitation. Utilizes critical point data of known human falls to modify and train fall behaviors in an unknown environment
[68]	Rapid execution of ML techniques and dynamic installation of ML processes locally or remotely	Future efforts will be directed at lowering platform overhead, particularly in the software sector, and expanding it to include more task deployment models
[69]	Better energy consumption	Addressing compression and hardware-specific limitations. Comparison of various hardware options and technology nodes
[70]	Better speed	Constraints related to hardware availability are that both systems surpass Tesla V100 GPUs but not significantly faster
[71]	Effective countermeasures against cyber-attacks ML-based detection with negligible latency, power consumption, and hardware overhead	Complexity should solve
[72]	Real-time apple detection in orchards Feasibility on inexpensive and power-efficient embedded hardware	Mean average detection accuracy achieved. Suitable for applications like unmanned ground vehicles in agricultural settings
[73]	Rapid creation, testing, and deployment of SoC platforms Focus on addressing computational demands of NNs for Industrial IoT Systems	Security concerns should be addressed
[74]	Increased hardware resource utilization, enhanced product quality inspection effectiveness, Decreased deployment cost, Adaptability to various industrial circumstances	Scalability needs to address
[75]	Reduced overall energy usage, novel performance trade-offs, optimization opportunities, Adaptation to multiple hardware platforms and service needs	Real-world deployment
[76]	Energy-efficient Edge AI, adaptively selecting DNN based on difficulty levels	Incorporating feedback loops for continuous improvement could enhance the accuracy of DNN selection
[77]	Improved edge intelligence performance, addressing resource constraints and latency	Conducting extensive field trials and considering a more comprehensive range of edge computing scenarios
[78]	Minimizing the maximum delay across all users and the number of iterations required for convergence	Additional experiments and sensitivity analyses could be conducted to evaluate the robustness of the RL algorithms across various
[79]	Reducing energy consumption	Conducting experiments with a diverse set of datasets or real-world data could provide a more comprehensive assessment
[80]	Better server utilization, Better complexity analysis	More experiments are needed
[81]	Good performance	Need for more realistic traffic and mobility models
[82]	Security, availability, and efficiency were demonstrated, focusing on high-level security and credibility	Further testing and optimization of the blockchain architecture, considering larger-scale deployments, and addressing potential scalability issues
[83]	Improved overall performance, secure data uploading and sharing, efficient data trading	Investigating and addressing privacy concerns and conducting user studies to understand the willingness of individuals to engage in decentralized data trading
[84]	Effectiveness in low-latency edge intelligence	Multidevice application scenario support is not possible with the suggested framework
[85]	Improving efficiency, Reducing waiting times	Practical deployment or scalability
[86]	Improving entertainment value	Conducting experiments across various game genres and refining the AI to adapt to different gameplay styles
[87]	Better data transmission	Evaluating the strategy in a broader range of scenarios and considering adaptive approaches for different environments
[88]	Effectiveness	Continuous monitoring and adaptation of the protection system to evolving threats, incorporating dynamic threat modeling
[89]	Reduced system-wide costs increased the number of patients benefiting from MEC	Conducting experiments with more real-world scenarios, considering variability in patient behaviors and health conditions
[90]	Efficient solution close to Nash equilibrium for task assignment	Incorporating adaptive mechanisms in the algorithms to handle dynamic changes and uncertainties in the collaborative environment

Table 10 (continued)

Ref	Advantage	Limitation and solutions
[91]	Improving QoE	They should consider reliability
[92]	Balanced classification and low-latency inference increased the precision of classification	Hardware requirements Conducting experiments with a broader range of facial expression datasets to assess the generalizability and robustness of the proposed architecture
[93]	Faster operational speed for real-time strawberry detection, negative correlation between parameters and speed	Further testing and optimization in different environmental conditions, considering lighting changes and occlusions in field conditions
[94]	Outperformed profit and low response latency baseline approaches, effective reinforcement learning in optimization	Conducting experiments with diverse edge federation scenarios and refining the reinforcement learning model to adapt to changing conditions
[95]	Addressing communication inefficiencies in AI deployment in resource-constrained edge environments	Exploring lightweight communication strategies and optimizations to ensure practical implementation in edge environments with limited resources

Table 11 Development phases, tools/methods, and datasets utilized in Edge AI supporting technologies category

Development phase	Used tool or method	Used dataset
Implementation	AI methods(SVM, CNN, LSTM) [67, 69, 71, 73, 74, 77, 79, 83, 84] [88]	CMU Panoptic and Studio dataset[67] Picasso, and People-Art datasets[68] FPHAB, OpenEDS, DetNet, and MobileNetV2 datasets[69]
	YOLO classifier[68]	CIFAR-10 dataset [84]
	QKeras and Timeloop + Accelergy [69]	
	Bayesian learning framework[75]	
	Raspberry Pi[84]	
	Feasible solution (FS) algorithm[84]	
	Nash equilibrium[89]	
	Non-cooperative game[90]	
	Arbitrary-oriented spatial pooling (ASP) and scalable frequency pooling (SFP) [92]	
Simulation	AI methods(Q-learning DBSCAN) [78] [80] [81] [87] [91] [94]	–
	OMNeT + + [81]	
	SimuLTE simulator[81]	
Both simulation and implementation	TensorFlow [70]	Real-world benchmarks[76]
	Raspberry Pi 3[72]	
	blockchain method[82]	
	Monte-Carlo tree search (MCTS) [85]	
	AI methods[95]	
	YOLO[93]	

6 Results and comparison

In this section, we expound upon our findings from analyzing the articles in Sect. 5. Furthermore, we tackle the five RQs outlined in Sect. 3.

To evaluate the articles, we classified them into three primary groups: applications, requirements, and supporting technologies. The breakdown shows that applications comprise 40% of the collection, requirements represent

26%, and supporting technologies constitute 34%, as depicted in Fig. 7.

Regarding the RQs, the following analytical findings have also been provided:

6.1 RQ1: What are the use cases of Edge AI for big data?

Big data analytics facilitates the gathering and analysis of data at the edge, enabling real-time processing at the point of information production. This simplicity empowers scientists to integrate big data and AI effectively. Edge AI finds various applications, including facial recognition and real-time traffic updates on interconnected devices, smartphones, and semi-autonomous vehicles. Additionally, Edge AI extends its scope to support security cameras, robotics, smart speakers, drones, wearable health monitors, and robots. The versatility of Edge AI technology enables a myriad of applications, including its use in mobile medical equipment within the healthcare industry, enhancing patient monitoring, testing, and care [40, 96].

Subsequently, this section delves into specific use cases of Edge AI for big data, providing a comprehensive overview based on the insights from the reviewed articles.:

6.1.1 Autonomous vehicles

The capability of autonomous vehicles to process essential data for safe real-time driving positions them as a prime example of the utility of edge computing. The substantial volume of data generated by autonomous vehicles and the potential latency and connectivity challenges in transmitting data to the cloud underscores the importance of real-time edge analysis. Estimates indicating terabyte-scale data creation by autonomous cars highlight the limitations of relying solely on 5G for timely transmission and processing for self-driving cars[12]. Edge AI-equipped autonomous vehicles showcase superior decision-making speed and accuracy compared to humans. They excel in identifying road elements and streamlining route navigation, ultimately leading to faster and safer transportation [8, 26, 97].

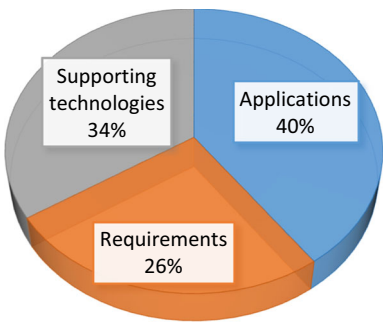

Fig. 7 Percentage of three categories of Edge AI

6.1.2 Smart cities

Civic authorities harness edge computing to build smart communities and improve road management through intelligent traffic systems. This application extends across multiple areas, including managing vehicle data to identify and alleviate congested locations. Edge computing assists civic authorities in promptly interpreting data from sensors on electricity grids, public infrastructure, and private buildings, enabling swift assessment and action [13, 98].

6.1.3 Robust security

Enhanced security stands out as a significant benefit for both business and consumer installations with the integration of edge computing. Organizations leverage edge computing for real-time surveillance and authorization methods, such as biometric scanning and video monitoring, ensuring that only authorized users engage in permitted actions. For example, businesses can implement biometric security solutions utilizing optical technology to verify access rights swiftly, enhancing overall security [64, 65].

6.1.4 Healthcare and smart hospital

The healthcare sector reaps significant benefits from edge computing in managing data generated by various medical devices. Edge devices adeptly process data from endpoint medical devices, discerning what information should be stored, discarded, or promptly addressed. Real-time data processing and edge computing are pivotal in healthcare delivery, improving operational efficiency and patient care. Edge AI applications contribute to high-accuracy thermal monitoring, inventory control, remote patient monitoring, and disease prediction [31, 99].

6.1.5 Manufacturing and industrial processes

The Industrial Internet of Things (IoT) has ushered millions of connected devices to manufacturing facilities, gathering data on production lines, equipment performance, and final goods. Edge computing plays a vital role in efficiently managing this data, especially when transporting data to centralized servers is impractical. Executives in the industrial sector harness edge computing as a critical component of an IoT ecosystem to monitor and control energy usage in facilities, ensure precise manufacturing processes, conduct product quality inspections, and predict mechanical failures [40, 41].

6.1.6 Virtual and augmented reality

Edge computing caters to the low-latency requirements of augmented and virtual reality applications, guaranteeing real-time processing of extensive datasets. This is vital for averting delays in image and instruction delivery. Edge computing enables a wide range of experiences, from assisting employees in their tasks and bolstering educational efforts to delivering distinctive and personalized experiences in entertainment and commerce. It effectively tackles challenges associated with bandwidth limitations and privacy concerns [4, 53, 79].

6.1.7 Streaming services and content delivery

Edge computing fulfills the low-latency demands of video streaming and content delivery, enriching user experiences through features such as search functionalities, content recommendations, and interactive capabilities. Media companies utilize edge computing to deliver original content, live events, and regional content seamlessly, ensuring a positive user experience as over-the-top streaming platforms gain prominence [2, 9, 66].

6.1.8 Smart homes

The growing volume of data generated by AI-enabled devices in smart homes, ranging from virtual assistants to security systems, poses challenges for provider networks. Home edge computing addresses this issue by storing more data locally, reducing dependence on centralized servers. This approach ensures real-time responses and enhances privacy for users[4, 8].

6.1.9 Privacy

Edge AI operations primarily process data locally, reducing the need for extensive data transmission to external locations. This local processing minimizes the risk of data misuse or mismanagement, although it does not eliminate potential security threats like hacking[13, 62, 64].

6.1.10 Reduction in internet bandwidth and cloud costs

Edge AI's local data analysis significantly reduces bandwidth consumption, offering cost-effective alternatives to using expensive cloud resources for in-the-moment fieldwork[1, 14, 57].

6.1.11 Energy efficiency

Edge AI contributes to reduced energy consumption by processing data locally, aligning with the efficient power consumption of edge computing devices. This contrasts with the higher energy requirements of AI processing in cloud data centers[39, 51].

6.1.12 Drones

Drones benefit from Edge AI in various applications, including mapping, traffic monitoring, and construction. Incorporating AI enables drones to analyze acquired data effectively, assisting in tasks such as face and object identification, predictive maintenance, real-time tracking, and maintenance[65, 100].

6.1.13 Traffic

The transportation sector utilizes Edge AI for applications such as autonomous ships and aircraft, handling large volumes of data to improve safety and accurately calculate passenger counts. Edge AI is pivotal in precise transportation management, optimizing efficiency and safety [65].

6.2 RQ2: How do big data drive Edge AI?

Technologies such as data analytics, ML, DL, and AI are driving innovation across businesses. Edge computing, AI, and IoT have evolved beyond novelty; they provide insights to create new services or lower costs. While AI algorithms demand significant data and processing resources, they have traditionally been implemented on the cloud. However, as AI grows, there's a rising reliance on cloud computing. Stakeholders recognize the necessity of deploying essential processing tasks directly onto devices rather than relying solely on the cloud. This approach enables them to serve more customers, especially as consumers spend more time on smart devices [26, 54, 84].

Furthermore, there are instances where conducting AI-based data computations and decisions locally, specifically on devices near the network's edge, becomes imperative for prompt action, such as in autonomous vehicles and fraud detection. The COVID-19 pandemic has placed immense stress on hospitals and healthcare workers, potentially resulting in a tenfold increase in patient populations. In such emergencies, finding methods that enable a single doctor to remotely monitor multiple patients becomes crucial. Various tools, including sensors attached to COVID-19 patients and relevant tests, generate vast amounts of data. In such scenarios, leveraging cutting-edge technologies capable of gathering, evaluating, mining, and deploying AI models becomes essential [52].

Subsequently, these models may be deployed on the cloud for applications like predictive maintenance. However, AI models must be situated at the edge in critical

situations, such as a sudden deterioration in a patient's condition requiring immediate notification to the doctor. This ensures swift decision-making without dependence on network connectivity or extensive data transfer across the network. Consequently, the market for edge computing is poised for continued growth in the future. Mission-critical and time-sensitive decisions can be executed swiftly, reliably, and with heightened security through AI at the edge [2, 3, 65].

Due to the rapid expansion of mobile computing and IoT applications, numerous mobile and IoT devices are interconnected, generating vast volumes of data at the network edge. This extensive data collection results in exceptionally high network bandwidth usage and latency in cloud data centers. Pushing the boundaries of AI to the network edge is essential to harness big data's potential fully. At viso.ai, we offer the Viso Suite, a comprehensive platform for Edge AI vision. Our platform empowers global corporate leaders to develop, deploy, and utilize edge computing-based computer vision applications [49, 78].

Edge AI, also called edge intelligence, combines edge computing with AI, executing AI algorithms on hardware or edge devices and processing data locally. By offering on-device AI, Edge AI provides quick reaction times with low latency, enhanced privacy, increased resilience, and efficient network traffic utilization. Emerging technologies such as ML, neural network acceleration, and model compression drive the adoption of Edge AI. Through ML edge computing, various sectors stand to benefit from a novel, reliable, and scalable AI system. Although still in its nascent stages, Edge AI is expected to advance AI development by bringing AI capabilities closer to real-world applications [6, 91, 101].

6.3 RQ3: What evaluation factors usually apply to Edge AI for big data?

In Edge AI, evaluation metrics are pivotal in enabling exploration and inquiry. These metrics serve as benchmarks to gauge the effectiveness of Edge AI systems in handling large-scale data. Researchers can significantly contribute to the progression of Edge AI solutions by scrutinizing the evaluation criteria utilized in their studies and comparing them with those employed in related research within their domain. Based on the categorization established by the requirements category, the following section provides a succinct definition of each evaluation metric identified in the analyzed articles. This sheds light on the landscape of evaluation metrics in Edge AI for big data. Delay: Delay refers to the time it takes for data to travel from a source (e.g., a sensor) to a destination (e.g., a processing unit or server) in an Edge AI system. It includes both processing and communication delays [47, 102].

- Latency: Latency is the time delay introduced during an Edge AI system data processing. It measures the time to perform a specific computation or task [21, 103].
- Bandwidth: The maximum amount of data transmitted over a communication channel in a given period. In Edge AI, it is essential to determine how much data can be sent between devices or nodes [20, 104].
- Accuracy: Accuracy measures how well an ML model or algorithm makes correct predictions or classifications. It quantifies correctly predicted instances' ratios to total cases [20].
- Recall: Recall is a criterion that assesses a model's capacity to locate each pertinent instance in a dataset when performing classification tasks. The ratio of real positives to true positives and false negatives determines this [98].
- F-Measure: The F-Measure is a single metric combining precision and recall to assess a model's performance in classification tasks. It is often used when there is an uneven distribution of classes[98, 102].
- Precision: Precision is a metric that measures the accuracy of positive predictions made by a model. It is the ratio of true positives to the sum of true positives and false positives [98, 102].
- Throughput: Throughput measures the rate at which a system can process data or tasks. Edge AI indicates how many computations or operations can be performed per unit of time [49, 103].
- Energy Consumption: Energy consumption refers to the amount of electrical power consumed by Edge AI devices or systems during operation. It is crucial in battery-powered or energy-efficient Edge AI applications [20, 21, 48].
- System Cost: System cost encompasses the financial investment required to build and deploy an Edge AI system. It includes hardware, software, development, maintenance, and operational expenses[20, 21, 75].

As depicted in Fig. 8, within the requirements category, time (including delay and latency) constitutes 31%, system cost accounts for 22%, energy consumption stands at 14%, accuracy represents 17%, throughput is at 7%, and other

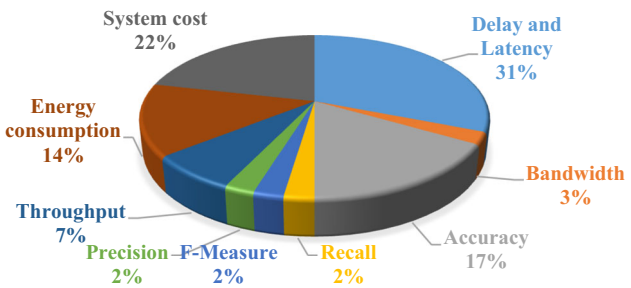

Fig. 8 Evaluation factors applied to the requirements category

metrics such as recall, precision, F-measure, and bandwidth each have a smaller percentage. In summary, Fig. 8 highlights that time, accuracy, and system cost emerge as the most crucial evaluation factors for Edge AI in handling big data, with energy consumption following closely. This suggests that enhancing the speed and efficiency of Edge AI processing poses a more significant challenge and priority than improving accuracy and reliability across diverse environments.

6.4 RQ4: What measurement environments are used for evaluating the Edge AI for big data?

Figure 9 demonstrates that most articles implemented their ideas and suggestions, which accounts for 64%. Also, 15% of the articles simulated their ideas, and 21% used both methods, i.e., implementation and simulation.

As shown in Fig. 10, it can be said that 21% of the articles used the dataset to evaluate their design, but 79% did not. A list of datasets that had the most repetitions is as follows:

- CIFAR-10 [4, 53]
- Internet Movie Database [9]
- Rotten Tomatoes [9]
- Polarity [9]
- PAMAP2 [31]
- MNIST [47]
- CTU dataset[65]
- Picasso Dataset [68]
- People-Art Dataset[68]
- FPHAB[69]
- OpenEDS[69]
- DetNet[69]
- MobileNetV2[69]
- Xception[38]
- NASNetMobile[38]
- InceptionResNetV2[38]
- Realistic Islanding Microgrid[49]
- Dogs vs. Cats dataset[10]

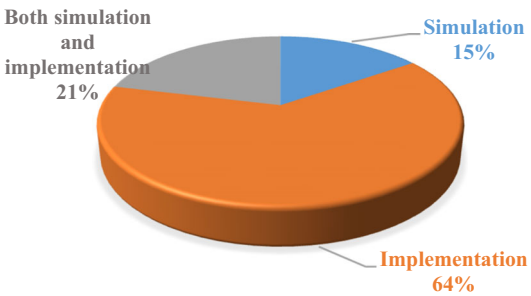

Fig. 9 Percentage of the development phase

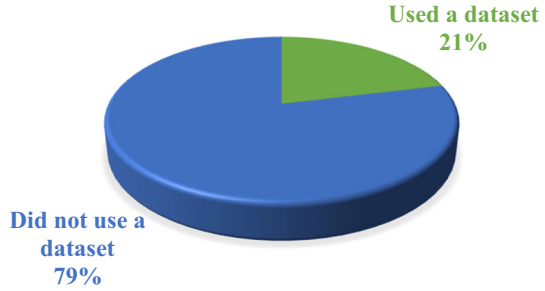

Fig. 10 Percentage of dataset usage

6.5 RQ5: What are Edge AI's challenges and open issues for big data?

Edge AI plays a pivotal role in real-world AI applications, as the conventional cloud computing approach may not be suitable due to the computational and data-intensive nature of these applications. Edge AI has emerged as a critical technology frontier, shaping the future of various industries. Presently, there is a competitive race to lead in this domain. This competition stems from the escalating volume of data devices generated, necessitating increased data processing at the edge. Furthermore, privacy concerns advocate for processing data locally, eliminating the need to transmit sensitive personal information in the cloud [1, 14, 57].

Edge AI is on the verge of becoming a must-have technology, given the enormous growth in data we currently face, because it can divert data from cloud data centers that are already overloaded. Combining edge computing and Edge AI will have many potential advantages. For instance, edge computing and Edge AI significantly decrease the time it takes to process data by machine and the quantity that travels to the cloud. Because of the reduced overall data transfer, it is more affordable and quick than the cloud. The future of computing is being revolutionized by edge computing and Edge AI. Older techniques will eventually be phased out in favor of edge computing and Edge AI. Some people will adjust to the advantages of this new technology. The future of computers has already been altered by affordability and increased security, and many people prefer the edge. It is only a matter of time before Edge AI follows the same path as edge computing [9, 49, 79].

Considering Edge AI's advantages and becoming more popular daily, some challenges and open issues need to be solved in the future. The challenges of the future are as follows:

6.5.1 Poor data quality

For the discovery and application of Edge AI, the poor quality of the data from the leading internet service providers worldwide is a serious obstacle. The output of an Edge AI system may likewise be poor quality if the sensor data are insufficient. False positives or negatives might result from this, and either could have dire repercussions. On the other hand, lost chances may result if the data are of low quality because the sensors are not appropriately maintained [14, 52].

6.5.2 Limited ML power

ML requires more processing power on edge computing hardware platforms. The computation performance of an edge or IoT device is the maximum for Edge AI infrastructure. Large, complex Edge AI models must often be streamlined to improve accuracy and efficiency before deployment to the Edge AI hardware [52].

6.5.3 Lost data

Data loss or erasure after processing is one concern associated with Edge AI. Edge AI has the benefit of deleting data after processing, which reduces costs. When the data are no longer valuable, AI identifies this and deletes it. The data may not necessarily be worthless in this scenario, which is an issue. An Edge AI system must be carefully designed and programmed to prevent data loss before implementation. Numerous Edge devices eliminate unnecessary data after collection, but if relevant data are dumped, that data are lost, and any analysis will be flawed [63].

6.5.4 Security

At the local level, security is at risk; however, there is a security benefit at the cloud and corporate levels. It is useless for a business to use a cloud provider with high levels of security if their local network is vulnerable to attack. Even if cloud-based security improves, most breaches are still caused by human mistakes, locally used programs, and passwords. Some digital specialists assert that edge computing's decentralized structure enhances its security. Locally gathered data, however, necessitates security for more sites [64, 65].

6.5.5 Computing power

Although Edge AI is fantastic, it still cannot match the computational power offered by cloud-based AI. As a result, only a few AI functions can be carried out on an Edge device. Cloud computing will still create and serve large models, but smaller models can be inferred on-device using edge devices. Small transfer learning tasks can be handled by edge devices as well [3, 51].

6.5.6 Insufficient hardware specifications

Edge computing is highly dependent on the hardware. It is made worse by the lack of standardized units in the Edge AI hardware currently on the market. In addition, several factors must be considered, including use cases, power usage, memory requirements, processors, etc. [52].

6.5.7 Integration of several components

One component of the AI model is the hardware. Using various models and frameworks to build apps is widespread among developers. It may be challenging to integrate, though. Businesses may also employ third-party platforms, which would necessitate a new interface with the software and equipment used for Edge AI [66].

6.5.8 Insufficient skills

Uses for Edge AI are constantly evolving, much like the industries accepting it. Knowing the most recent hardware selection, tool integration, model optimization for deployment and testing, etc., is necessary to address this requirement. It can be challenging to find a team of knowledgeable individuals about both Edge AI and the evolving technology stack [4, 6, 79].

7 Conclusion

This paper conducted an SLR on Edge AI for big data, employing a structured approach to analyze available articles technically. The research studies were categorized into three main classifications: applications, requirements, and supporting technologies. Additionally, the article introduced three taxonomies for classification purposes. The subsequent SLR explored the nuances of Edge AI concepts for big data. The findings revealed that the time factor held the highest percentage, indicating its paramount importance in evaluating Edge AI for big data. The breakdown of each factor's percentage is as follows: time (including delay and latency) at 31%, throughput at 7%, system cost at 22%, accuracy at 17%, recall, F-measure, and precision each at 2%, bandwidth at 2%, and energy consumption at 14%. In future research, addressing emerging topics such as poor data quality, limited ML power, data loss, security concerns, less computing power, lack of hardware standards, integration challenges with

multiple elements, and limited expertise could enhance the understanding of Edge AI for big data.

Acknowledgements Not applicable

Funding No funding was received.

Data availability Not applicable.

Declarations

Conflict of interest There is no conflict of interest.

References

1. Misra S, Tyagi AK, Piuri V, Garg L (2022) Artificial intelligence for cloud and edge computing. Springer, Berlin
2. Saleh H, Saber W, Rizk R (2022) Mobile computation offloading in mobile edge computing based on artificial intelligence approach: a review and future directions. In: The 8th international conference on advanced machine learning and technologies and applications (AMLTA2022). Springer International Publishing, Cham
3. Chang Z, Liu S, Xiong X, Cai Z, Tu G (2021) A survey of recent advances in edge-computing-powered artificial intelligence of things. IEEE Internet Things J 8(18):13849–13875
4. Hua H, Li Y, Wang T, Dong N, Li W, Cao J (2022) Edge computing with artificial intelligence: A machine learning perspective. ACM Comput Surveys
5. Zhou Z, Chen X, Li E, Zeng L, Luo K, Zhang J (2019) Edge intelligence: paving the last mile of artificial intelligence with edge computing. Proc IEEE 107(8):1738–1762
6. Deng S, Zhao H, Fang W, Yin J, Dustdar S, Zomaya AY (2020) Edge intelligence: the confluence of edge computing and artificial intelligence. IEEE Internet Things J 7(8):7457–7469
7. Hemmati A, Zarei M, Rahmani AM (2024) Big data challenges and opportunities in Internet of Vehicles: a systematic review. Int J Pervasive Comput Commun 20(2):308–342. https://doi.org/10.1108/IJPCC-09-2023-0250
8. Hemmati A, Rahmani AM (2022) The Internet of Autonomous Things applications: A taxonomy, technologies, and future directions. Internet Things 20:100635
9. Sankar H, Subramaniyaswamy V, Vijayakumar V, Kumar SA, Logesh R, Umamakeswari A (2020) Intelligent sentiment analysis approach using edge computing-based deep learning technique. Softw Pract Exp 50(5):645–657
10. Zhou C, Liu Q, Zeng R (2020) Novel defense schemes for artificial intelligence deployed in edge computing environment. Wirel Commun Mob Comput 2020:1–20
11. Hosseinzadeh M, Hemmati A, Rahmani AM (2022) Federated learning-based IoT: a systematic literature review. Int J Commun Syst 35(11):e5185
12. Hosseinzadeh M, Hemmati A, Rahmani AM 6G-enabled internet of things: vision, techniques, and open issues. Comput Model Eng Sci.
13. Lv Z, Chen D, Lou R, Wang Q (2021) Intelligent edge computing based on machine learning for smart city. Futur Gener Comput Syst 115:90–99
14. Labba C, Atitallah RB, Boyer A (2022) Combining artificial intelligence and edge computing to reshape distance education (Case Study: K-12 Learners). In: Rodrigo MM, Matsuda N, Cristea AI, Dimitrova V (eds) Artificial Intelligence in Education: 23rd International Conference, AIED 2022, Durham, UK, July 27–31, 2022, Proceedings, Part I. Springer International Publishing, Cham, pp 218–230. https://doi.org/10.1007/978-3-031-11644-5_18
15. Zhang L, Wu J, Mumtaz S, Li J, Gacanin H, Rodrigues JJPC (2019) Edge-to-edge cooperative artificial intelligence in smart cities with on-demand learning offloading. In: 2019 IEEE Global Communications Conference (GLOBECOM).
16. Rahman MA, Hossain MS, Showail AJ, Alrajeh NA, Ghoneim A (2023) AI-enabled IIoT for live smart city event monitoring. IEEE Internet Things J 10(4):2872–2880
17. Bui K-HN, Jung JJ (2019) Computational negotiation-based edge analytics for smart objects. Inform Sci 480:222–236. https://doi.org/10.1016/j.ins.2018.12.046
18. Neto AR, Silva TP, Batista T, Delicato FC, Pires PF, Lopes F (2021) Leveraging edge intelligence for video analytics in smart city applications. Information. https://doi.org/10.3390/info12010014
19. Chen N, Qiu T, Zhao L, Zhou X, Ning H (2021) Edge intelligent networking optimization for internet of things in smart city. IEEE Wirel Commun 28(2):26–31
20. Nasir M, Muhammad K, Ullah A, Ahmad J, Baik SW, Sajjad M (2022) Enabling automation and edge intelligence over resource constraint IoT devices for smart home. Neurocomputing 491:494–506
21. Lin Y-H (2019) Novel smart home system architecture facilitated with distributed and embedded flexible edge analytics in demand-side management. Int Trans Electr Energy Syst 29(6):e12014
22. Li C, Yang H, Sun Z, Yao Q, Zhang J, Yu A, Vasilakos AV, Liu S, Li Y (2023) High-precision cluster federated learning for smart home: an edge-cloud collaboration approach. IEEE Access 11:102157–102168
23. Zhang S, Li W, Wu Y, Watson P, Zomaya A (2018) Enabling edge intelligence for activity recognition in smart homes. In: 2018 IEEE 15th international conference on mobile Ad Hoc and sensor systems (MASS).
24. Benadda B, Benabdellah A (2022) Hardware design and integration of low-cost edge AI smart power management and home automation. In: 2022 International conference on artificial intelligence of things (ICAIoT).
25. Chen Y-Y, Chen M-H, Chang C-M, Chang F-S, Lin Y-H (2021) A Smart Home energy management system using two-stage non-intrusive appliance load monitoring over fog-cloud analytics based on tridium's niagara framework for residential demand-side management. Sensors 21(8):2883. https://doi.org/10.3390/s21082883
26. Su CL, Lai WC, Zhang YK, Guo TJ, Hung YJ, Chen HC (2020) Artificial intelligence design on embedded board with edge computing for vehicle applications. In: 2020 IEEE third international conference on artificial intelligence and knowledge engineering (AIKE).w
27. Sonmez C, Tunca C, Ozgovde A, Ersoy C (2021) Machine learning-based workload orchestrator for vehicular edge computing. IEEE Trans Intell Transp Syst 22(4):2239–2251
28. Liu T, Tang L, Wang W, He X, Chen Q, Zeng X, Jiang H (2022) Resource allocation in DT-assisted internet of vehicles via edge intelligent cooperation. IEEE Internet Things J 9(18):17608–17626
29. Jiang X, Yu FR, Song T, Leung VCM (2021) Edge intelligence for object detection in blockchain-based internet of vehicles: convergence of symbolic and connectionist AI. IEEE Wirel Commun 28(4):49–55
30. Alladi T, Kohli V, Chamola V, Yu FR, Guizani M (2021) Artificial intelligence (AI)-empowered intrusion detection architecture for the internet of vehicles. IEEE Wirel Commun 28(3):144–149

31. Aazam M, Zeadally S, Flushing EF (2021) Task offloading in edge computing for machine learning-based smart healthcare. Comput Netw 191:108019
32. Hayyolalam V, Aloqaily M, Ozkasap O, Guizani M (2021) Edge intelligence for empowering IoT-based healthcare systems. IEEE Wirel Commun 28:6–14
33. Li J, Cai J, Khan F, Rehman AU, Balasubramaniam V, Sun J, Venu P (2020) A secured framework for SDN-based edge computing in IoT-enabled healthcare system. IEEE Access 8:135479–135490
34. Guo Y, Liu F, Cai Z, Chen L, Xiao N (2020) FEEL: a federated edge learning system for efficient and privacy-preserving mobile healthcare. In: Proceedings of the 49th international conference on parallel processing. Association for Computing Machinery: Edmonton, AB, Canada
35. Mansour RF, Amraoui AE, Nouaouri I, Díaz VG, Gupta D, Kumar S (2021) Artificial intelligence and internet of things enabled disease diagnosis model for smart healthcare systems. IEEE Access 9:45137–45146
36. Nguyen T, Gia TN (2023) Novel smart homecare IoT system with edge-AI and blockchain. In: Namasudra Suyel, Akkaya Kemal (eds) Blockchain and its applications in industry 4.0. Springer Nature Singapore, Singapore, pp 293–317. https://doi.org/10.1007/978-981-19-8730-4_10
37. Yadav R, Zhang W, Elgendy IA, Dong G, Shafiq M, Laghari AA, Prakash S (2021) Smart healthcare: RL-based task offloading scheme for edge-enable sensor networks. IEEE Sens J 21(22):24910–24918
38. Rangarajan AK, Ramachandran HK (2021) A preliminary analysis of AI based smartphone application for diagnosis of COVID-19 using chest X-ray images. Expert Syst Appl 183:115401
39. Liao H, Zhou Z, Zhao X, Zhang L, Mumtaz S, Jolfaei A, Ahmed SH, Bashir AK (2020) Learning-based context-aware resource allocation for edge-computing-empowered industrial IoT. IEEE Internet Things J 7(5):4260–4277
40. Ren S, Kim JS, Cho WS, Soeng S, Kong S, Lee KH (2021) Big Data platform for intelligence industrial IoT sensor monitoring system based on edge computing and AI. In: 2021 International conference on artificial intelligence in information and communication (ICAIIC).
41. Foukalas F, Tziouvaras A (2021) Edge artificial intelligence for industrial internet of things applications: an industrial edge intelligence solution. IEEE Ind Electron Mag 15(2):28–36
42. Long H, Miao Y, Gaoxiang W, Hassan MM, Humar I (2019) iRobot-factory: an intelligent robot factory based on cognitive manufacturing and edge computing. Future Generat Comput Syst 90:569–577. https://doi.org/10.1016/j.future.2018.08.006
43. Ghahramani M, Qiao Y, Zhou MC, O'Hagan A, Sweeney J (2020) AI-based modeling and data-driven evaluation for smart manufacturing processes. IEEE/CAA J Automatica Sinica 7(4):1026–1037
44. Lin CC, Deng DJ, Chih YL, Chiu HT (2019) Smart manufacturing scheduling with edge computing using multiclass deep Q network. IEEE Trans Industr Inf 15(7):4276–4284
45. Sathesh A (2020) Artificial intelligence based edge computing framework for optimization of mobile communication. J ISMAC 2(3):160–165. https://doi.org/10.36548/jismac.2020.3.004
46. Hao Y, Miao Y, Hu L, Hossain MS, Muhammad G, Amin SU (2019) Smart-Edge-CoCaCo: AI-enabled smart edge with joint computation, caching, and communication in heterogeneous IoT. IEEE Netw 33(2):58–64
47. Zhang W, Zhang Z, Zeadally S, Chao HC, Leung VCM (2019) MASM: a multiple-algorithm service model for energy-delay optimization in edge artificial intelligence. IEEE Trans Industr Inf 15(7):4216–4224
48. Zhu S, Ota K, Dong M (2022) Energy-efficient artificial intelligence of things with intelligent edge. IEEE Internet Things J 9(10):7525–7532
49. Dong W, Yang Q, Li W, Zomaya AY (2021) Machine-learning-based real-time economic dispatch in islanding microgrids in a cloud-edge computing environment. IEEE Internet Things J 8(17):13703–13711
50. Zeb S, Rathore MA, Hassan SA, Raza S, Dev K, Fortino G (2023) Toward AI-enabled NextG networks with edge intelligence-assisted microservice orchestration. IEEE Wirel Commun 30(3):148–156
51. Fragkos G, Kemp N, Tsiropoulou EE, Papavassiliou S (2020) Artificial intelligence empowered UAVs data offloading in mobile edge computing. In: ICC 2020 - 2020 IEEE international conference on communications (ICC).
52. Debauche O, Mahmoudi S, Mahmoudi SA, Manneback P, Bindelle J, Lebeau F (2020) Edge computing and artificial intelligence for real-time poultry monitoring. Proc Comput Sci 175:534–541. https://doi.org/10.1016/j.procs.2020.07.076
53. Chen D, Xie LJ, Kim B, Wang L, Hong CS, Wang LC, Han Z (2020) Federated learning based mobile edge computing for augmented reality applications. In: 2020 international conference on computing, networking and communications (ICNC).
54. Guo Y, Zhao R, Lai S, Fan L, Lei X, Karagiannidis GK (2022) Distributed machine learning for multiuser mobile edge computing systems. IEEE J Select Topics Signal Process 16(3):460–473
55. Paissan F, Ancilotto A, Farella E (2022) PhiNets: a scalable backbone for low-power AI at the edge. ACM Trans Embedded Comput Syst 21:1–18
56. Fleischer B, Shukla S, Ziegler M, Silberman J, Oh J, Srinivasan V, Choi J, Mueller S, Agrawal A, Babinsky T, Cao N, Chen CY, Chuang P, Fox T, Gristede G, Guillorn M, Haynie H, Klaiber M, Lee D, Lo SH, Maier G, Scheuermann M, Venkataramani S, Vezyrtzis C, Wang N, Yee F, Zhou C, Lu PF, Curran B, Chang L, Gopalakrishnan K (2018) A scalable multi-TeraOPS deep learning processor core for AI trainina and inference. In: 2018 IEEE symposium on VLSI circuits.
57. Zhang W, Chen X, Liu Y, Xi Q (2020) A distributed storage and computation k-nearest neighbor algorithm based cloud-edge computing for cyber-physical-social systems. IEEE Access 8:50118–50130
58. Liu J, Zhang Q (2020) To improve service reliability for AI-powered time-critical services using imperfect transmission in MEC: an experimental study. IEEE Internet Things J 7(10):9357–9371
59. Wu W, He L, Lin W, Mao R (2021) Accelerating federated learning over reliability-agnostic clients in mobile edge computing systems. IEEE Trans Parallel Distrib Syst 32(7):1539–1551
60. Qiu Y, Liang J, Leung VCM, Chen M (2023) Online security-aware and reliability-guaranteed ai service chains provisioning in edge intelligence cloud. IEEE Trans Mobile Comput, 1–16
61. Mutalemwa LC, Shin S (2020) A classification of the enabling techniques for low latency and reliable communications in 5G and beyond: AI-enabled edge caching. IEEE Access 8:205502–205533
62. Zhang J, Zhao Y, Wang J, Chen B (2020) FedMEC: improving efficiency of differentially private federated learning via mobile edge computing. Mobile Netw Appl 25(6):2421–2433
63. Shahbazi Z, Byun Y (2021) Improving transactional data system based on an edge computing–blockchain–machine learning integrated framework. Processes 9:92

64. Makkar A, Ghosh U, Rawat DB, Abawajy JH (2022) FedLearnSP: preserving privacy and security using federated learning and edge computing. IEEE Consumer Electron Magazine 11(2):21–27
65. Kozik R, Choraś M, Ficco M, Palmieri F (2018) A scalable distributed machine learning approach for attack detection in edge computing environments. J Parallel Distribut Comput 119:18–26
66. Manoharan P, Walia R, Iwendi C, Ahanger TA, Suganthi ST, Kamruzzaman MM, Bourouis S, Alhakami W, Hamdi M (2022) SVM-based generative adverserial networks for federated learning and edge computing attack model and outpoising. Expert Syst. https://doi.org/10.1111/exsy.13072
67. Chang WJ, Hsu CH, Chen LB (2021) A pose estimation-based fall detection methodology using artificial intelligence edge computing. IEEE Access 9:129965–129976
68. Karras K, Pallis E, Mastorakis G, Nikoloudakis Y, Batalla JM, Mavromoustakis CX, Markakis E (2020) A hardware acceleration platform for AI-based inference at the edge. Circ Syst Signal Process 39(2):1059–1070. https://doi.org/10.1007/s00034-019-01226-7
69. Sarwar SS, Parmar V, Li Z, Lee H-HS, de Salvo H, Suri M (2023) Memory-oriented design-space exploration of edge-AI hardware for XR applications. arXiv
70. Kulkarni S, Tsyplikhin A, Krell MM, Moritz CA (2020) Accelerating simulation-based inference with emerging AI hardware. In: 2020 international conference on rebooting computing (ICRC).
71. Wang H, Sayadi H, Dinakarrao SMP, Sasan A, Rafatirad S, Homayoun H (2021) Enabling micro AI for securing edge devices at hardware level. IEEE J Emerg Select Topics Circ Syst 11(4):803–815. https://doi.org/10.1109/JETCAS.2021.3126816
72. Mazzia V, Khaliq A, Salvetti F, Chiaberge M (2020) Real-time apple detection system using embedded systems with hardware accelerators: an edge AI application. IEEE Access 8:9102–9114
73. Hielscher L, Bloeck A, Viehl A, Reiter S, Staiger M, Bringmann O (2021) Platform generation for edge AI devices with custom hardware accelerators. In: 2021 IEEE 19th international conference on industrial informatics (INDIN).
74. Hu P, He C, Zhu Y (2022) The Scheme and system architecture of product quality inspection based on software-defined edge intelligent controller (SD-EIC) in industrial internet of things. In: 2022 IEEE International conference on smart internet of things (SmartIoT).
75. Ayala-Romero JA, Garcia-Saavedra A, Costa-Pérez X, Iosifidis G (2023) EdgeBOL: a bayesian learning approach for the joint orchestration of vRANs and mobile edge AI. IEEE/ACM Transactions on Networking, 1–0.
76. Jayakodi NK, Doppa JR, Pande PP (2021) A general hardware and software Co-design framework for energy-efficient edge AI. In: 2021 IEEE/ACM international conference on computer aided design (ICCAD).
77. Deng C, Fang X, Wang X, Law K (2022) Software orchestrated and hardware accelerated artificial intelligence: toward low latency edge computing. IEEE Wirel Commun 29(4):110–117
78. Wang S, Chen M, Liu X, Yin C, Cui S, Vincent Poor H (2021) A machine learning approach for task and resource allocation in mobile-edge computing-based networks. IEEE Internet Things J 8(3):1358–1372
79. Wang T, Ke H, Jolfaei A, Wen S, Haghighi MS, Huang S (2022) Missing value filling based on the collaboration of cloud and edge in artificial intelligence of things. IEEE Trans Industr Inf 18(8):5394–5402
80. Munir MS, Abedin SF, Hong CS (2019) Artificial intelligence-based service aggregation for mobile-agent in edge computing. In: 2019 20th Asia-pacific network operations and management symposium (APNOMS).
81. Vita FD, Bruneo D, Puliafito A, Nardini G, Virdis A, Stea G (2018) A deep reinforcement learning approach for data migration in multi-access edge computing. In: 2018 ITU Kaleidoscope: machine learning for a 5G future (ITU K).
82. Zhaofeng M, Xiaochang W, Jain DK, Khan H, Hongmin G, Zhen W (2020) A blockchain-based trusted data management scheme in edge computing. IEEE Trans Industr Inf 16(3):2013–2021
83. Li X, Huang X, Li C, Yu R, Shu L (2019) EdgeCare: leveraging edge computing for collaborative data management in mobile healthcare systems. IEEE Access 7:22011–22025
84. Li E, Zeng L, Zhou Z, Chen X (2020) Edge AI: on-demand accelerating deep neural network inference via edge computing. IEEE Trans Wireless Commun 19(1):447–457
85. Lv Z, Qiao L, Verma Kavita S (2021) AI-enabled IoT-edge data analytics for connected living. ACM Trans Internet Technol 21(4):1–20. https://doi.org/10.1145/3421510
86. Ishii R, Ito S, Thawonmas R, Harada T (2019) A fighting game AI using highlight cues for generation of entertaining gameplay. In: 2019 IEEE Conference on Games (CoG).
87. Yang Y, Wang W, Yin Z, Xu R, Zhou X, Kumar N, Alazab M, Gadekallu TR (2022) Mixed game-based AOI optimization for combating COVID-19 with AI bots. IEEE J Sel Areas Commun 40(11):3122–3138
88. Li H, Wu J, Xu H, Li G, Guizani M (2022) Explainable intelligence-driven defense mechanism against advanced persistent threats: a joint edge game and AI approach. IEEE Trans Dependable Secure Comput 19(2):757–775
89. Ning Z, Dong P, Wang X, Hu X, Guo L, Hu B, Guo Y, Qiu T, Kwok RYK (2021) Mobile edge computing enabled 5G Health monitoring for internet of medical things: a decentralized game theoretic approach. IEEE J Sel Areas Commun 39(2):463–478
90. Long S, Long W, Li Z, Li K, Xia Y, Tang Z (2021) A game-based approach for cost-aware task assignment with QoS constraint in collaborative edge and cloud environments. IEEE Trans Parallel Distrib Syst 32(7):1629–1640
91. Yang S, Xu K, Cui L, Ming Z, Chen Z, Ming Z (2021) EBI-PAI: toward an efficient edge-based IoT platform for artificial intelligence. IEEE Internet Things J 8(12):9580–9593
92. Yirui W, Zhang L, Zonghua G, Hu L, Wan S (2023) Edge-AI-driven framework with efficient mobile network design for facial expression recognition. ACM Trans Embedded Comput Syst 22(3):1–17. https://doi.org/10.1145/3587038
93. Zhang Y, Jiya Yu, Chen Y, Yang W, Zhang W, He Y (2022) Real-time strawberry detection using deep neural networks on embedded system (rtsd-net): An edge AI application. Comput Electron Agric 192:106586
94. Baghban H, Rezapour A, Hsu CH, Nuannimnoi S, Huang CY (2022) Edge-AI: IoT request service provisioning in federated edge computing using actor-critic reinforcement learning. IEEE Trans Eng Manag, 1–10.
95. Mwase C, Jin Yi, Westerlund T, Tenhunen H, Zou Z (2022) Communication-efficient distributed AI strategies for the IoT edge. Futur Gener Comput Syst 131:292–308
96. Hemmati A, Zarei M, Souri A (2023) Blockchain-based internet of vehicles (BIoV): a systematic review of surveys and reviews. Secur Privacy. https://doi.org/10.1002/spy2.317
97. Hemmati A, Zarei M, Rahmani AM (2023) A systematic review of congestion control in internet of vehicles and vehicular ad hoc networks: techniques, challenges, and open issues. Int J Commun Syst. https://doi.org/10.1002/dac.5625
98. Hosseinzadeh M, Hemmati A, Rahmani AM (2022) Clustering for smart cities in the internet of things: a review. Cluster

Comput 25(6):4097–4127. https://doi.org/10.1007/s10586-022-03646-8
99. Hemmati A, Rahmani A (2022) Internet of medical things in the COVID-19 Era: a systematic literature review. Sustainability 14:12637
100. Hemmati A, Zarei M, Souri A (2023) UAV-based internet of vehicles: a systematic literature review. Intell Syst Appl 18:200226
101. Hemmati A, Arzanagh HM, Rahmani AM (2023) A taxonomy and survey of big data in social media. Concur Comput: Pract Exp. https://doi.org/10.1002/cpe.7875
102. Gasmi R, Harous S (2022) Robust connectivity-based internet of vehicles clustering algorithm. Wireless Pers Commun 125(4):3153–3185
103. Aggarwal S, Goswami D, Hooda M, Chakravarty A, Kar A, Vasudha (2020) Recommendation systems for interactive multimedia entertainment. In: Hemanth J, Bhatia M, Geman O (eds) Data visualization and knowledge engineering: spotting data points with artificial intelligence. Springer International Publishing, Cham, pp 23–48. https://doi.org/10.1007/978-3-030-25797-2_2
104. Bousbaa FZ, Kerrache CA, Lagraa N, Hussain R, Yagoubi MB, Tahari AEK (2022) Group data communication in connected vehicles: a survey. Vehicular Communications 37:100518