Artificial Intelligence Review (2025) 58:93
https://doi.org/10.1007/s10462-024-11033-5

# Edge deep learning in computer vision and medical diagnostics: a comprehensive survey

**Yiwen Xu**[1] · **Tariq M. Khan**[1] · **Yang Song**[1] · **Erik Meijering**[1]

Accepted: 17 November 2024 / Published online: 17 January 2025


**Abstract**
Edge deep learning, a paradigm change reconciling edge computing and deep learning, facilitates real-time decision making attuned to environmental factors through the close integration of computational resources and data sources. Here we provide a comprehensive review of the current state of the art in edge deep learning, focusing on computer vision applications, in particular medical diagnostics. An overview of the foundational principles and technical advantages of edge deep learning is presented, emphasising the capacity of this technology to revolutionise a wide range of domains. Furthermore, we present a novel categorisation of edge hardware platforms based on performance and usage scenarios, facilitating platform selection and operational effectiveness. Following this, we dive into approaches to effectively implement deep neural networks on edge devices, encompassing methods such as lightweight design and model compression. Reviewing practical applications in the fields of computer vision in general and medical diagnostics in particular, we demonstrate the profound impact edge-deployed deep learning models can have in real-life situations. Finally, we provide an analysis of potential future directions and obstacles to the adoption of edge deep learning, with the intention to stimulate further investigations and advancements of intelligent edge deep learning solutions. This survey provides researchers and practitioners with a comprehensive reference shedding light on the critical role deep learning plays in the advancement of edge computing applications.

**Keywords** Edge computing · Deep learning · Computer vision · Medical diagnostics · Lightweight neural networks

## 1 Introduction

Edge computing (Fig. 1), an emerging computing paradigm, is increasingly impacting the modern technological landscape by executing computational tasks near the data source. This not only increases response speed but also brings improvements in security, privacy

Yiwen Xu and Tariq M. Khan have contributed equally to this study.

Extended author information available on the last page of the article

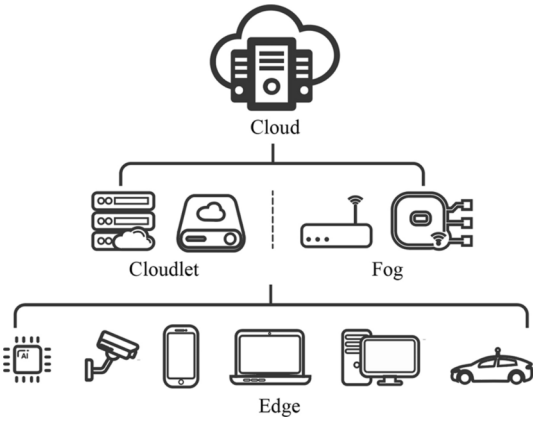


**Fig. 1** Illustrative overview of edge computing in relation to cloud computing

protection, scalability, and distributed processing. Edge computing (Cao et al. 2020) aims to process data on devices proximate to the data-generating source (edge devices) rather than transmitting it to distant cloud servers. In fact, edge computing and cloud computing are not mutually exclusive (Shi et al. 2016). Instead, the edge serves as a complement and extension to the cloud, reducing communication latency, protecting data privacy, and offering faster responses. Variations of edge computing, such as fog computing (Bonomi et al. 2012) and cloudlets (Babar et al. 2021), converge on a central philosophy: decentralising computational capabilities closer to the data origins. This not only alleviates the computational burden on cloud centres, but also provides the necessary support for real-time or near-real-time applications, catering to the increasing demands of the modern digital society (Chen and Ran 2019).

Deep learning (DL), particularly the development of advanced convolutional neural networks (CNNs), has made significant strides in many areas of computer vision, offering powerful tools for image and video data processing (Voulodimos et al. 2018; Dong et al. 2021; Esteva et al. 2021; Zhang and Qie 2023; Jiang et al. 2023). Typically, DL models are created through intensive training on high-performance computing hardware before being deployed on devices. However, even though the time required to execute a model is significantly shorter than to train it, the inherent computational complexity of modern models often poses challenges when deploying them to resource-constrained edge devices. To tackle this, researchers are shifting focus towards model compression and lightweight model design, ensuring that deep learning models operate efficiently on edge devices (Kumar et al. 2021; Zhang et al. 2021a; Huang et al. 2023a).

In recent years, with the advancement of edge devices and the rapid development of DL technologies, we have entered a new era marked by ubiquitous and continuously increasing levels of on-device intelligence. The amalgamation of edge computing and deep learning, often termed Edge DL (Shuvo et al. 2022), heralds new possibilities for various applications. Edge DL refers to the deployment of DL algorithms and models on edge devices, capitalising on the benefits of both local computation and artificial intelligence. This convergence has brought about advancements in autonomous driving and intelligent traffic control (Wang et al. 2021a; Liu et al. 2023a; Wan et al. 2022), industrial automation (Yi et al. 2022; Hang et al. 2022; Chen et al. 2022b), agriculture monitoring (Dang et al. 2020; Albattah et al. 2023), retail monitoring (Lachhab 2023; Kanjula et al. 2022), wildlife conser-

vation (Gotthard and Broström 2023; Arshad et al. 2020), smart cities (Avvenuti et al. 2022; Di Benedetto et al. 2022), sport analytics (Cui and Hu 2022; Cioppa et al. 2020), and medical diagnostics (Kara et al. 2023; Iqbal et al. 2023; Khan et al. 2024; Matloob Abbasi et al. 2024; Khan et al. 2024). Specifically, in medical diagnostics, edge computing can provide instantaneous feedback through medical image analysis and patient monitoring, enhancing diagnostic efficiency and patient safety (Greco et al. 2020; Idlahcen et al. 2024). Edge DL represents a paradigm shift in the deployment of deep neural networks (DNNs), bringing artificial intelligence (AI) closer to the data source and facilitating real-time, autonomous decision-making that is sensitive to and informed by the surrounding environment.

According to recent reports, the global edge computing in healthcare market was valued at USD 5.28 billion in 2023 (Research, Accessed 25 July 2024), and it is projected to reach USD 12.9 billion by 2028, growing at a compound annual growth rate (CAGR) of 26.1% from 2022 to 2028 (Markets and Markets, Accessed 25 July 2024). The adoption rate of DL algorithm-based medical devices has also seen a significant increase, demonstrating steady growth (Intel, Accessed 25 July 2024; Nvidia, Accessed 25 July 2024). These advancements highlight the potential of DL algorithms in enhancing the efficiency, and accessibility of medical diagnostics.

In this paper, we present a comprehensive survey of Edge DL with a focus on applications in computer vision in general and medical diagnostics in particular (Fig. 2). Providing a more in-depth discussion of the state-of-the-art specifically in these domains, our survey complements previous works exploring the fusion of edge computing and DL (Table 1). We begin by elucidating the fundamental concepts, terminologies, and technical merits of edge computing (Sect. 2), as well as categorising and exploring existing edge devices (Sect. 3). Subsequently, our focus shifts to discussing strategies for model compression and lightweight model design (Sect. 4), ensuring efficient CNN operation on edge devices. Next, we review the applications of edge computing in computer vision in general (Sect. 5) and

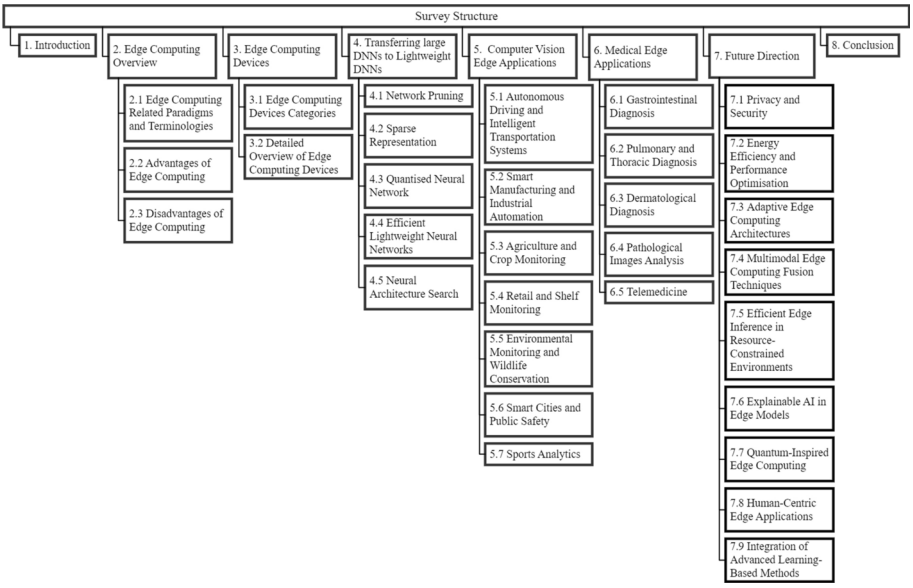


**Fig. 2** Schematic overview of the survey

**Table 1** Related surveys on edge computing and deep learning

| Paper | Summary | Scope | | | | |
|---|---|---|---|---|---|---|
| | | Hardware | Lightweight | Compression | Vision | Medical |
| Chen and Ran (2019) | Reviews recent deep learning methods for edge computing, focusing on IoT applications and the benefits of edge over cloud computing | ✓ | ✓ | ✓ | ✓ | ✗ |
| Greco et al. (2020) | Discusses the role of IoT in the shift towards AI at the edge in healthcare | ✗ | ✗ | ✗ | ✓ | ✓ |
| Liu et al. (2022) | Reviews deep learning for edge AI, including model optimisation and potential future work | ✓ | ✓ | ✓ | ✓ | ✗ |
| Mendez et al. (2022) | Investigates edge intelligence, its motivations, challenges, and prospective evolution | ✓ | ✗ | ✓ | ✓ | ✗ |
| Murshed et al. (2021) | Details machine learning implementation at the edge, focusing on practical operational aspects | ✓ | ✓ | ✓ | ✓ | ✗ |
| Wang et al. (2020c) | Surveys deep learning in edge computing smart applications | ✗ | ✗ | ✗ | ✓ | ✗ |
| Wang et al. (2020a) | Explores the synergy of edge computing and deep learning, addressing custom frameworks and future research | ✓ | ✗ | ✓ | ✓ | ✗ |
| Ours | Examines the deployment of computer vision and deep learning in edge computing with an emphasis on lightweight models and model compression, detailing advances, application, and enhancements in medical diagnostics and care | ✓ | ✓ | ✓ | ✓ | ✓ |

For each reference, the table summarises their focus areas and limitations, and indicates whether they discuss hardware, lightweight models, compression methods, computer vision applications, and medical applications

medical diagnostics in particular (Sect. 6). Finally, we highlight pivotal research challenges and future opportunities (Sect. 7) and summarise the main conclusions (Sect. 8). Building on existing literature, our survey highlights the important role of DL in advancing edge computing applications and aims to provide a comprehensive reference for researchers and practitioners.

## 2 Edge computing overview

Edge computing is an emerging computing paradigm that has attracted a great deal of attention in recent years (Chen and Ran 2019; Mendez et al. 2022). The proliferation of the Internet of Things (IoT), 5G, and other cutting-edge technologies has caused a rapid increase in data generation and consumption. In this setting, edge computing emerges as an effective strategy to handle the deluge of distributed data (Shi et al. 2016). Unlike conventional cloud computing, edge computing brings data processing closer to its origin, thus achieving reduced latency and increased processing efficiency (Cao et al. 2020). This approach is particularly relevant for latency-sensitive applications. For instance, autonomous vehicles require real-time data processing to make split-second decisions, ensuring safety and efficiency. Similarly, real-time patient monitoring systems in intensive care units rely on low-latency processing to provide immediate feedback and alerts. Furthermore, facilitating local processing on edge devices helps reduce data transmission and storage costs, simultaneously alleviating the computational burden on cloud infrastructures (Murshed et al. 2021). In this section, we provide a succinct overview of salient edge computing techniques and exemplars, underscoring the paramount advantages of edge computing, especially in the realm of medical imaging, where these advantages are particularly pronounced.

### 2.1 Edge computing related paradigms and terminologies

First we delve into various paradigms and terminologies relevant to edge computing, providing a foundational understanding of the architectural frameworks and methodologies that characterise this computing paradigm. This paves the way for a deeper exploration of model compression, computer vision, and medical diagnostics in subsequent sections.

**Cloudlets** (Babar et al. 2021) are small-scale cloud service hubs located at the edge of the internet, closer to end-users and devices. They serve as a bridge between large cloud data centres and edge devices through small-scale cloud servers, offering reduced latency, improved bandwidth efficiency, and localised computing environments. Cloudlets can be seen as "mini-clouds" that provide a subset of the services of cloud data centres, with the added advantage of geographical proximity to end-users.

**Fog computing** (Bonomi et al. 2012) extends the functionalities of cloud computing through a decentralised architecture that provides localised data processing, storage, and analysis at the network edge. Implemented through a distributed network of nodes, which may include routers, gateways, and other networking devices with computing and storage capabilities, it offers a more distributed approach to deliver a wider range of services. Fog computing addresses the bandwidth limitations of traditional cloud architectures and complements Cloudlets within the overall framework of edge computing.

**Mobile (multi-access) edge computing (MEC)** facilitates real-time data processing at the network edge, reducing latency and bandwidth use. By bringing computational resources closer to mobile devices, MEC enhances user experiences and supports applications that require real-time processing. A recent review article (Shahzadi et al. 2017) dives into the architecture and applications of MEC, offering insight into its role in evolving network infrastructures.

**Edge deep learning (Edge DL)** uses edge computing resources to perform deep learning inference and training tasks closer to data sources. This approach addresses the computa-

tional and latency challenges associated with centralised cloud-based deep learning frameworks. A recent survey (Wang et al. 2020a) provides a comprehensive overview of edge deep learning from a broad perspective, discussing the challenges and opportunities in this emerging field.

**Edge analytics** is an approach to data analysis performed on data at the source of generation, such as sensors and other edge devices, rather than sending the data back to a centralised data store. This reduces latency and bandwidth usage while enabling real-time insight. Various issues, challenges, opportunities, promises, and future directions of edge analytics have recently been discussed in a detailed review (Nayak et al. 2022).

**Cloud-edge collaboration** involves the integration of cloud computing and edge computing to take advantage of the strengths of both paradigms. This collaboration enables efficient data processing, analytics, and storage, optimises resource utilisation, and improves application performance. A recent comprehensive survey (Yao et al. 2022) delves into the various aspects of cloud-edge collaboration, including frameworks and architectures that facilitate seamless interaction between cloud and edge resources.

**Edge model training** facilitates the creation and enhancement of machine learning (ML) models directly on edge devices. Examples include EdgeMove (Dong et al. 2023), a scheme designed to accelerate model training between edge devices and servers by minimising communication overheads. Federated learning (FL), on the other hand, allows for model training on local data while preserving data privacy, albeit at the cost of computational and communication resources (Brecko et al. 2022; Abreha et al. 2022). Furthermore, distributed model training (DMT) optimises the distribution of training data among edge nodes, thus enhancing the efficiency of model training and the use of network resources (Hu et al. 2020).

**Edge model deployment** is crucial for real-time processing. IBM underscores the necessity of a model management system to efficiently handle various models in edge computing.[1] Similarly, Microsoft Azure extends ML inference from the cloud to on-premise or edge scenarios through Azure Stack Edge, ensuring seamless deployment and operation of models at the edge[2] AdaptiveNet (Wen et al. 2023) enables post-deployment neural architecture adaptation for diverse edge environments to ensure stable service quality. The simplification of DL model deployment at the edge is further promoted by NVIDIA's Triton Inference Server, which standardises the deployment process across various devices[3]

**Edge model inference** is the execution of trained ML models on edge devices to deduce insights from new data (Jiang et al. 2018). This is crucial for real-time or near-real-time analytics, especially in scenarios where low latency and reduced data transmission are vital. The effectiveness of this approach hinges on the ability to execute complex computational tasks efficiently within the limited resource confines of edge devices. Accordingly, to optimise the performance of model inference, strategies such as model compression (He et al. 2017b; Han et al. 2015; Matsubara et al. 2022; Chen et al. 2022a; Wang et al. 2020e) and adopting lightweight models (Mauri et al. 2022; Koonce and Koonce 2021; Wang and Huang 2021; Paluru et al. 2021) are frequently implemented.

[1] IBM: IBM Edge Computing Solutions. https://www.ibm.com/edge-computing.

[2] Microsoft: Azure Stack Edge. https://azure.microsoft.com/en-au/products/azure-stack/edge.

[3] NVIDIA: Simplifying AI Model Deployment at the Edge with NVIDIA Triton Infer- ence Server. https://developer.nvidia.com/blog/simplifying-ai-model-deployment-at-the-edge-with-triton-inference-server/.

### 2.2 Advantages of edge computing

Edge computing, characterised by its decentralised data processing, represents a significant change from traditional cloud-centric models (Shi et al. 2016). This paradigm is gaining prominence due to its potential impact across various domains, especially in environments with stringent requirements for real-time processing, data security, and bandwidth efficiency (Murshed et al. 2021). The adoption of edge computing can bring about substantial improvements in processing speeds, security, network utilisation, scalability, operational reliability, and cost efficiency (Cao et al. 2020; Mendez et al. 2022). These attributes are particularly relevant in sectors such as healthcare, where fast, secure and reliable data processing is paramount. The following survey delineates these advantages, underscoring their implications and applications in medical and related fields.

**Real-time data processing:** The primary advantage of edge computing is its ability to process data in real time at the data source. This is critical for applications that require swift responses. In healthcare, for example, immediate diagnostics and analysis in ambulances or intensive care units using edge devices can lead to interventions crucial for patient survival. Similarly, in medical imaging, processing images from MRI or CT scans directly on edge devices can accelerate the diagnostic process, potentially enabling quicker treatment planning. This rapid processing capability directly influences patient treatment outcomes, especially in time-sensitive scenarios (Sait et al. 2019; Liu et al. 2023b; Chen et al. 2019; Lingappa and Parvathy 2022).

**Data privacy and security:** In an era where data breaches are becoming more common, edge computing can offer a more secure alternative to traditional cloud computing. By processing and storing data locally, it helps reduce the exposure of data during network transmission. This is particularly important in industries with high data sensitivity, such as healthcare. Patient health information, medical histories, and diagnostic images contain highly sensitive information. Edge computing ensures that these data remain within the confines of the hospital's local network, significantly reducing breach risks and ensuring compliance with stringent health data regulations like the Health Insurance Portability and Accountability Act (HIPAA). Furthermore, local processing allows for quicker detection and response to potential security threats, enhancing the overall security framework. This local processing approach provides a reliable and robust framework for managing sensitive health data, safeguarding patient privacy and maintaining trust in healthcare systems (Cao et al. 2023; Singh and Chatterjee 2021; Alwakeel 2021).

**Bandwidth efficiency:** Edge computing can help mitigate network congestion by processing data locally, reducing the bandwidth required for data transmission. This efficiency is crucial in industries where large data files are common, such as healthcare. For example, transmitting high-resolution medical images, such as digital histopathology slides, can consume considerable bandwidth. With edge computing, these images are processed locally, eliminating the need for a constant, high-volume data transfer to the cloud. This not only conserves network bandwidth, but also ensures faster, more reliable medical imaging services, leading to more efficient healthcare delivery (Dong et al. 2020; Dave et al. 2021; Zheng et al. 2021).

**Scalability and flexibility:** Edge computing's scalability enables organisations to expand their computing capabilities as needed without significant infrastructure overhauls. This scalability is essential for industries experiencing rapid growth or facing fluctuating

demands. In healthcare, the ability to handle varying patient loads and integrate new medical technologies is crucial. The flexibility of edge computing is evident in scenarios like remote monitoring and care, supporting an increase in chronic disease treatment and enhancing healthcare access, especially in remote or rural areas. Processing data from IoT, e-health devices, and wearable technology at the edge allows healthcare providers to offer more effective continuous care and management of chronic diseases (Sun et al. 2020; Oueida et al. 2018).

**Reliability and accessibility:** Edge computing can provide more reliable data processing capabilities, especially in environments with limited or inconsistent internet connectivity. This reliability is crucial in many sectors, particularly in healthcare, where consistent access to patient data and medical applications can be lifesaving. For example, in remote or rural healthcare settings, reliable access to medical records and diagnostic tools is essential for patient care. Edge computing enables this by allowing local data processing and analysis, ensuring uninterrupted healthcare services despite connectivity issues. This local processing capability is particularly valuable in telemedicine and remote patient monitoring, where reliable data access is key to providing quality care to patients in remote locations (Srivastava et al. 2023; Ullah et al. 2021; Abdellatif et al. 2021).

**Cost-effectiveness:** Finally, edge computing offers cost-effective solutions by reducing the need for extensive data transmission and the dependency on centralised cloud storage. Particularly in the healthcare sector, this approach can save costs associated with managing and storing large amounts of medical data on cloud servers. Cloud services typically charge based on usage, including storage space and data transmission volumes. By keeping an appropriate amount of data storage and processing at the edge, hospitals and healthcare providers can establish more economical and efficient data processing schemes. In addition, economies of scale make the costs of deploying and maintaining edge computing hardware continuously decrease. Moreover, edge computing can be integrated with cloud computing to balance the initial investment and maintenance costs of local devices. As edge computing improves operational efficiency and response speed, healthcare professionals can gain faster access to critical information, make timely treatment decisions, and reduce treatment delays. In emergency situations, this rapid response capability can save lives and reduce long-term treatment costs. In summary, this localised data processing method can help optimise the use of IT resources in healthcare, making healthcare services more efficient and cost-effective (D'Agostino et al. 2019; Gu et al. 2022; Tang and Xin 2023; Mahenge et al. 2019; Jebadurai et al. 2021).

### 2.3 Disadvantages of edge computing

**Limited computational resources:** Compared to centralised cloud servers, edge devices often have limited resources, including processing power, memory, and storage space. These limitations can hinder the ability to run complex and computationally intensive deep learning models on edge devices (Varghese et al. 2020; Murshed et al. 2021). This issue is particularly significant in medical diagnostics, where the data involved is typically very large. For instance, medical imaging data such as MRI or CT scans often have extremely high resolution and detail, resulting in large data volumes. Moreover, due to the limited computational power, running large deep learning models for medical image analysis on edge devices may not be feasible. This limitation necessitates the use of data and model

compression techniques or lightweight models, which may compromise the accuracy and effectiveness of the diagnostic process (Ou et al. 2021; Goceri 2021b).

**Heterogeneity of hardware and software:** Edge computing devices come in a wide variety, with significant differences in performance and capabilities across different platforms, as well as varying requirements for software compatibility. This heterogeneity increases the complexity of development and maintenance, requiring specialised optimisation and adaptation strategies to ensure efficient operation across different devices. For example, while field-programmable gate arrays (FPGAs) offer flexibility and lower power consumption, their programming and optimisation are challenging. Although there are solutions to mitigate these issues, such as Open Neural Network Exchange (ONNX),[4] ExecuTorch,[5] and TensorFlow Lite,[6] which facilitate model conversion and optimisation across different hardware platforms, these solutions still face challenges in medical scenarios. Specifically, medical image data not only have high resolution but also vary in type, and devices like FPGAs and application-specific integrated circuits (ASICs) may encounter software compatibility and driver issues when processing these images, further increasing the difficulty and cost of application development and implementation (Solanki et al. 2021; Gtifa and Sakly 2023; Tabassum et al. 2023).

**Standardization issues:** Due to hardware differences and lack of uniform standards among edge devices, inconsistent computational results may be produced, affecting the overall system's accuracy and reliability (Feng et al. 2022). In medical environments, ensuring the standardisation of edge device hardware and software protocols is crucial to guarantee the consistency and accuracy of diagnostic results, as well as to comply with regulatory requirements. Standardising edge devices not only helps improve system integration efficiency but also ensures the accuracy and reliability of diagnostic outcomes, thereby enhancing the overall quality of medical diagnostics. Additionally, standardisation can reduce development and maintenance costs, enabling healthcare institutions to deploy and manage edge computing devices more effectively. It also simplifies compliance with medical data regulations like HIPAA and General Data Protection Regulation (GDPR), further improving the efficiency and quality of medical services.

## 3 Edge computing devices

Edge computing devices serve as the core components of edge computing frameworks and play a crucial role in their implementation. The performance and diverse functionality of these devices directly impact the feasibility and efficiency of edge computing solutions. Different edge computing scenarios require various types of edge devices, ranging from simple microcontrollers to computing platforms capable of processing large deep learning models. These devices together form a complex and feature-rich technological ecosystem. Therefore, a thorough understanding of the classification, specifications, application scenarios, and cost-effectiveness is essential for the design and implementation of edge computing solutions in medical diagnostics.

[4] ONNX. https://onnx.ai/.

[5] ExecuTorch. https://pytorch.org/edge.

[6] TensorFlow Lite. https://www.tensorflow.org/lite.

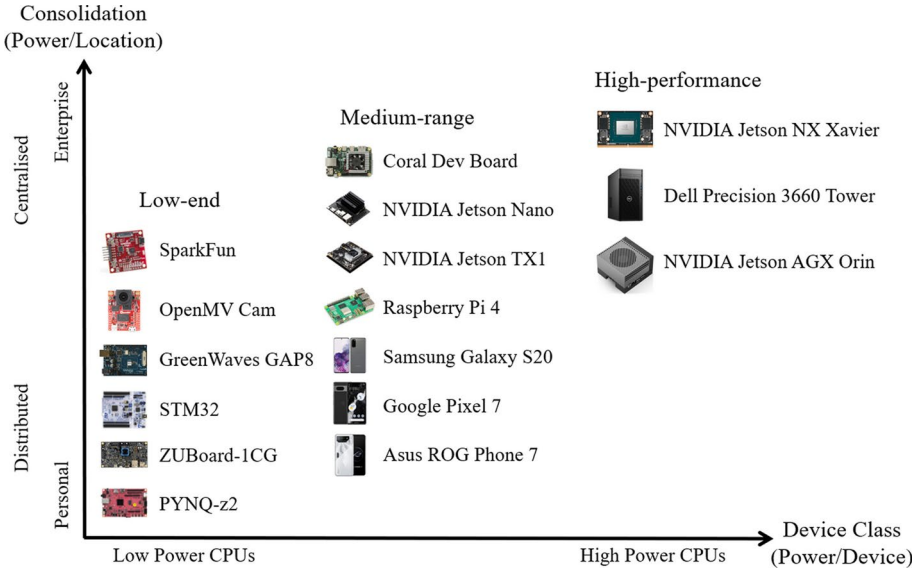


**Fig. 3** Three categories of edge computing devices

In this section, we present a classification of edge computing devices based on their computational capabilities, dividing them into low-end, mid-range, and high-performance devices. Each category of devices is designed for specific application scenarios, from executing basic data preprocessing tasks to supporting complex deep learning model training. Subsequently, we introduce some typical devices and discuss their performance and purposes. We emphasise the importance of edge device diversity and demonstrate how to fully leverage the potential of edge computing by selecting the appropriate devices.

### 3.1 Edge computing device categories

We propose to categorise edge computing devices according to their computational capabilities into low-end, medium-range, and high-performance devices (Fig. 3) and briefly discuss each category.

**Low-end edge devices** are characterised by their limited computational resources and energy efficiency, making them suitable for lightweight applications (Dutta and Bharali 2021). These devices typically lack dedicated multiple graphics processing units (GPUs) or advanced central processing units (CPUs). Due to their computational constraints, they are more suited for inference tasks than for training DL models. These devices usually have power efficient processors, such as the ARM Cortex M series[7] or the lower-end Cortex A series,[8] and give priority to minimising power usage and achieving small form factors. Memory capacities are quite constrained, often ranging from a few hundred kilobytes (kB) to a few megabytes (MB) and may not include advanced accelerators. Therefore, these devices are highly suitable for performing basic edge computing tasks in environments with

[7] ACM: Cortex M Series. https://www.arm.com/products/silicon-ip-cpu/cortex-m/cortex-m4.

[8] ACM: Cortex A Series. https://www.arm.com/products/silicon-ip-cpu/cortex-a/cortex-a720.

limited resources. Sensor data preprocessing, fundamental DL inference, and low-complexity IoT applications are a few examples (Nunez-Yanez and Howard 2021; Shen et al. 2022; Dutta and Bharali 2021). Due to their energy efficiency and affordability, these devices are also well suited for projects that have limited financial resources or require installation in remote locations with a lack of power supply (Ray 2022; Schizas et al. 2022). This makes them particularly valuable in lightweight medical diagnostic applications. For example, in wearable health monitoring systems, low-end devices can effectively handle sensor data preprocessing, enabling real-time health tracking in remote or resource-constrained environments. Their high energy efficiency and affordability make them ideal for tasks such as basic medical data acquisition and simple inference.

**Medium-range edge devices** offer a balanced level of computational power, making them capable of handling more advanced tasks compared to low-end devices. They often come with multicore processors and some level of GPU acceleration, which allows them to perform more computationally intensive tasks Jolles (2021), Zhao et al. (2015), Mao et al. (2017). Although they may lack the substantial computational resources required to train DL models, their enhanced processing capabilities make them well suited for more advanced inference tasks. Unlike low-end devices, these devices typically have memory capacities ranging from one to several gigabytes (GB). For performance enhancement, certain devices in this category incorporate specialised AI accelerators (Sipola et al. 2022). These devices exhibit versatility and are well suited for a wide array of AI applications, including image recognition (Chavan et al. 2021; Sati et al. 2021; Kang et al. 2018), object detection (Nazeer et al. 2022; Mittapalli et al. 2023; Kaymak and Aysegul 2018), and the creation of compact DL models (Momin et al. 2023; Heo et al. 2020; Gonzalez-Huitron et al. 2021) for use in robotics (Alexey et al. 2021; Tang et al. 2017; Mathe et al. 2022), smart cameras (Kyrkou 2020; Karaman et al. 2021), and other edge computing scenarios requiring a moderate amount of computational capability (Mittal 2019; Miori et al. 2017; Lertsinsrubtavee et al. 2017; Zhao et al. 2018). In the medical field, medium-range devices are particularly well-suited for deploying AI models that perform image-based diagnostics, such as automated pathology screening. Devices such as the NVIDIA Jetson Nano, Raspberry Pi, and smartphones have already been utilised in portable medical imaging devices to execute pretrained diagnostic models in real-time, striking a good balance between performance and power consumption. (Wang et al. 2019b; Raghavan et al. 2020; Goceri 2021b; Paluru et al. 2021; Shrivastava et al. 2023)

**High-performance edge devices** are designed to handle complex computational tasks, including DL model training. They feature robust architectures with powerful CPUs, GPUs, and significant memory resources. Unlike low-end and medium-range devices, high-performance edge devices can handle training and inference tasks effectively, making them suitable for scenarios that require high computational power. The memory capacities of these devices are substantial, ranging from several GB to terabytes (TB), allowing them to efficiently process intricate models and datasets. These devices provide exceptional performance in applications that require extensive AI tasks, such as training DL models, handling sophisticated computer vision tasks, and executing advanced AI applications (Kirillov et al. 2023; Ma et al. 2024; Ho et al. 2020). They are suitable for use in sophisticated robots (Krupnik et al. 2023; Wang et al. 2023; Makoviychuk et al. 2021) and data centres where top-notch performance is crucial (Angus et al. 2022; Oro et al. 2011; Barisoni et al. 2020; Shen et al. 2019). These devices can handle large medical images (such as MRI or

CT scans) and perform in-depth image analysis in real-time. Enabling local training of deep learning models, they can leverage test-time domain adaptation and continual learning to accommodate the ever-growing medical data, ensuring higher diagnostic accuracy and reliability. This is crucial in handling complex medical scenarios, especially when faced with constantly evolving patient data or environments.

This discussion underscores the fact that medium-range and low-end edge devices are typically more constrained in their training capabilities because of limited computational resources. However, their compact size and portability make them suitable for deployment in space-constrained (Dutta and Bharali 2021; Xie et al. 2018; Lachhab 2023; Achakir et al. 2023) or mobile scenarios (Shahzadi et al. 2017; Abbas et al. 2017; Mach and Becvar 2017). Moreover, although they may not be able to effectively handle training of DL models, they are generally capable of achieving (near) real-time inference, a critical aspect for certain computer vision applications (Nazeer et al. 2022; Tang et al. 2017; Ray 2022). The ability to process data locally also enhances privacy and security by minimising data transmission, thus reducing the risk of data leakage or tampering (Cao et al. 2023; Singh and Chatterjee 2021; Alwakeel 2021; Ranaweera et al. 2021; Zhang et al. 2018; Ali et al. 2021).

### 3.2 Detailed overview of edge computing devices

Following the above categorisation, we now present concrete examples of edge computing devices across the spectrum of computational capabilities, from low-end devices to high-performance devices (Table 2). We summarise the fundamental hardware specifications of these devices, including the CPU, accelerator, and memory attributes, providing a clearer perspective on their computational capacities and potential use cases.

**Low-end edge computing devices** such as the SparkFun Edge[9], STM32 microcontroller series,[10] XUP PYNQ-Z2 board,[11] and Intel Neural Compute Stick 2 (NCS2)[12] have shown significant multifunctionality and applicability. The STM32 series of microcontrollers is widely used in embedded systems, especially in the TinyML (tiny machine learning) application domain (Cum 2022; Schizas et al. 2022; Dutta and Bharali 2021). The SparkFun Edge board optimised for ultra-low power consumption is ideal for edge DL applications where efficiency is paramount. Built around the ARM Cortex-M4F[13] and integrated with TensorFlow Lite for microcontrollers, it supports DL tasks directly on the device. Its on-board sensors and bluetooth connectivity make it perfect for IoT and wearable devices, enabling real-time data processing and communication in compact, power-sensitive projects.

The STM32 X-CUBE-AI extension package[14] further expands the capabilities of STM32, allowing users to easily deploy trained neural networks to microcontrollers. This feature makes the STM32 an ideal choice for tasks such as image recognition and video analysis. In industrial automation and smart home systems, STM32 combined with deep learning algorithms can enhance efficiency and accuracy, optimising monitoring and control processes.

[9] SparkFun Edge. https://github.com/sparkfun/SparkFun_Edge.

[10] ST: STM32. https://www.st.com/en/microcontrollers-microprocessors/stm32-arm-cortex-mpus.html.

[11] AMD: XUP PYNQ-Z2. https://www.xilinx.com/support/university/xup-boards/XUPPYNQ-Z2.html.

[12] Intel: Neural Compute Stick 2, https://www.intel.com/content/www/us/en/developer/articles/tool/neural-compute-stick.html.

[13] ARM: Cortex-M4. https://developer.arm.com/Processors/Cortex-M4.

[14] ST: AI Expansion Pack for STM32Cube. https://www.st.com/en/embedded-software/x-cube-ai.html.

STM32 also supports various communication protocols, such as SPI, I2C, UART, and CAN, making it an ideal choice for integrating various peripherals and modules. Additionally, STM32 microcontrollers support multiple programming environments, including Eclipse-based STM32CubeIDE and Arduino, offering significant flexibility and ease of use.

Leveraging the STM32H743VI microcontroller with a Cortex-M7 processor, the OpenMV Cam H7 R2[15] is a pivotal tool in edge-based machine vision applications. Its design facilitates real-time image processing, enabling tasks such as face recognition and object tracking. The device integrates a camera module for direct image capture and supports multiple peripherals for expanded functionality. Compatible with the OpenMV IDE, it offers a Python-based development environment, which streamlines the deployment of complex vision algorithms.

Another commonly used edge device is the AUP PYNQ-Z2 board.[16] PYNQ-Z2, based on the Xilinx Zynq SoC,[17] combines the powerful functionality of an ARM processor with the flexibility of a field-programmable gate array (FPGA). It provides significant hardware acceleration for advanced image processing and complex signal processing. A notable feature of PYNQ-Z2 is its support for AMD Vitis AI,[18] designed to accelerate AI inference on FPGAs and other programmable logic devices. This allows optimising and deploying trained DL models to the PYNQ-Z2 board, enabling efficient edge AI applications. It makes PYNQ-Z2 particularly advantageous in AI projects that require custom hardware acceleration, such as real-time video analysis and intelligent sensor applications. Additionally, PYNQ-Z2 supports Python programming and the open source PYNQ framework, greatly simplifying the development process and making it an ideal choice for researchers and developers for rapid prototyping.

The Intel Neural Compute Stick 2 (NCS2) is focused on enhancing the DL inference capabilities of edge devices. Equipped with the Movidius Myriad X Vision Processing Unit (VPU),[19] it accelerates the execution of neural network models without significant increases in power consumption, which is crucial for applications requiring real-time image and video analysis, such as intelligent surveillance and automated detection systems.

Furthermore, the Eta Compute ECM3532[20] combines an ARM Cortex M3 processor, optimising energy efficiency and computational performance in edge AI applications. This dual-core approach enables the device to support advanced deep learning tasks, within the stringent power constraints typical of wearable and IoT devices. GreenWaves GAP8[21] stands out for its ultra-low power consumption and ability to perform embedded DL tasks efficiently on the edge. Utilising an 8-core computational cluster with a hardware accelerator, it is optimised for processing image and audio algorithms, including CNN inference, with exceptional energy efficiency. The BeagleBone AI,[22] designed around the Texas Instru-

[15] OpenMV Cam H7 R2. https://openmv.io/products/openmv-cam-h7-r2.

[16] AUP PYNQ-Z2. https://www.amd.com/en/corporate/university-program/aup-boards/pynq-z2.html.

[17] Xilinx Zynq SoC. https://www.xilinx.com/products/silicon-devices/soc/zynq-7000.html.

[18] Vitis AI. https://www.xilinx.com/products/design-tools/vitis/vitis-ai.html.

[19] Movidius Myriad X VPU. https://www.intel.com/content/www/us/en/products/details/processors/movidius-vpu/movidius-myriad-x/products.html.

[20] Eta Compute ECM3532. https://media.digikey.com/pdf/Data%20Sheets/Eta%20Compute%20PDFs/ECM3532_AI_Sensor_PB_1.0.pdf.

[21] GreenWaves GAP8. https://greenwaves-technologies.com/low-power-processor.

[22] BeagleBone AI. https://www.beagleboard.org/boards/beaglebone-ai.

**Table 2** Examples of edge computing devices categorised by computational capabilities (Prices are approximate and may vary by region and retailer)

| Type | Devices | Operating system | Framework | Core language | CPU | Accelerator | Memory | Price (USD) |
|---|---|---|---|---|---|---|---|---|
| Low-End | SparkFun Apollo3 Blue | – | TensorFlow | Python | ARM Cortex-M4F | – | 384 KB | ∼ $20 |
| | OpenMV Cam H7 R2 | MicroPython | – | MicroPython | ARM Cortex-M7 | – | 1 MB | ∼ $80 |
| | Eta Compute ECM3532 | – | – | C/C++ | ARM Cortex-M3 | – | 256 KB | ∼ $60 |
| | GreenWaves GAP8 | FreeRTOS, PULP OS, PMSIS | TensorFlow, Pytorch, Keras, ONNX | Python, C/C++ | Nona-Core RISC-V | – | 16 MB | ∼$100 |
| | BeagleBone AI | Linux | TensorFlow, Pytorch, Caffe, ONNX | Python, C/C++ | Dual ARM Cortex-A15 + Dual ARM Cortex-M4 | – | 1 GB | ∼$300 |
| | STM32MP157F-DK2 | Linux | ONNX | Python, C/C++ | Dual Cortex-A7 + ARM Cortex-M4 | – | 4 GB | ∼$200 |
| | PYNQ-Z2 | Linux | TensorFlow, Pytorch, ONNX | Python | Dual ARM Cortex-A9 | – | 512 MB | ∼$120 |
| | ZUBoard-1CG | Linux | TensorFlow, Pytorch, ONNX | Python | ARM Cortex-A53 + ARM Cortex-R5F | – | 1 GB | ∼$150 |

**Table 2** (continued)

| Type | Devices | Operating system | Framework | Core language | CPU | Accelerator | Memory | Price (USD) |
|---|---|---|---|---|---|---|---|---|
| Medium-Range | Google Coral Dev Board | Linux | TensorFlow | Python, C++ | Quad ARM Cortex-A53 + ARM Cortex-M4F | GC7000 Lite Graphics + Google Edge TPU | 1 GB/4 GB | ∼$130 |
| | NVIDIA Jetson Nano | Linux | TensorFlow, Pytorch, Keras, JAX, PaddlePaddle, MXNet, Caffe, ONNX | Python | Quad ARM Cortex-A57 | NVIDIA Maxwell 128 CUDA Cores | 4 GB | ∼$150 |
| | NVIDIA Jetson TX1 | Linux | TensorFlow, Pytorch, Keras, JAX, PaddlePaddle, MXNet, Caffe, ONNX | Python | Quad ARM Cortex-A57 | NVIDIA Maxwell 256 CUDA Cores | 4 GB | ∼$200 |
| | Raspberry Pi 5 | Linux | TensorFlow, Pytorch, Caffe, Keras | Python, C++ | Quad ARM Cortex-A76 | Broadcom VideoCore VII | 2 GB/4 GB /8 GB | ∼ $80 |
| | NVIDIA Jetson TX2 | Linux | TensorFlow, Pytorch, Keras, JAX, PaddlePaddle, MXNet, Caffe, ONNX | Python | Quad ARM Cortex-A57 + NVIDIA Dual Denver 2 | NVIDIA Pascal 256 CUDA Cores | 8 GB | ∼$250 |
| | Samsung Galaxy S20 | Android | TensorFlow, Pytorch, Caffe, Keras, ONNX | Java, Python | Qualcomm SM8250 | Qualcomm Adreno 650 | 8 GB | ∼$250 |
| | Google Pixel 7 | Android | TensorFlow, Pytorch, Caffe, Keras, ONNX | Java, Kotlin, Python | Google Tensor G2 | Mali-G710 MP7 | 8 GB | ∼$350 |
| | Asus ROG Phone 7 | Android | TensorFlow, Pytorch, Caffe, Keras, ONNX | Java Python | Qualcomm SM8550 | Qualcomm Adreno 740 | 16 GB | ∼$800 |

**Table 2** (continued)

| Type | Devices | Operating system | Framework | Core language | CPU | Accelerator | Memory | Price (USD) |
|---|---|---|---|---|---|---|---|---|
| High-Performance | Jetson NX Xavier | Linux | TensorFlow, Pytorch, Keras, JAX, PaddlePaddle, MXNet, Caffe, ONNX | Python | 6-Core NVIDIA Carmel | NVIDIA Volta 384 CUDA Cores | 16 GB | ~$700 |
| | Jetson AGX Orin | Linux | TensorFlow, Pytorch, Keras, JAX, PaddlePaddle, MXNet, Caffe, ONNX | Python | 12-Core Arm Cortex-A78AE | NVIDIA Ampere 2048 CUDA Cores | 64 GB | ~$1000 |
| | Dell Precision 3660 Tower | Linux, Windows | TensorFlow, Pytorch, Keras, JAX, PaddlePaddle, MXNet, Caffe, ONNX | Python | Intel Core i9-13900K | NVIDIA RTX A4000 | 32 GB | ~$3000 |

ments AM5729 Sitara processor,[23] offers a platform for developers to explore AI integration in edge computing. It features a dual ARM Cortex-A15 processor,[24] supported by Embedded Vision Engines (EVEs) for DL, and extensive connectivity options including Gigabit Ethernet and WiFi. The ZUBoard 1CG,[25] serves as a versatile platform for edge computing applications such as embedded vision. Moreover, it offers high-speed storage and wireless connectivity options, satisfying the demanding requirements of industrial, healthcare, and multimedia applications.

**Medium-range edge computing devices** include Google Coral Dev Board, a compact, powerful platform designed for Edge DL, which integrates Google's Edge Tensor Processing Unit (TPU) coprocessor[26] and is capable of performing fast DL inferencing on small form factor devices. This makes it ideal for prototyping AI products and solutions that require processing efficiency. The board supports TensorFlow Lite models, facilitating the development of AI applications such as healthcare, retail, and the smart home, by enabling local real-time processing of deep learning workloads (Imran et al. 2020; Winzig et al. 2022).

The NVIDIA Jetson series,[27] including Jetson Nano, TX1, and TX2, are key devices designed for different levels of edge computing needs. Jetson Nano, the base model in the series, is mainly suitable for lightweight machine vision and data processing tasks, such as simple object recognition and video stream processing. Its low power consumption and compact design make it an ideal choice for deployment in resource-constrained environments. Jetson TX1 and TX2 offer more powerful computing capabilities, suitable for applications that require higher image processing capacity, such as in augmented reality (AR) and virtual reality (VR), autonomous vehicle perception systems, drone navigation systems, chest CT, and dermatology detection (Mittal 2019; Cass 2020; Wang et al. 2019b; Shrivastava et al. 2023; Abubeker and Baskar 2023).

The Raspberry Pi series,[28] a widely used single-board computer, is preferred for its cost-effectiveness, low power consumption, and robust community support. Despite its limited processing capabilities, it is suitable for simple applications, such as sensor data processing and lightweight image recognition tasks (Zhao et al. 2015; Jolles 2021; Kaymak and Aysegul 2018; Paluru et al. 2021).

Additionally, smartphones represent a typical example of mid-range edge devices, equipped with processors capable of efficiently running DL models for various tasks including image processing and video analysis. Integrated with dedicated GPUs, digital signal processors (DSPs), and DL accelerators such as neural processing units (NPUs), they offer excellent task processing capabilities and efficient energy management. The Qualcomm SM8250 and SM8550[29] are top-notch smartphone chipsets that can provide devices with advanced inference performance for DL models. In addition, they support 5G connectivity,

[23] Texas Instruments AM5729. https://www.ti.com/product/AM5729.

[24] ARM: Cortex-A15. https://developer.arm.com/Processors/Cortex-A15.

[25] ZUBoard 1CG. https://www.avnet.com/wps/portal/us/products/avnet-boards/avnet-board-families/zuboard-1cg.

[26] TensorFlow models on the Edge TPU. https://coral.ai/docs/edgetpu/models-intro.

[27] NVIDIA Jetson Modules. https://developer.nvidia.com/embedded/jetson-modules.

[28] Raspberry Pi. https://www.raspberrypi.com.

[29] Snapdragon 865 5G Mobile Platform. https://www.qualcomm.com/products/mobile/snapdragon/smartphones/snapdragon-8-series-mobile-platforms/snapdragon-865-5g-mobile-platform.

enabling faster internet speeds and improved network performance. This facilitates effective support for distributed edge training and inference paradigms, such as federated learning.

Quectel's SC66 smart module[30] is a multi-functional and widely applicable edge module. Equipped with Qualcomm's SDM660 chipset[31] and the Snapdragon Neural Processing Engine (SNP),[32] it is designed for high data rate, multimedia capabilities, and advanced deep learning-based use cases. SC66 supports various features, such as quick charge technology, making it highly suitable for industrial and consumer-grade applications. Combining high-speed wireless connectivity and an embedded Global Navigation Satellite System (GNSS) receiver, it is capable of serving a wide range of edge applications (Lestariningati 2018; Sadique 2013; Mwansa et al. 2022; Cum 2022).

Furthermore, the Apple A16 Bionic chip,[33] introduced in the iPhone 14 Pro models, is built on 4 nm process technology, offering improved performance and efficiency compared to its predecessors. It features a 6-core CPU with two high-performance and four efficiency cores, a 5-core GPU for improved graphics, and a 16-core neural engine for advanced DL tasks, doubling the DL capabilities to 17 trillion operations per second. This chip significantly boosts performance for real-time DL-based image and video processing, while optimising power consumption for extended battery life.

**High performance edge computing devices** include the NVIDIA Jetson Xavier series[34] and Jetson Orin series.[35] The Jetson Xavier and Orin series uses the NVIDIA Volta and Ampere GPU architecture, featuring CUDA (compute unified device architecture) cores and Tensor cores, as well as larger memory capacity, all designed specifically for DL acceleration. These devices are suited for running more complex DL models and allow model fine-tuning during deployment (Bhardwaj et al. 2022; Kortli et al. 2022). A performance comparison of the YOLOv3 model on NVIDIA Jetson Xavier NX, NVIDIA Jetson Nano, and Raspberry Pi 4 + NCS2 (Table 3) highlights the advantages of high-performance devices like the Jetson Xavier NX in terms of FPS and inference time, while also showcasing the

**Table 3** Performance comparison of the YOLOv3 model on Raspberry Pi 4 + NCS2, Jetson Nano, and Jetson Xavier NX

| Model | Devices | Mean confidence (%) | FPS | CPU Usage (%) | Memory usage (GB) | Energy consumption (W) | Inference time (s) |
|---|---|---|---|---|---|---|---|
| YOLOv3 | Raspberry Pi 4 + NCS2 | 99.3 | 2.5 | 4.3 | 0.33 | 6.0 | 690 |
| | Jetson Nano | 99.7 | 1.7 | 26.5 | 1.21 | 7.9 | 967 |
| | Jetson Xavier NX | 99.7 | 6.1 | 22.5 | 1.51 | 15.2 | 256 |

The tests were conducted on video data consisting of 1596 frames with a frame size of $768 \times 436$ (Feng et al. 2022)

[30] Quectel LTE SC66 series. https://www.quectel.com/product/lte-sc66-smart-module-series.

[31] Qualcomm-SDM660.https://www.qualcomm.com/products/technology/processors/application-processors/sdm660.

[32] SNP Engine. https://developer.qualcomm.com/sites/default/files/docs/snpe/overview.html.

[33] Apple debuts iPhone 14 Pro and iPhone 14 Pro Max. https://www.apple.com/au/newsroom/2022/09/apple-debuts-iphone-14-pro-and-iphone-14-pro-max/.

[34] Jetson Xavier series. https://www.nvidia.com/en-au/autonomous-machines/embedded-systems/jetson-xavier-series/.

[35] Jetson Orin series. https://www.nvidia.com/en-au/autonomous-machines/embedded-systems/jetson-orin/.

balance of performance and energy efficiency offered by mid-range devices like the Jetson Nano.

The Dell Precision 3660 Tower Dell (Accessed 28 January 2024b) exemplifies edge devices of workstation type, which typically offer exceptional performance and high configurability. It typically offer exceptional performance and high configurability.Equipped with the latest generation Intel Xeon processors[36] and NVIDIA Quadro GPUs,[37] it can meet the training and deployment needs of resource intensive DL applications. Additionally, it provides flexible expansion options and comprehensive security features, including the Trusted Platform Module (TPM) 2.0 and hardware security modules,[38] to protect sensitive data. Furthermore, some workstations support multi-GPU configurations. These workstations are specifically designed to handle highly parallel computing tasks, such as large-scale DL model training. Integrating multiple high-performance GPUs, such as NVIDIA's Tesla[39] or Quadro series, these workstations can significantly accelerate the training and inference of the DL model. Workstations that support multiple GPUs also offer high-bandwidth memory configurations and high-speed data transfer interfaces to ensure efficient data transfer between GPUs, maximising parallel computing efficiency. Furthermore, these workstations are usually equipped with advanced cooling systems and power management features to ensure stability and reliability during extended periods of operation.

This section categorises edge computing devices from low-end to high-performance, emphasising their computational capabilities and hardware specifications to highlight their applicability across various applications. With advancements in technology, not only traditional embedded systems and specialised computing platforms but also an increasing number of smartphones, laptops, automobiles and even smart home devices have been optimised for deep learning, extending their use into the domain of edge computing (Murshed et al. 2021). This diversity ensures that users can select the most appropriate technology based on their specific computational needs and application requirements, effectively leveraging the power of edge computing in diverse environments.

## 4 Transferring large DNNs to lightweight DNNs

The need to implement DNNs on devices with limited resources has led to the creation of lightweight neural network architectures (Khan et al. 2022b, 2024; Farooq et al. 2024). To achieve an optimal balance between computational efficiency and model complexity, these architectures are well-suited for edge devices. Compared to their larger counterparts, lightweight neural networks provide a number of benefits, including a smaller memory footprint, accelerated inference speed, and reduced energy consumption (Khan et al. 2022c; Javed et al. 2024; Naqvi et al. 2023; Khan et al. 2022a). We examine the main methods (Fig. 4) to obtain lightweight neural networks suited for implementation in edge devices. These methods include network pruning, sparse representation, quantised neural networks, efficient lightweight neural networks, and neural architecture search (see Table 4 for a comparison

[36] Intel Xeon https://www.intel.com/content/www/us/en/products/details/processors/xeon.html.

[37] NVIDIA Quadro GPUs https://www.nvidia.com/en-us/design-visualization/quadro/.

[38] TMP https://support.microsoft.com/en-us/topic/what-is-tpm-705f241d-025d-4470-80c5-4feeb24fa1ee.

[39] NVIDIA Tesla GPUs https://www.nvidia.com/en-gb/data-center/tesla-v100/.

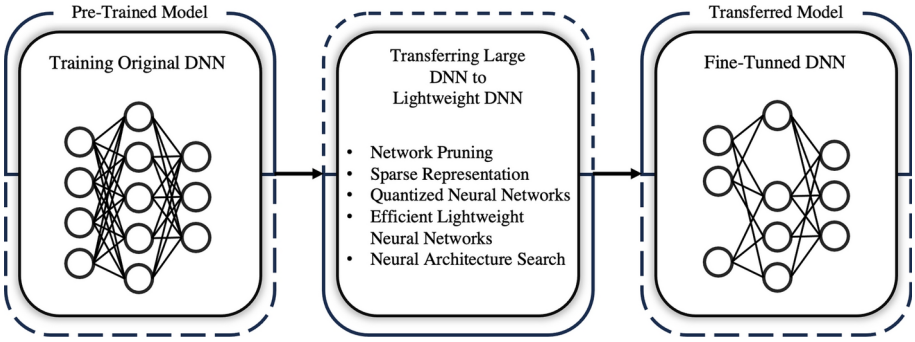


**Fig. 4** Methods to convert large DNNs to Lightweight DNNs

**Table 4** Comparison of methods for converting large DNNs to lightweight DNNs

| Method | Advantages | Disadvantages | Applications |
|---|---|---|---|
| Network pruning | Reduces model size and complexity; maintains accuracy; suitable for post-training optimisation | Can be computationally intensive; requires retraining; pruned networks might need specialised hardware for efficient inference | Image classification, object detection, and scenarios with over-parameterized models |
| Sparse representation | Reduces memory and computational requirements; improves generalisation; makes models more interpretable | Determining optimal sparsity level is challenging; may require specialised algorithms for training and inference | Applications where interpretability and efficiency are crucial, such as medical imaging and real-time analytics |
| Quantised neural networks | Significant reduction in memory and computation; faster inference; lower energy consumption | Potential loss of accuracy; requires quantisation-aware training; bit-width selection is nontrivial | Edge devices, mobile applications, and real-time systems requiring fast and efficient inference |
| Efficient lightweight neural networks | Designed for efficiency from the ground up; high accuracy with low computational cost; no need for post-processing optimisation | May not perform as well as larger networks on very complex tasks; design is task specific | Mobile and embedded devices, IoT applications, and scenarios with stringent resource constraints |
| Neural architecture search (NAS) | Automates the design of optimal architectures; can achieve state-of-the-art performance; considers multiple objectives (accuracy, latency, etc.) | Computationally expensive; may require significant resources and time; complexity of search space can be high | Diverse applications including image recognition, language processing, and any task needing tailored architectures |

of these methods, highlighting their respective advantages, disadvantages, and potential applications).

## 4.1 Network pruning

Neural networks enable systems to autonomously discern patterns from data (LeCun et al. 2015). However, achieving high performance in DNN models often entails a trade-off, as it leads to an increased number of neurons and synaptic connections, which in turn significantly increases the computational and spatial complexity of neural network models.

Various studies (Han et al. 2015; He et al. 2017b, 2018; Hu et al. 2016) reveal a substantial presence of redundant neurons and connections within numerous DNN architectures. Identification and removal of these redundant elements have been shown to significantly decrease both computational demands and model size while maintaining accuracy. Addressing this challenge, pruning is a technique designed to automatically pinpoint redundant neurons and connections, facilitating model compression and acceleration. We discuss four different types of pruning often used for model compression, namely weight pruning, neuron pruning, channel pruning, and filter pruning.

### 4.1.1 Weight pruning

Weight pruning in CNNs is a technique to reduce the size of the model by eliminating certain weights (parameters) while maintaining performance to some extent. In a neural network, weights represent the strength of connections between neurons. Pruning involves identifying and removing some of these weights, effectively setting them to zero, and then retraining the network to recover performance. Unstructured pruning (Han et al. 2015) aims to reduce the complexity and energy consumption of large DNNs on edge devices, specifically during the inference phase. This technique achieves reductions of $9\times$ and $13\times$ in complexity in AlexNet (Krizhevsky et al. 2012) and VGG16 (Simonyan and Zisserman 2014), respectively. It has three main stages. First, a network is trained to acquire knowledge about the connection using conventional training techniques (Bottou 2010). Afterwards, a pre-established threshold is used to remove connections with weights below the threshold. Finally, the weights acquired during the initial training process are utilised to establish the starting weights of the pruned neural network. Pruning leads to the weight matrix becoming sparse as numerous weights are eliminated. A compressed sparse row or column structure is utilised to effectively store this sparse matrix.

### 4.1.2 Neuron pruning

Neuron pruning goes beyond weight pruning by not only removing individual weights but entire neurons (or filters) along with their associated weights (Fig. 5). Neurons in CNNs correspond to feature maps or filters that capture specific patterns or features in the input data. Similar to weight pruning, neuron pruning is typically performed iteratively. After pruning,

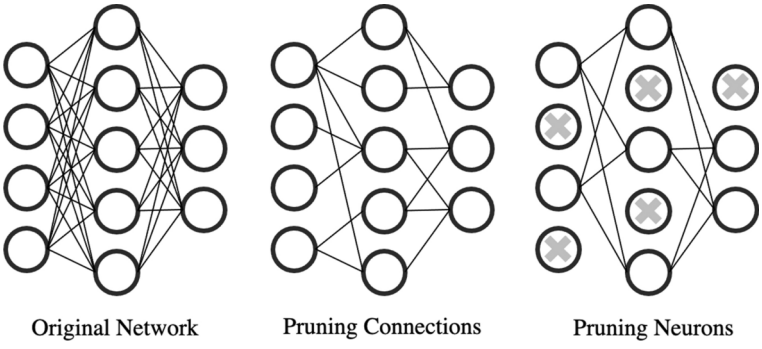


**Fig. 5** Difference between pruning connections and pruning neurons

the model needs to be fine-tuned to recover any performance loss caused by the removal of neurons. Neuron pruning uses certain criteria to determine which neurons are less important or redundant and can be pruned from the network. To address problems with unstructured pruning (Han et al. 2015) and improve the computational and space efficiency of DNN models, a structured pruning technique can be used (Hu et al. 2016). Contrary to weight-based pruning, this technique specifically targets neurons that exhibit a high frequency of zero activations after the ReLU (rectified linear unit) layer. The technique is based on the recognition of significant duplication of neurons across DNN models. Redundancy in neural networks not only leads to higher computational and space requirements but also worsens overfitting. The average percentage of zeros (APoZ) metric can be used to measure the proportion of zero activations in a neuron following the ReLU layer.

### 4.1.3 Channel pruning

Weight pruning can decrease the space complexity of a neural network model. However, implementing the resulting sparse structure on hardware platforms might be difficult, and accelerating the inference process usually requires the use of specialised hardware. To overcome this, channel pruning (He et al. 2017b) can be used as an alternate technique for structured pruning specifically for convolutional layers (Han et al. 2015). Channel pruning is distinct from weight pruning, as it introduces structured sparsity by directly reducing the number of channels in a feature map rather than inducing unstructured sparsity in convolutional layers. Structured sparsity allows for efficient implementation on both CPU and GPU platforms. Inference-based channel pruning (He et al. 2017b) aims to minimise the reconstruction error of a layer's feature map. The optimisation process comprises two primary stages: channel selection and feature map rebuilding. In the first stage, the most representative channels are selected, and redundant channels are eliminated to decrease the complexity of the model. To address pruning mistakes, a linear least-squares approach is used to reconstruct the output feature map using the pruned feature map data. This bipartite procedure guarantees that the pruned model maintains its prognostic effectiveness while reducing computational and memory demands.

Channel pruning enhances the scalability of neural networks by reducing the number of channels in each layer, leading to a more compact model that can be efficiently deployed across various hardware platforms. He et al. (2017b) demonstrated that channel pruning could reduce the number of parameters in ResNet-50 by 50% while maintaining 99% of its original accuracy on ImageNet. This reduction in complexity makes the pruned model highly scalable and suitable for deployment on both CPUs and GPUs.

The structured sparsity introduced by channel pruning makes it flexible for implementation on different types of hardware without requiring specialized inference engines. Li et al. (2020) applied channel pruning to MobileNetV2, achieving a $2\times$ reduction in the number of FLOPs while maintaining a top-1 accuracy of 70.6% on ImageNet. This flexibility allows the pruned model to be used in a wide range of applications, from mobile devices to high-performance servers.

Recent advancements in channel pruning have focused on optimizing models for edge devices by further reducing computational complexity and memory usage. He et al. (2020) proposed a learning-based channel pruning method that achieved a $3\times$ reduction in the number of parameters for MobileNetV3, while maintaining 95% of its original accuracy on

ImageNet. This makes it highly suitable for real-time applications on edge devices such as smartphones and IoT devices.

Channel pruning has been combined with other compression techniques, such as quantisation and knowledge distillation, to create highly efficient models for deployment in resource-constrained environments. Chen et al. (2021) integrated channel pruning with quantisation, achieving a top-1 accuracy of 72.8% on ImageNet for a pruned and quantised version of ResNet-34, with a $4\times$ reduction in the number of parameters and a $3\times$ reduction in FLOPs. This hybrid approach ensures that the model remains efficient and accurate and suitable for mobile and embedded applications.

Furthermore, channel pruning has been applied to object detection models, where maintaining high accuracy is crucial while reducing the size of the model and the computational cost. Li et al. (2021b) applied channel pruning to YOLOv3, achieving a $2.5\times$ reduction in the number of FLOPs while maintaining 98% of the original mean average precision (mAP) in the COCO data set. This makes the pruned model highly efficient for real-time object detection tasks on edge devices.

### 4.1.4 Filter pruning

Filter pruning involves identifying and eliminating redundant or less impactful filters within the network without compromising overall performance (Fig. 6). Filter pruning is motivated by the recognition that neural networks often contain filters that contribute minimally to the network's representational capacity. Removing these less crucial filters can result in a more streamlined model with improved efficiency, making it particularly valuable for deployment in resource-constrained environments such as edge devices or mobile applications. A common strategy involves evaluating the importance of filters based on metrics such as weight magnitudes, activation patterns, or gradient information during the training process. Filters identified as less critical are subsequently pruned from the network. Another approach incorporates regularisation techniques tailored to encourage sparsity in the network. L1 regularisation, for example, penalises nonessential parameters, making them more likely to be pruned during the training phase. Deep compression with filter pruning has been shown to result in significant model compression without sacrificing accuracy, particularly in tasks such as image classification (Han et al. 2015). Others have formulated filter pruning as an optimisation problem and efficiently solved it through a three-step algorithm (He et al.

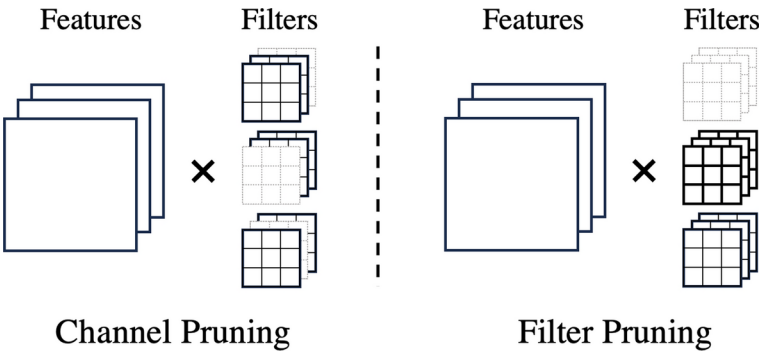


**Fig. 6** Channel pruning versus filter pruning

2017b). Recent advances in filter pruning include the AMC (AutoML for Model Compression) framework, which leverages reinforcement learning to dynamically determine which filters to prune (He et al. 2018). The choice of a specific pruning method often depends on the specific application, resource constraints, and the desired trade-off between model size and performance.

In recent times, there have been numerous groundbreaking developments in filter pruning methods that aim to improve the effectiveness and efficacy of deep neural networks. Soft filter pruning is one such technique; it permits dynamic modifications to be made during training, allowing previously pruned filters to be reinstated in subsequent training epochs. The method illustrated by He et al. (He et al. 2019) provides greater versatility and adaptability when it comes to managing a wide range of dynamic and changing datasets. Moreover, Wang *et al.* (Wang and Li 2021) have devised an energy-aware pruning approach that prioritises the optimisation of neural network energy consumption and computational efficiency. This characteristic makes the strategy highly suitable for implementation in environments with energy restrictions, such as mobile devices.

Genetic algorithms have been used to optimise filter selection in a more exploratory way, further expanding the scope of filter pruning (Li and Wong 2022). This methodology emulates a natural selection process in which only the most efficient filters are preserved, revealing distinctive network configurations that traditional techniques may fail to identify (Li and Wong 2022). In addition,(Zhao and Liu 2023) have proposed a Bayesian sparsity pruning technique that employs probabilistic models to evaluate the redundancy of filters. This approach is particularly valuable in situations where there are ambiguous or noisy data. It offers a statistically reliable framework for making pruning decisions (Zhao and Liu 2023).

Filter pruning improves the scalability of neural networks by reducing the number of filters, leading to a more compact model that can be efficiently deployed across various hardware platforms. He et al. (2017b) demonstrated that filter pruning could reduce the number of parameters in ResNet-50 by 50% while maintaining 99% of its original accuracy on ImageNet. This reduction in complexity makes the pruned model highly scalable and suitable for deployment on both CPUs and GPUs. The structured sparsity introduced by filter pruning makes it flexible for implementation on different types of hardware without requiring specialised inference engines. Li et al. (2020) applied filter pruning to MobileNetV2, achieving a $2\times$ reduction in the number of FLOPs while maintaining a top-1 accuracy of 70.6% on ImageNet. This flexibility allows the pruned model to be used in a wide range of applications, from mobile devices to high-performance servers.

Recent advances in filter pruning have focused on optimising models for edge devices by further reducing computational complexity and memory usage. He et al. (2020) proposed a learning-based filter pruning method that achieved a $3\times$ reduction in the number of parameters for MobileNetV3, while maintaining 95% of its original accuracy on ImageNet. This makes it highly suitable for real-time applications on edge devices, such as smartphones and IoT devices. Filter pruning has been combined with other compression techniques, such as quantisation and knowledge distillation, to create highly efficient models for deployment in resource-constrained environments. Chen et al. (2021) integrated filter pruning with quantisation, achieving a top-1 accuracy of 72.8% on ImageNet for a pruned and quantised version of ResNet-34, with a $4\times$ reduction in the number of parameters and a $3\times$ reduction in FLOPs. This hybrid approach ensures that the model remains efficient and accurate and suitable for mobile and embedded applications.

### 4.2 Sparse representation

Sparse representation, characterised by the presence of a relatively small number of non-zero elements (Fig. 7), has emerged as a crucial aspect of neural network optimisation. The motivation behind promoting sparsity in neural networks lies in the observation that not all connections and activations are equally essential for effective learning and representation. Sparse representations facilitate a more compact and interpretable model, reducing redundancy, and improving generalisation.

Various techniques have been proposed to induce sparsity in neural networks during training and inference. A common technique is to apply L1 regularisation in the training process to penalise nonessential parameters and encourage the model to set many of them to zero. Another technique is the use of activation functions, such as ReLU, that naturally lead to sparse activations and can contribute to a sparser network.

The introduction of sparsity in neural networks offers several advantages. First, sparse networks have lower computational and memory requirements, making them more efficient and suitable for deployment in resource-constrained environments. Second, sparse models are more interpretable, as nonzero weights and activations can be directly associated with influential features and connections. And third, sparsity acts as a form of regularisation, preventing overfitting and resulting in a more robust model.

Despite the advantages, sparse representation in neural networks poses challenges, such as determining an optimal level of sparsity and addressing potential performance trade-offs. The ongoing research aims to develop more efficient pruning algorithms, adaptive sparsity-inducing techniques, and strategies to mitigate sparse model drawbacks.

Another progress is the emergence of structured pruning approaches that not only eliminate individual weights but also prune entire filters or layers in a network. This strategy preserves structural integrity while substantially decreasing the complexity. Chen and Zhang (2023) applied structured pruning to ResNet-50, achieving a $3\times$ reduction in the number of parameters and a $2\times$ reduction in FLOPs while maintaining 99% of the original accuracy on ImageNet. This makes the model highly efficient for real-time image classification tasks on edge devices.

Furthermore, scholars have investigated the possibility of incorporating sparsity into additional network optimisation techniques. Wang and Lee (2024) implemented a hybrid approach of quantisation and sparsity on MobileNetV2, achieving a top-1 accuracy of 70.8% on ImageNet with a $4\times$ reduction in model size and a $3\times$ reduction in inference latency. This makes it suitable for deployment in mobile devices for applications such as real-time image recognition and augmented reality.

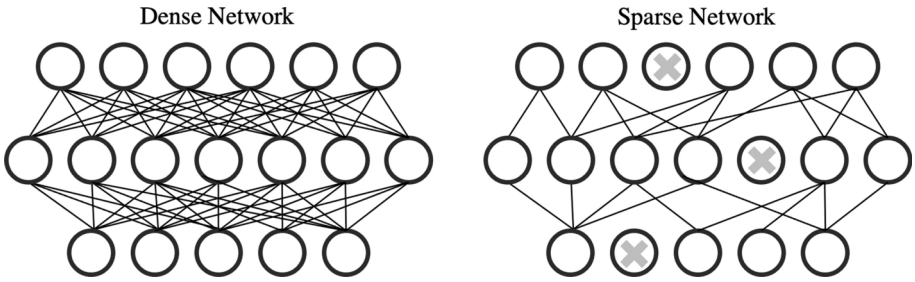


**Fig. 7** Dense versus sparse representation

Furthermore, an innovative approach to dynamically modify sparsity levels during training was proposed by Kim et al. in the form of an adaptive sparsity method.It used an adaptive sparsity approach on BERT, achieving a 2× reduction in model size and a significant decrease in computational cost while maintaining 98% the original precision on the GLUE benchmark. This makes it feasible to implement BERT on mobile devices for real-time text processing tasks.

Furthermore, certain works concentrate on the theoretical aspects of sparsity, in addition to these algorithmic developments. Liu et al. conduct theoretical research on the limitations of generalisation errors in sparse neural networks. Their study sheds light on how sparsity enhances the resilience and generalisation capabilities of models (Liu and Sharma 2024).

### 4.3 Quantised neural network

Quantised neural networks (QNNs) are networks in which the precision of the model parameters is reduced to a discrete set of quantised values (Fig. 8). Quantisation involves mapping high-precision weights to lower bit-width representations, such as 8-bit integers. QNNs offer a trade-off between model accuracy and computational efficiency, making them particularly attractive for resource-constrained environments.

Several quantisation techniques have been developed to reduce the bit-width of weights and activations in neural networks. The most common technique is to reduce the precision of the weights. For example, weights originally represented as 32-bit floating-point numbers can be quantised to 8-bit integers. In addition to quantising the weights, the activations can be quantised as well, further contributing to the overall compression of the model. The most extreme form of quantisation is to binarise the weights and activations (1-bit). Although challenging, binary quantisation significantly reduces memory requirements and computational complexity. Finally, mixed precision quantisation allows different layers of the neural network to be quantised to different bit-widths based on their sensitivity to precision, optimising the trade-off between accuracy and efficiency.

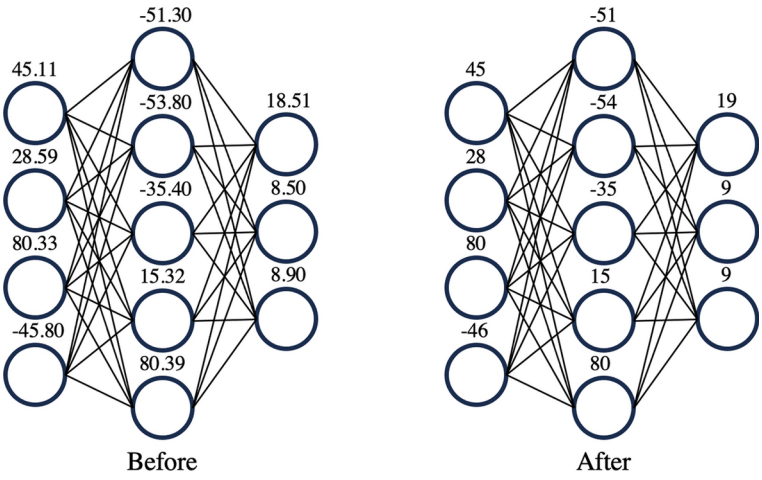


**Fig. 8** Quantisation of a neural network. Here, the original model parameter values before quantisation are floating-point numbers, which are quantised to integers

The use of QNNs has several advantages. Quantisation substantially decreases the memory requirements of neural network models, allowing deployment on memory-constrained devices. Furthermore, with lower bit-width representations, quantised models lead to faster inference times, making them suitable for real-time applications. In addition, quantised models require less computational resources, resulting in reduced energy consumption, which is particularly crucial for edge devices and mobile applications.

On the other hand, QNNs come with challenges and considerations. Reduction in precision can lead to loss of model accuracy, especially for tasks that require fine-grained distinctions. To mitigate accuracy degradation, it is crucial to train models with quantisation in mind and employ quantisation-aware training techniques (Chung et al. 2020; Shen et al. 2021). Choosing the appropriate bit-width for quantisation is a nontrivial task and often depends on the specific characteristics of the neural network and the target platform.

Recent research has focused on developing advanced quantisation methods and training strategies to overcome the challenges associated with QNNs. Techniques like mixed-precision quantisation (Pandey and Kumar 2023), learning-aware quantisation (Huang et al. 2023b), and hardware-aware quantisation (Huang et al. 2024) aim to push the boundaries of model efficiency while maintaining acceptable levels of accuracy.

An approach worth mentioning is adaptive quantisation, which modifies the bit-widths dynamically throughout the training process in response to the weight and activation distribution. The aforementioned approach enables greater precision in regulating the quantisation procedure, potentially resulting in improved efficacy when faced with intricate tasks (Zhou and Zhang 2023). The utilisation of quantisation intervals that are learnt during training as opposed to fixed is an additional significant development. The method outlined in the article by Lee et al. allows the quantisation procedure to adjust to the unique attributes of the data, potentially improving the reliability and precision of the quantised model (Lee and Kim 2024).

Moreover, researchers have also investigated the use of quantisation with other methods for compressing networks, such as pruning and knowledge distillation. Smith and colleagues propose an integrated framework that integrates various strategies to significantly reduce the size of the model and computing needs, while maintaining a reasonable level of accuracy (Smith and Gupta 2023).

Research on application-specific quantisation techniques has recently gained attention. In the realm of natural language processing, specific quantisation strategies have been created to address the distinct difficulties presented by huge language models. Nguyen et al. present a technique tailored for transformer designs that effectively manages the trade-off between efficiency and performance in these computationally demanding models (Nguyen and Pham 2024).

Mixed precision quantisation has been widely adopted for deploying models on mobile and edge devices due to its balance between accuracy and efficiency. Pandey and Kumar (2023) demonstrated that a mixed-precision quantised version of ResNet-50 achieved a top–1 accuracy of 75. 3% on ImageNet while reducing model size by $4\times$ and inference latency by $3\times$ compared to the full-precision model. This makes it suitable for real-time image recognition tasks on mobile devices. Nguyen et al. (2024) applied quantization techniques to BERT, a popular transformer model, reducing its memory footprint by $3\times$ and maintaining 97% of its original accuracy on the GLUE benchmark. This makes it feasible to deploy BERT on mobile devices for real-time text processing tasks. Huang et al. (2024)

implemented a hardware-aware quantised version of MobileNetV2, achieving a top-1 accuracy of 72.2% on ImageNet with only 220 million FLOPs and significant energy savings. This approach is particularly beneficial for edge devices with limited power resources.

### 4.4 Efficient lightweight neural networks

Many neural network architectures exist nowadays that are lightweight by design rather than by reducing a given large network using pruning, sparsification, or quantisation. Here, we survey the most prominent and efficient lightweight neural networks.

#### 4.4.1 MobileNet

MobileNet (Howard et al. 2017) is a family of neural network architectures designed for mobile and embedded devices. The first iteration, MobileNetV1, was introduced in 2017, pioneering the use of depthwise separable convolutions for efficient model design. It struck a balance between the size of the model and its accuracy, making it suitable for on-device tasks. MobileNetV2 followed in 2018, improving performance through inverted residual blocks and linear bottlenecks. It offered better accuracy and efficiency, particularly for real-time applications on mobile hardware. MobileNetV3 continued this trend in 2019 with optimisations such as the h-swish activation function, further improving both accuracy and efficiency. These models are well-regarded for their practicality in various mobile vision tasks.

Howard et al. (2017) demonstrated that MobileNetV1 achieved a top-1 accuracy of 70.6% on ImageNet with 569 million FLOPs, significantly reducing computational cost compared to traditional CNNs. MobileNetV2 further enhanced scalability, achieving a top-1 accuracy of 72.0% on ImageNet with only 300 million FLOPs (Sandler et al. 2018). Howard et al. (2019) demonstrated that MobileNetV3-Large achieved a top-1 accuracy of 75.2% on ImageNet with only 219 million FLOPs, making it highly suitable for mobile and edge applications such as real-time image recognition and augmented reality. Natarajan et al. (2021) used MobileNetV3 to detect pneumonia in chest X-rays. MobileNetV3 achieved a sensitivity of 93.3% and specificity of 92.1% while significantly reducing inference time compared to previous models, making it practical for real-time diagnostic tools on edge devices.

MobileNet variants have exhibited considerable success in various practical domains, highlighting their performance in edge-specific environments. In agriculture, Mohanty et al. (2020) employed MobileNetV2 for the real-time detection of plant diseases utilising images obtained from smartphones. This application demonstrates MobileNetV2's efficacy in low-latency, on-device processing, allowing farmers to make prompt, informed decisions in the field without reliance on high-performance computing resources. The model's capacity to operate on devices with constrained computational resources while preserving high accuracy allows implementation in rural or resource-deficient environments. Likewise, Wang et al. (2021b) utilised MobileNetV3 on drones for instantaneous object detection in precision agriculture. Their study illustrated that the lightweight architecture of MobileNetV3 enabled drones to execute precise detection tasks without overburdening onboard processors, thereby ensuring extended flight durations and enhanced power efficiency.

In healthcare, in addition to pneumonia detection, MobileNet has been utilised in numerous diagnostic instruments. For example, Rajpurkar et al. (2019) utilised MobileNetV2 to identify anomalies in chest X-rays within the framework of a comprehensive mobile health initiative. This application highlights MobileNet's efficacy in providing real-time diagnostic insights in regions with restricted access to conventional medical imaging tools. Furthermore, Su et al. (2020) employed MobileNetV3 for real-time ECG monitoring on wearable devices, demonstrating its capacity to manage continuous time-series data with minimal latency, rendering it appropriate for the early identification of cardiac events in routine environments.

MobileNet variants have been extensively utilised in robotics within autonomous systems. For instance, Zhao et al. (2020) utilised MobileNetV2 for real-time object recognition on mobile robots, illustrating its applicability for edge devices with constrained computational resources and battery longevity. The model's low memory usage and fast inference speed enabled the robots to execute real-time decisions independently of cloud processing, thereby enhancing the system's reliability and responsiveness in dynamic settings. The literature case studies demonstrate the adaptability of MobileNet variants to diverse real-world edge applications, confirming their versatility and efficacy in both industrial and consumer sectors.

### 4.4.2 EfficientNet

EfficientNet (Tan and Le 2019) takes a different approach to optimisation. Using compound scaling, EfficientNetV1 uniformly scales network width, depth, and image resolution, thus creating a spectrum of models ranging from EfficientNetB0 to EfficientNetB7. EfficientNetV1 is typically synonymous with the foundational model architecture outlined in the paper. This methodology quickly garnered attention for its ability to strike a balance between efficiency and accuracy, particularly excelling in image classification tasks.

The variations within the EfficientNet family, from B0 to B7, diverge in their architectures, depth, and computational complexity. Beginning with B0, the models progressively increase in size and complexity. EfficientNetB0 is the smallest and simplest model, having fewer layers and parameters. On the contrary, EfficientNetB7 emerges as the largest, featuring a deeper and wider architecture capable of capturing intricate data patterns.

Typically, transitioning from B0 to B7 results in improved performance in tasks such as image classification or object detection. However, this improvement comes at the expense of increased model capacity, manifested in larger parameter counts and computational requirements. The decision on which variant to employ depends on the task-specific demands, available computational resources, and the desired balance between the size, speed, and accuracy of the model.

The compound scaling strategy of EfficientNet has demonstrated significant efficacy across numerous practical applications, where the trade-off between computational efficiency and accuracy is paramount. For instance, Wang et al. (2020d) illustrated the application of EfficientNetB3 in healthcare, particularly for the detection of diabetic retinopathy in retinal images. Their study demonstrated that the lightweight architecture of EfficientNet attained high sensitivity (92.1%) and specificity (90.3%) while reducing inference time, making it suitable for deployment on edge devices like portable diagnostic tools, where rapid and precise decision-making is crucial. In another study, Li et al. (2021a) utilised Effi-

cientNetB4 in autonomous vehicles for object detection and scene comprehension, where real-time performance and minimal latency are essential for safety and functionality. EfficientNetB4 delivered exceptional accuracy while optimising energy consumption, enabling efficient operation on the constrained processing power of embedded systems in vehicles.

To further illustrate EfficientNet's relevance to edge scenarios, Tan and Le (2020a) presented EfficientNet-Lite0, which was explicitly designed for mobile and edge applications by reducing model complexity while maintaining performance. EfficientNet-Lite0 achieved a top-1 accuracy of 75.1% on ImageNet with merely 5.4 million parameters and 390 million FLOPs, making it exceptionally suitable for applications like real-time image recognition on smartphones or augmented reality systems, where power efficiency and rapid inference are paramount. Furthermore, Tan and Le (2021) showed that EfficientNetV2-S enhanced accuracy, achieving 83.9% top-1 accuracy on ImageNet, while significantly decreasing training duration, rendering it beneficial for contexts necessitating swift deployment, such as in retail, healthcare, and autonomous systems, where fast training and inference are paramount.

### 4.4.3 Group convolutional networks

Group-convolutional networks (Table 5), with their various iterations and innovative approaches, offer efficient alternatives to traditional deep neural networks. They are well suited for a wide range of applications, from mobile and edge devices to real-time tasks, where computational efficiency is a critical factor. The choice of architecture depends on the specific resource constraints and performance requirements of a given project.

**Table 5** Summary of group convolutional networks

| Method | Advantage | Disadvantage | Characteristic |
|---|---|---|---|
| ShuffleNet | Significantly reduces computational costs | May have slightly lower accuracy compared to more resource-intensive models | Leverages group convolutions and channel shuffling for efficiency. Improved in ShuffleNetV2 with residual connections |
| CondenseNet | Greatly reduces model size while maintaining competitive accuracy | Pruning and condensing process may require careful tuning | Follows a two-step pipeline for pruning and condensing networks. Highly efficient |
| MixNet | Uses mixed depthwise convolutions for efficiency-accuracy trade-offs | Customisation may require expertise in architecture design | Offers flexibility and customisation to meet various resource constraints. Achieves state-of-the-art trade-offs |
| GhostNet | Introduces "ghost" modules for lightweight feature maps and reduced computation | Adaptation of the "ghost" modules may be task-specific | Maintains competitive performance with a significant reduction in parameters and computation. The application is well suited for limited computational resources |
| DiCENet | Dynamic inference channels adjust the number of active channels at runtime for efficient computation | May require additional complexity for dynamic channel management | Designed for real-time and mobile applications, prioritising adaptability and efficiency |
| MicroNet | Optimised for ultralow resource constraints | May have limitations in handling complex tasks due to resource constraints | Includes MicroNetV1 and MicroNetV2 for edge and IoT applications. Provides options for highly resource-constrained environments |

ShuffleNet (Zhang et al. 2018) is a family of neural network architectures that revolutionised the efficiency of models. ShuffleNetV1, introduced in 2017, leveraged group convolutions and channel shuffling to significantly reduce computational costs while maintaining competitive accuracy. It became the go-to choice for mobile and embedded devices where computational resources are limited. ShuffleNetV2, released in 2018, further improved efficiency by introducing residual connections, offering an even better balance between accuracy and computational efficiency. These models are known for their ability to provide impressive performance on resource-constrained platforms. ShuffleNet effectively balances accuracy and computational cost by using efficient architectural strategies. This approach ensures that the model can achieve high accuracy while minimising resource usage. In their experiments, Zhang et al. (2018) found that ShuffleNet consistently outperformed other lightweight models in terms of accuracy-efficiency trade-offs. For example, ShuffleNetV2 achieved a top-1 accuracy of 69.4% on ImageNet with only 150 million FLOPs, significantly reducing computational cost while maintaining competitive performance compared to models such as MobileNetV2 and CondenseNet.

CondenseNet (Huang et al. 2018), introduced in 2018, follows a two-step pipeline to prune and condense neural networks. Removing less important connections significantly reduces the size of the model while maintaining competitive accuracy. It is a highly efficient architecture, well-suited for applications with strict resource constraints. CondenseNet balances accuracy and computational cost effectively by pruning less important connections and condensing the network structure. This approach ensures that the model can achieve high accuracy while minimising resource usage. In their experiments, Huang et al. (2018) found that CondenseNet consistently outperformed other pruned models in terms of accuracy-efficiency trade-offs. For example, CondenseNet achieved a top-1 accuracy of 73.8% on ImageNet with only 2.5 million parameters, significantly reducing computational cost while maintaining competitive performance compared to models like MobileNet and ShuffleNet.

MixNet (Tan and Yu 2019), unveiled in 2019, uses mixed depthwise convolutions to reduce computation while preserving accuracy. It offers a flexible architecture that can be easily customised to meet different resource constraints. MixNet has earned recognition for achieving state-of-the-art efficiency-accuracy trade-offs, making it a versatile choice for various applications. MixNet balances accuracy and computational cost effectively employing mixed depth-wise convolutions. This approach ensures that the model can achieve high accuracy while minimising resource usage. In their experiments, Tan and Yu (2019) found that MixNet consistently outperformed other lightweight models in terms of accuracy-efficiency trade-offs. For example, MixNet-M achieved a top-1 accuracy of 77.0% on ImageNet with only 360 million FLOPs, significantly reducing computational cost while maintaining competitive performance compared to models such as MobileNetV3 and EfficientNet-Lite.

GhostNet (Han et al. 2020), developed in 2020, introduces “ghost” modules to produce lightweight feature maps, significantly reducing the number of parameters and computations. This innovative approach allows GhostNet to maintain competitive performance while being highly efficient, particularly suited for tasks where computational resources are limited. GhostNet balances accuracy and computational cost effectively by generating more feature maps from fewer computations. This approach ensures that the model can achieve high accuracy while minimising resource usage. In their experiments, Han et al. (2020) found that GhostNet consistently outperformed other lightweight models in terms

of accuracy-efficiency trade-offs. For example, GhostNet achieved a top-1 accuracy of 73.9% on ImageNet with only 141 million FLOPs, significantly reducing computational cost while maintaining competitive performance compared to models like MobileNetV2 and ShuffleNet.

DiCENet (Mehta et al. 2022), incorporates dynamic inference channels to adjust the number of active channels at runtime. This adaptability significantly reduces the computational and memory footprint, allowing for varying resource constraints. DiCENet is designed for real-time and mobile applications, prioritising efficiency without compromising performance. DiCENet balances accuracy and computational cost effectively by dynamically adjusting its inference channels. This approach ensures that the model can maintain high accuracy while minimising resource usage. In their experiments, Mehta et al. (2022) found that DiCENet consistently outperformed other dynamic models in terms of accuracy and efficiency trade-offs. For example, DiCENet achieved a top-1 accuracy of 74.8% on ImageNet with significantly lower computational cost compared to static models, making it an attractive choice for resource-constrained applications.

MicroNet (Banbury et al. 2021) is a collection of neural network architectures optimised for ultra-low resource constraints. MicroNet includes MicroNetV1 and MicroNetV2, specifically designed for edge devices and scenarios with extremely limited computational resources. These models provide a range of options for resource-constrained environments, making them highly suitable for edge and IoT applications. MicroNet balances accuracy and computational cost by employing efficient architectural strategies tailored for low-resource environments. In their experiments, Banbury et al. (2021) found that MicroNet models consistently outperformed other ultra-low-resource architectures in terms of accuracy while maintaining minimal computational requirements. For example, MicroNetV2 achieved a top-1 accuracy of 67.3% on ImageNet with only 150 million FLOPs, outperforming other lightweight models like TinyNAS (Wu et al. 2021) and EfficientNet-Lite (Tan and Le 2020b).

#### 4.4.4 Squeeze & excitation

Squeeze & excitation (SE) networks, including SqueezeNet, SENet, and SqueezeNeXt (Table 6), have made significant contributions to the field of deep learning. SqueezeNet focusses on model compression and efficiency, while SENet uses crucial attention mechanisms that enhance model performance, and SqueezeNeXt combines the best of both worlds,

**Table 6** Summary of squeeze & excitation networks

| Method | Advantage | Disadvantage | Characteristic |
|---|---|---|---|
| SqueezeNet | Exceptional model compression and efficiency | Lacks attention mechanisms for enhanced feature selection | Utilises “fire” modules for model compression. Ideal for resource constrained devices |
| SENet | Uses attention mechanisms for improved feature selection | May be computationally more expensive due to attention mechanisms | Employs “squeeze” and “excitation” steps for adaptive recalibration of channel importance. Achieves state-of-the-art accuracy |
| SqueezeNeXt | Combines efficiency of SqueezeNet with the SENet attention mechanisms | Slightly more complex compared to SqueezeNet | Integrates parallel pathways with varying channel sizes and channel-wise attention mechanisms for a balance between efficiency and accuracy |

offering a balance between computational efficiency and accuracy, which makes it a valuable choice for various applications with limited computational resources.

SqueezeNet (Iandola et al. 2016), introduced in 2016, is a pioneering SE network that stands out for its exceptional compression and efficiency. It achieves these characteristics by using small $1\times1$ convolutions, known as "fire" modules, to reduce the number of parameters while preserving the representational power of the network. This is a particularly popular choice for deployment on resource constrained devices, such as smartphones and IoT devices. Although it focuses on model compression, it does not incorporate the attention mechanisms of later SE models. SqueezeNet balances accuracy and computational cost, making it a practical choice for resource-constrained environments. In their experiments, Iandola et al. (2016) found that SqueezeNet maintained competitive accuracy levels while drastically reducing the number of parameters. For example, SqueezeNet achieved a $50\times$ reduction in parameters and a $2\times$ reduction in inference time compared to AlexNet, making it suitable for real-time applications.

SENet (Hu et al. 2018), introduced in 2017, addresses the limitations of conventional neural networks by introducing attention mechanisms. It employs a "squeeze" step to capture global statistics from feature maps, and an "excitation" step to adaptively recalibrate the importance of different channels. This self-attention mechanism significantly increases model performance by allowing the network to focus on relevant information and ignore less informative features. SENet has become a critical advancement in computer vision, achieving state-of-the-art accuracy in various image classification tasks. The squeeze-and-excitation mechanism is flexible and can be applied to various types of neural network, including CNN and RNN. This flexibility allows SENet to be used in a wide range of applications. Hu et al. (2018) showed that SENet could be effectively applied to different architectures such as ResNet, Inception, and MobileNet. For example, SENet improved the top-1 accuracy of MobileNetV2 from 71.8% to 72.9% on the ImageNet dataset.

SqueezeNeXt (Gholami et al. 2018) builds on the concepts of SqueezeNet and SENet. Introduced in 2017, it combines the efficiency of SqueezeNet with the SENet attention mechanism. The network employs parallel paths with varying channel sizes to capture a wide range of information, promoting both efficiency and accuracy. By integrating channel-wise attention mechanisms inspired by SENet, SqueezeNeXt effectively boosts model performance. This architecture has become a popular choice for applications where a balance between computational efficiency and accuracy is essential, particularly in scenarios with limited computational resources. SqueezeNeXt balances accuracy and computational cost by leveraging the efficiency of SqueezeNet and the performance boost from SENet's attention mechanisms. In their experiments, Gholami et al. (2018) found that SqueezeNeXt consistently outperformed other lightweight models in terms of accuracy and efficiency trade-offs. For instance, SqueezeNeXt-M achieved a top-1 accuracy of 71.2% on ImageNet with only 2.8 million parameters and 1.5 billion FLOPs, making it an attractive choice for resource-constrained environments.

### 4.4.5 Mobile transformers

Mobile transformers are transformer-based neural network architectures designed to be efficient and lightweight, making them suitable for deployment on mobile devices with limited computational resources. Transformers, originally introduced for natural language process-

ing tasks, have proven to be powerful for a wide range of applications, including computer vision.

MobileViT (Mehta and Rastegari 2022) combines the strengths of lightweight CNNs and heavy weight self-attention-based vision transformers (ViT) to create a lightweight and low-latency network for mobile vision tasks. It introduces a unique vision transformer designed for mobile devices, treating transformers as convolutions for a different perspective on global information processing. Experimental results show that MobileViT outperforms CNN-based networks (specifically MobileNetV3) and ViT-based networks (specifically DeIT) in various tasks and datasets. In the ImageNet-1k dataset, MobileViT achieves a top-1 accuracy of 78.4% with approximately 6 million parameters, surpassing MobileNetV3 by 3.2% and DeIT by 6.2% with a similar parameter count. In object detection on MS-COCO, MobileViT is 5.7% more accurate than MobileNetV3 with a comparable number of parameters.

EdgeViTs (Pan et al. 2022) address the computational and model size challenges associated with self-attention-based ViTs for mobile devices. They introduce a new family of lightweight ViTs with a focus on low on-device latency and high energy efficiency. EdgeViTs incorporate a local-global-local (LGL) information exchange bottleneck that efficiently integrates self-attention and convolutions. This design allows them to compete with the best lightweight CNNs in terms of the trade-off between accuracy and on-device efficiency. EdgeViTs are positioned as Pareto-optimal models, excelling in both accuracy-latency and accuracy-energy trade-offs. Models demonstrate strict dominance over other ViTs in almost all cases and compete with the most efficient CNNs when evaluated based on on-device latency and energy efficiency. EdgeViTs are designed to efficiently balance accuracy and computational cost. The LGL bottleneck enables models to achieve high accuracy while minimising computational overhead. In their experiments, Pan et al. (2022) found that EdgeViTs consistently outperformed other lightweight models in terms of accuracy-latency and accuracy-energy trade-offs. For example, EdgeViTs achieved a top-1 accuracy of 76.5% on ImageNet with a latency of only 18ms on a Pixel 4 phone, positioning them as a leading choice for mobile vision applications.

Mobile transformer architectures, including MobileViT and EdgeViTs, have proven effective in numerous real-world applications. For example, Chen et al. (2021) utilised mobile transformers for facial recognition tasks in low-power smart devices, including smartphones and smart doorbells. The model's transformer-based architecture showed exceptional performance in facial recognition with minimal power consumption, rendering it an ideal solution for battery-operated edge devices. This application demonstrates the capacity of mobile transformers to perform complex vision tasks, such as facial recognition, in real-time without relying on cloud processing, thereby safeguarding privacy and security. Mobile transformers have been employed in healthcare for remote patient monitoring systems. Xu et al. (2021) implemented a mobile transformer-based model for the real-time analysis of medical images, specifically for identifying abnormalities in chest X-rays and CT scans.

Another practical application is in autonomous retail, where mobile transformers are utilised for object detection in cashierless checkout systems. Lin et al. (2022) illustrated the application of EdgeViTs for real-time product identification with negligible latency in low-power embedded systems deployed in retail settings. These systems depend on efficient and precise object detection to guarantee seamless customer experiences, and the transformer-based architecture enabled them to surpass conventional CNN models while preserving a

minimal computational footprint. Mobile transformers have also been explored for augmented reality (AR) applications, especially for real-time object tracking. Kim et al. (2022) utilised EdgeViT to facilitate augmented reality systems in tracking multiple objects within dynamic environments with minimal latency. The low latency and high accuracy of mobile transformers demonstrate their potential to improve mobile gaming and augmented reality experiences on smartphones, which often have constrained computational resources.

### 4.5 Neural architecture search

Neural architecture search (NAS) is a subfield of DL that focusses on automating the design of neural network architectures (Elsken et al. 2019). Traditional neural network architecture design often involves human expertise and manual trial and error, which can be time consuming and may not always yield the best results. NAS aims to address these limitations by using search algorithms to discover optimal neural network architectures for specific tasks. We briefly discuss some notable NAS approaches and architectures (Table 7). For a comprehensive overview of NAS, the readers can refer to the survey by (Elsken et al. 2019).

PNASNet (Progressive Neural Architecture Search Network) (Liu et al. 2018), introduced in 2017, is a NAS method that employs a progressive search strategy. PNASNet's progressive search strategy starts with a smaller, simpler network and incrementally increases its complexity. This method ensures that the search process remains computationally feasible even as the architecture grows. For example, PNASNet achieved competitive performance in the CIFAR-10 dataset with significantly fewer computational resources compared to other NAS methods. Specifically, PNASNet was able to achieve a top-1 error rate of 3.41% on CIFAR-10 while using only 255 GPU h for the architecture search, compared to NASNet's 2000 GPU h (Liu et al. 2018). One of the strengths of PNASNet is its ability to balance accuracy with computational cost. During the search process, both these factors are considered, ensuring that the resulting architecture is not only accurate, but also efficient. For example, in the study by Liu et al. (2018), PNASNet achieved a top-1 accuracy of 82.9% in CIFAR-10 with significantly lower computational resources compared to other architectures such as AmoebaNet (Real et al. 2019b), which required much higher computational costs to achieve similar accuracy levels.

MNASNet (Mobile Neural Architecture Search Network) (Tan et al. 2019), proposed in 2019, is designed for mobile and edge devices with limited computational resources. MNASNet's architecture search is guided by a multi-objective optimisation framework that explicitly incorporates latency into the optimisation process. This ensures that the resulting architectures are scalable and can be efficiently deployed on devices with varying computational capabilities. For example, Tan et al. (2019) demonstrated that MNASNet achieved

**Table 7** Summary of NAS methods

| Method | Advantage | Disadvantage | Characteristic |
|---|---|---|---|
| PNASNet | Progressive search strategy | Computational cost may still be high | Uses progressive growth in architecture |
| MNASNet | Designed for mobile and edge devices | May require reinforcement learning expertise | Uses reinforcement learning for search |
| FBNet | Factorised search space for flexibility | Trade-off between performance and efficiency | Focuses on balancing performance and efficiency |
| AmoebaNet | Achieved state-of-the-art results | Evolutionary algorithm can be computationally intensive | Uses an evolutionary algorithm for search |

a top-1 accuracy of 74.0% on ImageNet with 4.9 billion multiply-add operations (MAdds), making it $1.5\times$ more efficient than MobileNetV2 (Sandler et al. 2018).

FBNet (Facebook Network) (Wu et al. 2019a), also from 2019, is an AI-assisted NAS approach developed by Facebook that targets efficient neural network architectures. FBNet's architecture search uses a factorised approach that decouples the selection of operations from their connectivity. This makes the search space more manageable and allows for efficient exploration of potential architectures. For instance, Wu et al. (2019a) demonstrated that FBNet achieved significant improvements in both performance and efficiency compared to manually designed models. Specifically, FBNet achieved a top-1 accuracy of 74.9% on ImageNet with 375 million FLOPs, which is $1.5\times$ more efficient than MobileNetV2 (Sandler et al. 2018).

AmoebaNet (Real et al. 2019a), introduced in the same year, is known to achieve state-of-the-art results in NAS using an evolutionary algorithm. Explores a wide search space, allowing for the discovery of complex and effective architectures. Together with other NAS approaches, it has contributed to the advancement of DL by automating the design of neural networks and improving their performance in various tasks. NAS continues to play a key role in obtaining networks that are more efficient, effective, and adaptable to various applications and hardware constraints. For example, Real et al. (2019a) demonstrated that AmoebaNet achieved a top-1 accuracy of 83.9% on ImageNet, which was state of the art at the time. The scalability of AmoebaNet is evident in its ability to generate architectures that perform well on both small datasets like CIFAR-10 and large datasets like ImageNet. AmoebaNet achieves a high level of accuracy using an evolutionary algorithm that optimises both performance and computational cost. For example, AmoebaNet achieved a top-1 accuracy of 96.7% on CIFAR-10 and 83.9% on ImageNet.

In real-world applications, edge computing models have to trade-off various aspects such as accuracy, latency, memory consumption, and power efficiency. An example is the work of Jiang et al. (2020), in which a lightweight model for autonomous drone navigation was designed using NAS. This model adeptly balanced precision and computational efficiency, enabling drones to execute real-time object detection and navigation in dynamic environments without dependence on cloud-based processing, thereby illustrating the relevance of NAS to practical edge computing applications. He et al. (2021) used NAS in a study optimising models for wearable medical technology. The architecture, developed via NAS, was specifically designed for real-time monitoring of physiological signals, including heart rate and oxygen levels, on edge devices. Through the implementation of techniques such as quantization and network pruning, they effectively diminished the model's size and power consumption while preserving high accuracy, making it suitable for continuous monitoring in low-power settings such as smartwatches and fitness trackers. In the automotive industry, NAS was used by Liu et al. (2021) to optimise models for advanced driver assistance systems (ADAS) and autonomous driving. Their NAS-generated models were implemented on in-vehicle edge devices, where the capacity to process visual information efficiently and rapidly was essential. The research indicated that NAS could generate architectures that satisfied the stringent latency and energy efficiency criteria essential for real-time decision-making in vehicles.

## 5 Computer vision edge applications

With the advancement of computer vision, lightweight models, and model optimisation techniques, an increasing number of edge applications are being explored across various types of edge devices. Here we survey several primary application domains (Fig. 9): autonomous driving and intelligent transportation, smart manufacturing and industrial automation, agriculture and crop monitoring, retail and shelf monitoring, environmental monitoring and wildlife conservation, smart cities and public safety, and sports analytics. Emphasis is placed on applications that leverage the real-time processing capabilities of edge devices to enhance operational efficiency and decision-making. By reviewing nonmedical applications of edge deep learning, we aim to highlight their potential contributions to advancing medical diagnostics.

### 5.1 Autonomous driving and intelligent transportation systems

Autonomous driving and intelligent transportation are critical domains that leverage advanced edge computer vision technologies to improve vehicle autonomy and traffic management. These technologies contribute significantly to real-time object detection, tracking and classification, which are vital to the safety and efficiency of transportation systems.

Wang et al. (2021a) proposed an improved object detection method based on YOLOv4 (Bochkovskiy et al. 2020) tailored for autonomous vehicles, focusing on the trade-off between speed and accuracy. Their method improved detection accuracy and inference speed by incorporating modifications in the backbone, neck, and predictor head of the YOLOv4 architecture. It showed a notable increase in average accuracy in the KITTI (Geiger et al. 2013) and BDD (Yu et al. 2020) datasets, demonstrating its potential for real-time applications in autonomous driving. Additionally, Mauri et al. (2022) proposed a lightweight

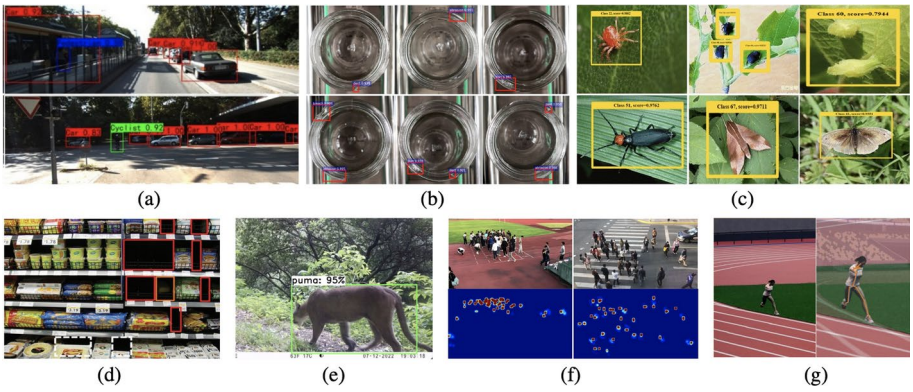


**Fig. 9** Examples of computer vision edge applications. **a** Detecting complex dynamic driving environments by fusing RGB-camera and LiDAR data (Liu et al. 2023a). **b** Detecting defects such as cracks and dents on seal surfaces of containers in filling lines using a lightweight network based on MobileNet (Li et al. 2018). **c** Identifying and classifying pests in crops via drones (Albattah et al. 2023). **d** Monitoring out-of-stock situations in retail environments using mobile robots (De Simone et al. 2023). **e** Detecting wildlife using lightweight models deployed on Raspberry Pi (Tulasi et al. 2023). **f** Improving automatic people counting in smart city environments through surveillance videos (Avvenuti et al. 2022). **g** Analyzing human motion through edge photographic equipment (Baumgartner and Klatt 2023)

CNN based on YOLOv5 (Jocher 2020) for real-time 3D object detection in road and railway environments, improving on real-time embedded constraints and enhancing its applicability.

Another approach, by Liu et al. (2023a), tackled the challenges of object detection by fusing data from RGB cameras and LiDAR. Their method uses the richness of semantic information from cameras with the accurate depth data from LiDAR sensors, improving the reliability of object detection units. The proposed siamese network structure and feature-layer fusion strategy significantly enhanced detection performance, particularly in complex and dynamic driving environments.

In the domain of traffic monitoring, Wan et al. (2022) introduced an edge computing-based video segmentation method. They optimised traditional video processing techniques to reduce redundancy and enhance real-time performance. By implementing spatiotemporal interest points and multimodal linear feature combinations, their method efficiently segments traffic videos, enabling effective monitoring and management of urban traffic.

Moreover, Fernández et al. (2021) developed TrafficSensor, a robust real-time traffic surveillance system using DL. The system uses a calibrated camera to track and classify vehicles on highways under various conditions, including poor lighting and adverse weather. Using advanced neural networks and tracking algorithms, TrafficSensor provides reliable data for infrastructure planning and traffic management, showcasing the capabilities of edge computing in intelligent transportation systems. Furthermore, the method introduced by Zou et al. (2022) involves shifting video analysis tasks to edge devices. This aims to facilitate low-latency and high-accuracy experiences in real-time traffic monitoring for connected vehicle networks.

### 5.2 Smart manufacturing and industrial automation

In modern manufacturing, edge computer vision has made significant strides in smart manufacturing and industrial automation. These advances are crucial to improving the precision, efficiency, and safety of manufacturing processes. A core aspect of this technological evolution is the development of sophisticated real-time defect detection systems, critical to maintaining product quality and optimising manufacturing workflows.

Yi et al. (2022) tackled the challenge of detecting bead-toe defects in tyre X-ray images. They proposed an innovative lightweight semantic segmentation network that first extracts texture features from various tyre regions and then employs a decoder to fuse these features. This effectively reduces the dimensions of feature maps to pinpoint bead-toe positions.

For the detection of surface defects in sanitary ceramics, Hang et al. (2022) proposed a lightweight real-time defect detection network that uses MobileNetV3 (Koonce and Koonce 2021) as the backbone. The network achieves multiscale detection of surface defects through a multilayer feature pyramid and combines a region proposal network with an anchor-free method. The detection head incorporates a channel attention structure and a low-level mixed feature classification strategy is employed to achieve higher accuracy in defect classification. This method demonstrated significant improvements, including at least 22.9% faster detection and 35.0% higher average precision, while reducing memory consumption by at least 8.4% compared to classical detection methods such as SSD (Liu et al. 2016), YOLO V3 (Redmon and Farhadi 2018), and Faster R-CNN (Girshick 2015).

Furthermore, Chen et al. (2022b) presented an improved YOLOv3 model to detect surface defects on polarisers. They modified the model by replacing DarkNet53 (Redmon et

al. 2016) with MobileNet (Howard et al. 2017) as the backbone, significantly reducing the number of network parameters and enhancing detection speed. The introduction of the Mixed Convolution Efficient Attention (MECA) module further improved detection accuracy. The modified network showed a detection speed of 121 frames/second and an mAP of 89.8%, demonstrating a 44% increase in speed compared to the standard YOLOv3.

To detect defects in industrial insert moulding using X-ray images, Wang and Huang (2021) developed a lightweight deep network based on the YOLOv5 model, incorporating the Ghost module to reduce the model size and a transformer module for enhanced spatial multiheaded attentional feature extraction. The network achieved an mAP of 93.6% and showed robustness under different luminance and noise conditions.

Xia et al. (2019) proposed SSIM-NET, a new print circuit board (PCB) defect detector combining the structural similarity index (SSIM) and MobileNetV3 (Howard et al. 2019). This two-stage detection algorithm first uses SSIM to identify suspicious regions, less susceptible to environmental factors such as variation in illumination. In the second stage, MobileNetV3 with a binary loss and focal loss is used to classify these regions, significantly reducing computational costs. The approach achieved 12× speed increase over Faster-RCNN without sacrificing accuracy.

Additionally, Dai et al. (2020) introduced an integrated detection framework for solder joint defects in PCBs using automatic optical inspection (AOI). They employed a generic DL method for localisation that is easily adaptable to various PCB configurations and soldering technologies. For classification, they introduced an active learning method to minimise the labeling workload. Their approach demonstrated fast, accurate localisation, and high classification accuracy with minimal user input.

To achieve accurate and real-time surface defect detection, Li et al. (2018) optimised the SSD network structure by integrating it with MobileNet, resulting in a method called MobileNet-SSD. This method was particularly effective in detecting typical defects such as breaches, dents, and burrs on container sealing surfaces in filling lines.

Bonam et al. (2023) investigated the application of lightweight CNN models for product defect detection in the manufacturing industry, focusing on utilising edge deep learning to address the issues of labour intensity, error-proneness, and unreliability in detecting defects in fabrics, surfaces, and castings. Specifically, the authors employed lightweight models such as MobileNetV2, ShuffleNetV2, and CondenseNetV2 for defect detection in these areas and successfully deployed these models on edge devices with limited computational capabilities, such as the NVIDIA Jetson Nano. The results demonstrated that these lightweight models achieved high accuracy and efficiency in defect detection, with MobileNetV2 achieving a test accuracy of 96.87% on the fabric defect dataset, ShuffleNetV2 achieving 99% on the surface defect dataset, and CondenseNetV2 achieving 98.08% on the casting defect dataset. This showcases the potential of lightweight CNN models to enhance the efficiency and reliability of defect detection in manufacturing environments.

These edge DL algorithms are increasingly being integrated with existing manufacturing processes, thereby enhancing precision and efficiency. However, this integration also presents certain challenges. Ensuring compatibility with legacy systems can be complex, often requiring modifications or upgrades to existing infrastructure. Additionally, the initial cost of implementing and deploying these edge deep learning algorithms can be substantial. Despite these challenges, the benefits of edge DL, including reduced waste, improved product quality, and increased production efficiency, make it an ideal choice for modern

manufacturing (Nain et al. 2022). Moreover, some continual learning approaches (Zhang et al. 2023; Dekhovich and Bessa 2024; Chang et al. 2024) confer additional advantages over traditional methods, making edge DL particularly well-suited for application in the manufacturing sector.

### 5.3 Agriculture and crop monitoring

Edge computer vision is also crucial in precision agriculture as it enables the monitoring of crops and the optimisation of agricultural techniques. Unmanned aerial vehicles (UAVs) equipped with advanced electronics and cameras capture live images of agricultural fields. Computer vision algorithms then analyse these photographs to assess the condition of crops (identifying overall health and areas of stress, disease, or nutrient deficiencies), detect pests (enabling early intervention and targeted treatments), and optimise watering (based on soil moisture levels and crop health indicators). This empowers farmers to make decisions based on data, resulting in higher crop production and better use of resources.

Dang et al. (2020) introduced an innovative method for the automatic detection and classification of diseases in radish fields using UAVs. It uses UAV-mounted cameras to capture high-quality field images, which are analysed by combining colour and texture feature extraction. The method employs K-means clustering to segment radish regions, followed by the use of a fine-tuned GoogleNet to detect early stages of Fusarium wilt. This offers a rapid and accurate alternative to manual inspection methods.

To identify and categorise insect pests in crops, Albattah et al. (2023) introduced an automated system based on UAVs to address the challenges of manual inspection and the need for timely pest management. They employed a lightweight UAV and a custom CornerNet (Law and Deng 2018) with DenseNet100 (Huang et al. 2017) as the foundational network. Specifically, the method involves three stages: acquiring the region of interest through sample annotations for model training, applying DenseNet100 for deep keypoint computation in the custom CornerNet, and finally employing the CornerNet model to identify and categorise various insect pests. The DenseNet network enhances feature representation, helping CornerNet to effectively detect insect pests as paired key points. The method was evaluated using the standard IP102 benchmark dataset (Wu et al. 2019b), which demonstrated its effectiveness and accuracy in identifying target insects, thus providing an essential tool to strengthen crop management and food yield through timely pest detection and intervention.

Albuquerque et al. (2020) proposed a DL method designed to automatically identify water from aerial footage captured by UAVs using the Mask R-CNN (He et al. 2017a). The method aims to improve the inspection and maintenance of irrigation systems, potentially reducing time and costs. The ability to accurately detect water in image frames allows for early identification of malfunctioning irrigation nozzles, crucial to correctly implementing crop field hydration plans. Such malfunctions can lead to insufficient or excessive watering, compromising the effectiveness of irrigation plans. Preliminary results demonstrated the feasibility of using advanced UAVs and neural networks in smart irrigation systems, offering a promising solution to ensure crop quality and productivity amid increasing food demand.

The application of UAV-acquired multispectral and thermal infrared imagery in precision irrigation management in a Cabernet Sauvignon orchard was showcased by Zúñiga

Espinoza et al. (2017). They evaluated the efficacy of these images in assessing various subsurface irrigation configurations at different depths and rates compared to standard surface irrigation. While no significant differences in fruit yield were observed between different irrigation techniques (pulse versus continuous) or depths, the study did find significant yield variations caused by deficient irrigation. Strong correlations were seen between vegetation indices (NDVI, GNDVI) and canopy temperature with both fruit yield and leaf stomatal conductance, demonstrating the potential of UAV-acquired imagery in real-time crop stress assessment. This work emphasises the efficacy of using thermal imagery as a rapid tool to estimate leaf stomatal conductance, crucial to optimising irrigation schedules and enhancing water use efficiency in viticulture.

An innovative method combining UAVs, multispectral imaging, and YOLOv3 to evaluate phenotypic traits in citrus crops was introduced by Ampatzidis and Partel (2019). The method overcomes the limitations of traditional plant breeding evaluation by offering low-cost, automated, high-throughput phenotyping. Using YOLOv3, the method detects, counts, and geolocates trees and tree gaps, categorises trees based on canopy size, develops individual tree health indices, and evaluates citrus varieties and rootstocks. In a study involving 4,931 citrus trees, the method achieved high precision and recall rates of 99.9% and 99.7%, respectively, for tree detection and counting, 85.5% overall accuracy for canopy size estimation, and 100% precision and 94.6% recall for detecting and locating tree gaps. The method significantly advances genomic selection and cultivar development in agriculture. The authors did not further optimise the YOLOv3 model, but rather deployed it directly on UAVs, demonstrating the feasibility of using an existing efficient model on resource-constrained edge devices. This highlights the inherent balance between performance and computational efficiency of YOLOv3.

It is noteworthy that the effectiveness of edge deep learning methods in different agricultural environments may vary due to factors such as crop type, environmental conditions, and specific agricultural practices (McEnroe et al. 2022). For instance, using drones in dense crop canopies may require more sophisticated methods or higher resolution images to maintain detection accuracy (Su et al. 2023). Nevertheless, these edge deep learning methods generally contribute to enhanced productivity and resource efficiency. Moreover, these methods demonstrate the potential for adaptability and scalability of drone-based solutions across various agricultural settings, ranging from small farms to large-scale operations. Particularly, large-scale farms can leverage economies of scale to reduce deployment costs and further enhance operational efficiency through standardisation.

### 5.4 Retail and shelf monitoring

Retailers utilise computer vision technology to monitor shelves and manage inventory. Cameras are installed on shelves to capture photos in real-time. Computer vision algorithms analyse the photos to monitor stock levels, detect misplaced items, and generate warnings for restocking. This optimises inventory management and increases the overall shopping experience for customers. We discuss several notable studies that have explored the integration of computer vision for retail and shelf monitoring applications.

Lachhab (2023) proposed the use of edge computer vision techniques, particularly edge-based object counting models, to automate the traditional manual inventory management process. Automation includes capturing images of fruits and vegetables shelves, identifying

boxes and their categories, and then using DL counting models to estimate the number of items in each box. This process aims to optimise store operations through continuous monitoring and analysis. The study employed object detection and density estimation methods, evaluating object counting approaches in four data scenarios: supervised learning, semi-supervised learning, few-shot learning, and zero-shot learning. Key findings include the YOLO model (especially YOLOv5) performing well in supervised learning due to its balance between speed and size. In semi-supervised learning, the Efficient-Teacher method enhanced the performance of the YOLO model using limited labeled data. Zero-shot learning, particularly the CLIP-Count method, is recommended in environments with limited data but ample computational resources, striking a balance between speed and error rate.

Advances in edge computer vision were also utilised by Kanjula et al. (2022), who introduced a pioneering AI-based people counting system for retail analytics. This cost-effective solution calculates conversion rates by correlating the number of visitors with actual transactions, providing valuable insights for retail optimisation, such as security checks and intelligent queuing. The project utilised Intel's OpenVINO toolkit for optimising neural network inference, demonstrating how advanced sensor technology and edge AI can transform the traditional retail environment. This approach not only improved the accuracy of people counting but also aligned with the trend of integrating AI-driven solutions into retail to achieve growth equivalent to online data analysis capabilities.

Fan et al. (2021) presented a low-power, cost-effective DL method using small IoT cameras to monitor the shelf status in retail environments, termed as CMSS (Camera Monitoring Store Shelves). This method addressed the inefficiency of manual inspection for restocking shelves, crucial for maintaining turnover in retail settings. The system operates on ultra-low-power FPGA chips, ensuring minimal power consumption at milliwatt levels, with total system power consumption below 6 mW. Their experimental results demonstrated the effectiveness of the system, with a maximum lateral recognition distance of 40 centimeters on shelves of the same width. Compared to existing methods, this solution offers significant advantages, including lower cost, reduced power consumption, and ease of maintenance, making it a highly feasible option for large-scale retail applications.

A sophisticated mobile object recognition system tailored for retail environments utilising video frame detection was proposed by Mustafa and Sethi (2005). The system was specifically designed to monitor and detect activities such as the movement of shopping carts and the operation of cash drawers by analysing the boundaries of these objects. The system, manually trained through pre-recorded video sequences, reliably identifies specific actions, aiding in the surveillance of potential suspicious activities. It emphasises the importance of understanding the specific context of the monitored activities, involving either human or nonhuman entities, and adapts to various surveillance requirements, including single or multiple camera setups. This method provides retail store owners with a practical solution to effectively filter and analyse surveillance footage, thereby enhancing the security and operational oversight in retail environments.

To address out-of-stock (OOS) issues in retail supermarkets, Achakir et al. (2023) introduced an edge computer vision system based on Faster R-CNN and the monocular depth estimation model MiDas to identify OOS products by analysing images of store shelves. The integration of Affine-SIFT descriptors with Faster R-CNN improved the accuracy of detecting products, especially in challenging categories like coffee and wine. This automated method, tested in multiple stores, not only accurately detects stockouts but also facilitates

timely restocking, ultimately improving sales and customer satisfaction. The system demonstrated a significant shift from manual to automated inventory monitoring, highlighting the potential of DL technologies in optimising retail operations and inventory management.

De Simone et al. (2023) proposed ROSCH, a humanoid autonomous mobile robotic platform based on the robotic operating system (ROS) framework, designed to detect OOS items on supermarket shelves. Specifically, ROSCH utilises the humanoid robot Pepper, powered by the NVIDIA Jetson Xavier NX, to identify empty and partially empty shelves using the YOLOv6 model. YOLOv6 (Li et al. 2022) is a single-stage anchor-free object detection model known for its excellent generalisation capabilities and fast convergence, particularly when training data is limited. To further enhance inference speed, the researchers optimised the model using the TensorRT framework. Integrating navigation with YOLOv6, ROSCH autonomously navigates supermarket aisles and notifies human operators to expedite restocking. The ROSCH system was tested in a supermarket in Salerno, Italy, demonstrating impressive performance, with its OOS detection being eight times faster than manual detection. The system aims to reduce sales losses due to OOS situations and improve the overall efficiency of supermarket operations.

Edge deep learning technology offers numerous benefits for retail environments, such as real-time inventory monitoring, which helps optimise inventory management processes while reducing labour costs. However, the initial investment in infrastructure, including devices and software, can be substantial. Despite this, the potential return on investment (ROI) remains considerable, as these technologies can enhance sales through improved inventory management and customer insights, ultimately optimising store operations and increasing profitability (Dell, Accessed 26 July 2024a; Kellermayr-Scheucher et al. 2022). Furthermore, it supports unmanned retail by providing more accurate shelf monitoring and automated customer services, offering customers a more efficient and convenient shopping experience. (Xu et al. 2020)

### 5.5 Environmental monitoring and wildlife conservation

Edge computer vision is also used in the field of environmental monitoring and wildlife conservation. Wildlife is documented through the use of camera traps (i.e., remotely triggered cameras) placed in natural habitats, which noninvasively collect both images and videos (see Fig. 10 for a camera trap-based edge device setup used for monitoring wildlife and conserving natural habitats). Computer vision systems deployed at the camera traps process these data to monitor animal movements, recognise endangered species, and detect any abnormal behaviour. Real-time monitoring facilitates ecological studies and improves the efficacy of wildlife protection efforts.

Gotthard and Broström (2023) used edge computer vision for wildlife conservation by employing object detection and classification models on edge devices equipped with camera sensors to monitor endangered species and detect potential intruders, primarily in the Ngulia sanctuary in Africa. The study evaluated three object detection models (SSD, FOMO, MobileNetV2, and YOLOv5) on three different microcontrollers. A key innovation is the deployment of wireless updates, enabling edge devices in remote locations to collect field data and iteratively improve through an active learning pipeline. This approach demonstrates the feasibility and effectiveness of edge computer vision in real-world conservation efforts.

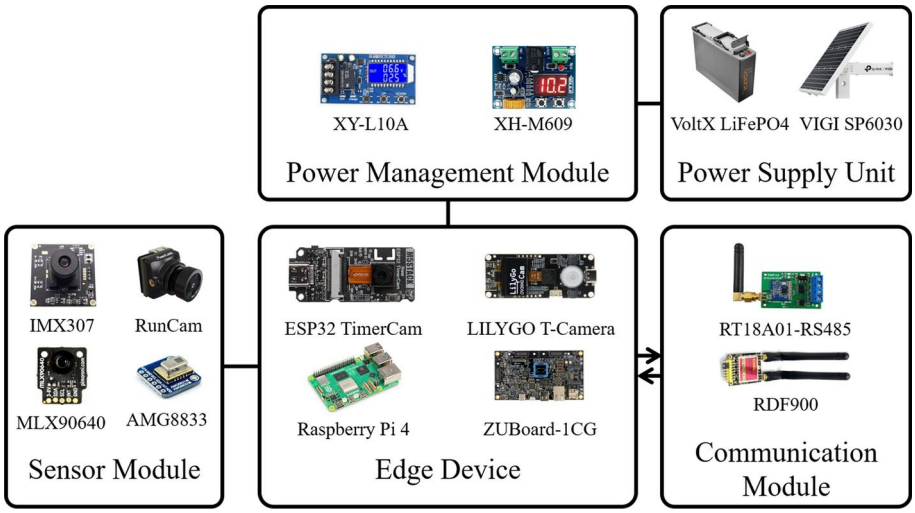


**Fig. 10** Schematic of an edge device for wildlife monitoring and conservation

Simões et al. (2023) developed a framework that employs MegaDetector, a computer vision model created by Microsoft's AI for Earth project, which is widely used in camera trap-based environmental monitoring and wildlife conservation. MegaDetector utilises a two-stage Faster R-CNN process to identify and classify objects in images by first detecting regions containing objects and then classifying these objects with confidence scores. The authors fine-tuned the Faster R-CNN model with an Inception-ResNet-v2 backbone to detect and classify 13 wildlife species relevant to the Parc National du Mercantour (PNM) in France. Under various environmental conditions, DeepWILD accurately detected, classified, and counted these 13 species. At an IoU of 0.5, the method achieved a mean Average Precision (mAP) of 96.88%, significantly enhancing the efficiency and accuracy of wildlife population estimation.

Focusing particularly on accurately determining the population density of deer in specific areas, Arshad et al. (2020) introduced a novel method for wildlife monitoring and counting. Their method, utilising CNNs, edge computing, and online tracking, addressed various challenges in automatic animal counting, such as wildlife movement, light fluctuation, and the issue of recounting the same animal in different images. The authors highlighted the shortcomings of traditional wildlife monitoring methods, including the high cost and logistical challenges of manual observation, and the limitations of trap cameras in manual data extraction and their inability to distinguish repeated appearances of the same animal. To overcome these issues, Arshad et al. proposed a video-based detection and tracking system capable of operating under various lighting conditions, accurately counting and tracking the movement of deer within the camera's field of view. This method not only enhances the accuracy of wildlife counting but also provides valuable data for understanding animal activities and behaviors, proving to have immense potential in conservation and ecological research.

Tulasi et al. (2023) presented an energy-efficient solution to enhance wildlife monitoring by deploying a lightweight network on existing camera traps. Specifically, their method employs the Raspberry Pi Zero 2W for on-site data processing and combines it with a detec-

tion and classification two-stage pipeline based on MobileNetV2. This approach offers timely wildlife detection and alerts, addressing major challenges of traditional camera traps, such as delayed data retrieval and processing. This not only improves the accuracy and efficiency of wildlife monitoring but also provides a scalable and sustainable solution for ecological research and conservation efforts, demonstrating its practical application through successful field deployment.

Another method for wildlife monitoring and analysis based on camera trap network captures was proposed by Gupta et al. (2022). Their method addresses the challenge of cluttered images in camera trap data, which typically leads to low detection rates and high false discovery rates in animal monitoring. To address this, the authors leveraged a camera trap database containing candidate animal proposals generated using multilevel graph cuts in the spatiotemporal domain. These proposals were then validated to differentiate animals from the background. The core of their method is the use of self-supervised CNNs to develop an animal detection model. Their extensive experimental results show that the proposed detection model achieves a high precision on standard camera trap datasets.

Focusing on the needs of animal conservation, particularly for endangered elephants in regions like India, Verma and Gupta (2018) presented a solution using deep learning models to prevent human-elephant collisions (HEC). They developed an automated early warning system based on visual cues, employing transfer learning models such as ResNet50, MobileNet, and InceptionV3. These models were tested on a comprehensive dataset, showing high accuracy in detecting elephants, potentially preventing tragic incidents on forest-crossing railway tracks. This study offers a low-cost, accurate, and efficient method for elephant detection, aligning with efforts to mitigate the ecological imbalance caused by increasing wildlife mortality rates.

In conclusion, edge deep learning offers innovative and effective tools for wildlife monitoring and conservation. Various studies have demonstrated its potential in enhancing the accuracy, efficiency, and scalability of monitoring systems, particularly in remote or challenging environments. However, ethical considerations and the potential impacts on wildlife behaviour must be taken into account to ensure that conservation efforts do not inadvertently harm the species they aim to protect. Noninvasive techniques, such as camera traps, can minimise human disturbance, thereby avoiding stress or changes in animal behaviour and movement patterns.

### 5.6 Smart cities and public safety

Edge computer vision also plays a significant role in improving public safety and urban planning within the framework of smart cities. Surveillance cameras, equipped with edge devices, employ real-time video stream analysis to identify abnormal behaviours, monitor traffic patterns, and improve overall security. The decentralised model guarantees prompt reactions to crises and facilitates municipal planning using real-time data. Several studies have explored the diverse applications and advantages of deploying edge computer vision systems in urban environments.

Avvenuti et al. (2022) introduced a neural network based on spatiotemporal attention for improving automatic people counting in urban smart city environments using surveillance video. Their method effectively counts and precisely locates people within video frames, embodying significant advancements in practical urban surveillance and crowd monitoring

through DL technologies. Leveraging the temporal correlation between video frames, it significantly reduces errors on the FDST benchmark through self-attention connectives and attention-based temporal fusion layers.

Utilising cost-effective off-the-shelf hardware equipped with computer vision capabilities, Di Benedetto et al. (2022) presented an embedded system for monitoring human activities and ensuring safety in critical environments, such as in the context of health emergencies or hazardous workplaces. Their DL-based embedded system performs tasks like people counting, social distance measurement, and personal protective equipment detection. Developed in response to challenges posed by the COVID-19 pandemic, the system operates both indoors and outdoors, reducing the need for manual supervision. Its effectiveness was validated through two novel datasets, one containing images from public squares in Pisa, Italy, and another featuring images with and without personal protective equipment. The results indicate that the system can accurately monitor compliance with safety rules, providing a practical solution for real-time safety monitoring in various scenarios.

Addressing the challenge of real-time facial recognition on resource-constrained devices, Deng et al. (2023) developed an efficient, compact DL model inspired by FaceNet. The model employs one-shot or few-shot learning to achieve effective feature embedding. The study demonstrated the model's effectiveness in recognising occluded faces using gray-scale input images in uncontrolled environments, making it suitable for real-time embedded applications like entrance security systems in urban settings.

Raj et al. (2023) presented a study on how drones and computer vision technologies could improve the quality of life for visually impaired persons (VIPs) in smart cities. They explored the potential of these technologies to assist VIPs in navigating both indoor and outdoor environments, enhancing their mobility and safety. Given the large number of people affected by visual impairments globally, the study emphasises the need for innovative solutions to aid VIPs. With the rapid development of smart city infrastructures, the authors suggest drones equipped with advanced computer vision and navigation systems as a promising solution. These drones could serve as mobile assistive tools, replacing traditional methods like guide dogs, and offer additional features like hazard detection and collision avoidance. This approach highlights the integration of emerging technologies into urban environments, aiming to make cities more inclusive and convenient for VIPs.

A significant study on the urban heat island effect in the rapidly urbanising Pearl River Delta region in China was presented by Chen et al. (2006). Focusing particularly on the impact on regional climate and socio-economic development, they analysed land use/cover types and brightness temperature using satellite imagery, introducing the Normalised Difference Bareness Index (NDBaI) for improved land analysis. The study specifically examined the rapidly developing city of Shenzhen to understand how its expansion affects temperature distribution and the urban heat island effect. The study provides valuable insights into the impact of urbanisation on local climate, offering crucial data for urban planning and mitigating the urban heat island effects of rapidly developing cities.

Cao et al. (2023) proposed an improved, lightweight, real-time detection algorithm for drone imagery, addressing the challenges of detecting small objects and reducing computational costs. The algorithm combines MobileNetV3 with the ECA attention mechanism and uses additional prediction heads to enhance small object detection. The algorithm showed outstanding performance and efficiency, making it a valuable tool for drones performing disaster search and rescue missions in urban areas.

Muhammad et al. (2018a) proposed a CNN-based fire detection method for surveillance videos. This approach leverages DL to extract features from video frames, identifying early signs of fire. Subsequently, the authors optimised the method using MobileNet, improving inference speed and computational efficiency while reducing model parameters and memory usage, making it more suitable for resource-constrained edge devices Muhammad et al. (2018b). The optimised method maintains high detection accuracy, significantly enhances real-time performance, and broadens its applicability. Experimental results demonstrate its effectiveness in various surveillance environments, quickly detecting flames and smoke during the early stages of fire, thus providing critical time for rescue operations.

### 5.7 Sports analytics

In sports analytics, edge computer vision is increasingly utilised to provide real-time insights during events. Edge devices process live footage captured by cameras installed in stadiums to generate performance metrics, analyse game dynamics, and monitor player movements. This information improves the overall sports experience for coaches, analysts, and even spectators, as illustrated by several recent studies.

Cui and Hu (2022) introduced a federated learning algorithm and a lightweight neural network-enhanced distributed computing model to address the limitations of centralised cloud computing in processing large-scale video surveillance data. This model processes video data at the edge, reducing transmission load and latency. The study also explored human action recognition, particularly in Taekwondo, using behavior detection algorithms combined with sensors to improve training quality. This approach aims to strengthen Taekwondo training methods and ensure safer practice routines. The study explores the feasibility and reliability of distributed computing and DL to sports training and intelligent video surveillance systems.

Focusing on improving the localisation of human-generated events, Cioppa et al. (2020) introduced a novel loss function for action discovery in sports videos, especially football. Their method considers the temporal context around actions and, when applied to SoccerNet, achieved a substantial improvement over the baseline and was applied to ActivityNet for general activity detection. Additionally, the method not only enhanced action recognition in SoccerNet but also demonstrated its versatility and effectiveness in broader activity localisation tasks. This provides advanced tools for deeper match analysis and automated content creation.

Addressing the limitations of monocular 3D human pose estimation (HPE) methods, Baumgartner and Klatt (2023) proposed a novel method for extracting 3D kinematic data from 2D sports videos. Their research introduced partial motion field registration for precise camera calibration, crucial for kinematic analysis in sports like middle and long-distance running. By generating a synthetic dataset in Unreal Engine 5, the study provided new benchmarks for evaluating 3D HPE technologies, paving the way for advanced, large-scale human motion analysis from existing video sources.

Van Hooren et al. (2023) performed markerless motion capture using computer vision technology to analyse running techniques. They evaluated the accuracy of DeepLabCut and OpenPose in tracking sagittal plane hip, knee, and ankle kinematics during running and compared them to a marker-based gold standard system. They emphasised the potential of computer vision in enhancing the reliability and effectiveness of running technique analysis.

A four-point camera calibration method for sports video capture was proposed by (Zhang and Izquierdo 2023). Their method uses conditional generative adversarial networks (cGANs) to generate semantically segmented video frames and then estimates four key points from a single segmented frame using a regression network. Evaluation on various datasets, including the 2014 FIFA World Cup and National Basketball, showed that this method is superior in accuracy and computational efficiency, and suitable for real-time applications in sports video camera calibration.

## 6 Medical edge applications

The application of edge computing technology in computer vision is becoming increasingly mature, demonstrating the unique advantages of edge computing in various scenarios. Recent advancements in edge computing within nonmedical domains have also impacted medical diagnostics, particularly through the adoption of optimised models and hardware initially designed for general computer vision tasks. For example, the YOLO series, as a highly efficient object detection algorithm, widely used in surveillance and autonomous driving, has now been successfully applied to medical tasks, such as polyp detection during colonoscopies (Ou et al. 2021; Redmon et al. 2016). Similarly, MobileNet, known for its efficiency on edge devices, has shown promising results in diagnosing diseases from medical images, achieving reduced latency without sacrificing accuracy (Sait et al. 2019; Goceri 2021b).

On the hardware side, edge devices such as the NVIDIA Jetson series and Raspberry Pi have proven capable of supporting medical diagnostic tasks that require real-time processing and low power consumption. For instance, the Jetson series has been utilised in computer vision-based frameworks for detecting skin and cervical cancers (Shrivastava et al. 2023; Wang et al. 2019b). Additionally, several approaches have integrated Raspberry Pi, smartphones, and lightweight networks for chest CT and dermatology detection, offering a low-cost and efficient diagnostic solution (Mieras et al. 2018; Dai et al. 2019; Raghavan et al. 2020; Masud et al. 2020; Goceri 2021b). The successful deployment of these devices highlights the potential of edge computing in medical scenarios, enabling efficient processing and diagnostics while delivering reliable results, even under resource-constrained conditions.

The exploration of nonmedical applications in the previous section underscores the potential of edge deep learning technologies. Autonomous driving and intelligent traffic control utilise edge devices to process sensor data locally, enabling real-time decision-making and adaptive responses. These methodologies, developed for efficient real-time processing, can be applied to medical diagnostics and emergency care, where immediate analysis of medical images or patient data is critical. Furthermore, sensor fusion techniques can assist in the multimodal data analysis and processing required in medical scenarios. Additionally, the long-term, low-power monitoring solutions used in agriculture can be adapted for continuous medical monitoring and remote diagnostics. In retail, edge deep learning supports intelligent inventory management and real-time customer interaction. Similarly, these methods can improve drug inventory management in medical facilities, ensuring efficient healthcare operations.

In summary, insights gained from nonmedical domains can inspire new approaches and improve existing methodologies in medical diagnostics, paving the way for a more integrated and effective use of edge deep learning in healthcare. Here we discuss various examples of medical edge computing (Fig. 11), including for the diagnosis of gastrointestinal, pulmonary and thoracic, and dermatological diseases, as well as for pathological images analysis and telemedicine applications, with a focus on ensuring real-time and efficient diagnostic services.

### 6.1 Gastrointestinal diagnosis

Gastroenterological diagnostics has benefited from advances in computer vision and edge computing technologies, as these assist in the real-time analysis and interpretation of endoscopic images for early detection, treatment, and monitoring of gastrointestinal diseases. Currently, gastroenterological diagnostics primarily rely on computer-aided diagnosis (CAD) systems, deployed on intermediate edge devices, primarily identifying and classifying anatomical markers through analysis of endoscopic images.

Accurate polyp detection is crucial for the early identification of adenomas and for mitigating the risks associated with cancer progression. However, traditional colonoscopy examinations have a polyp miss rate as high as 25% (Corley et al. 2014). Factors such as the varying skill levels of gastroenterologists, along with physical and mental fatigue, contribute to these oversights, leading to significant variability in adenoma detection rates (ADR) among practitioners (Kumar et al. 2017; Leufkens et al. 2012). CAD systems employing

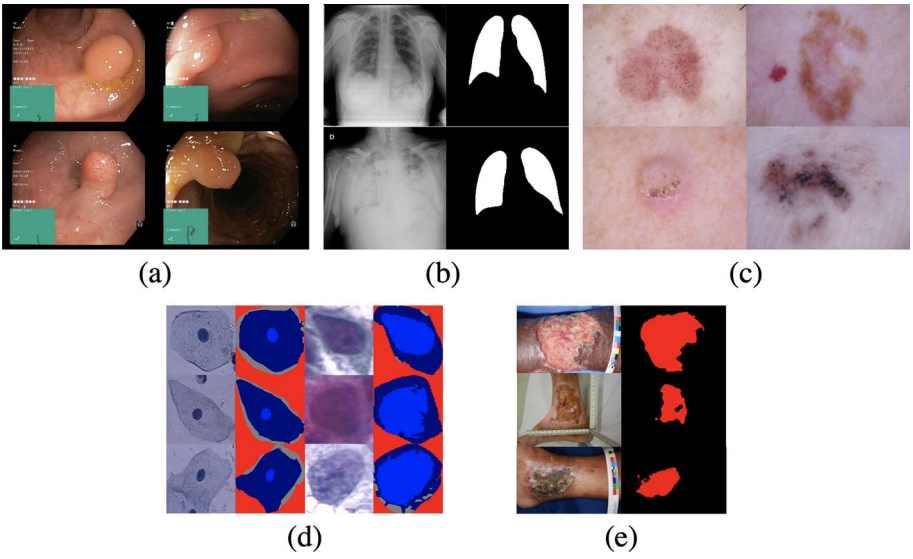


**Fig. 11** Examples of edge applications in the medical field based on computer vision. **a** Detecting intestinal polyps through endoscopes and DL-based computer-aided diagnosis (Urban et al. 2018). **b** Rapid screening method for COVID-19 deployed on Intel/Movidius Neural Compute Stick 2 (Liu et al. 2023b). **c** Fast identification of melanoma through mobile edge devices (Dai et al. 2019). **d** Detection of cervical cancer cells through embedded edge devices (Wang et al. 2019b). **e** Detection of chronic skin ulcers using portable edge devices (Chino et al. 2020)

DL algorithms can reduce the polyp-miss rate, particularly for endoscopists with lower ADR (Wang et al. 2019a, 2020b; Goceri 2024).

However, a challenge in CAD system-based polyp detection is the difficulty of deploying models on resource-constrained endoscopic hardware for real-time clinical prediction. DL models are typically trained and tested on dedicated deep learning accelerators, making it impractical to deploy them directly on medical diagnostic devices. Medical diagnostic devices often suffer from low performance due to weaker DL accelerators or the absence of specialised AI accelerators. For models that are difficult to deploy on low-performance devices, forced deployment resulting in low frame rates per second (FPS) still severely affects the efficiency of CAD. Therefore, from a systems perspective, lightweight models play a crucial role in facilitating high-accuracy and real-time polyp diagnosis.

A representative approach is PolypSeg+ (Wu et al. 2022), which utilises a lightweight architecture to enhance real-time application in clinical settings. It incorporates Adaptive Scale Context (ASC) and Efficient Global Context (EGC) to improve feature differentiation and detail preservation. This method effectively addresses several challenges in polyp segmentation, such as significant variations in polyp size and shape, low contrast between polyps and surrounding tissue, and blurred polyp boundaries. The model has been tested on the Kvasir-SEG (Pogorelov et al. 2017) and CVC-EndoSceneStill (Vázquez et al. 2017) datasets, showing that PolypSeg+ outperforms existing models in both speed and accuracy. Additionally, Ou et al. (2021) have proposed a lightweight network based on YOLOv5. By targeting optimisations for endoscopic video inputs, the number of convolutional kernels is halved, and unnecessary large object detection heads are removed. The model achieves accuracy close to that of YOLOv3-spp, yet its size is only 1/30 of YOLOv3-spp.

Although wireless capsule endoscopy (WCE) is a crucial method for small-bowel investigation, its time consuming and tedious nature poses challenges for physicians (ASGE Technology Committee et al. 2013). Leenhardt et al. (2020) introduced CAD-CAP, a large multicentre database designed for the development of CAD tools for WCE image reading. It includes 100,000 annotated training images and 20,000 test images. Seguí et al. (2016) designed a lightweight CNN achieving a high accuracy of 96% on CAD-CAP for the precise classification of nonpathological image features in the intestines, such as the wall of the bowel, bubbles, turbid substances, folds, and transparent spots.

Sahafi et al. (2022) introduced a capsule endoscopy device equipped with a Kendryte K210 chip, which enables real-time onboard analysis of gastrointestinal images through bidirectional communication with personal electronic devices such as smartphones or tablets. This innovation enhances the diagnostic efficiency of gastrointestinal diseases by employing a lightweight DNN to immediately identify abnormalities such as lesions or polyps during the examination. Moreover, the device also allows for task modifications and updates to the neural network via Bluetooth.

The Olympus EndoCapsule 10 System[40] sets a new standard for capsule endoscopy technology, integrating efficient image processing techniques with a user-friendly operating system. Equipped with a DL-based CAD system, Olympus ENDO-AID,[41] it leverages advanced optical and computer vision technologies to deliver clear, detailed gastrointestinal

[40] ENDOCAPSULE-10-System.https://www.olympus.com.au/medical/en/Products-and-Solutions/Products/Product/ENDOCAPSULE-10-System.html.

[41] ENDO-AID.https://www.olympus.com.au/medical/en/Products-and-Solutions/Products/Product/ENDO-AID.html.

images and evaluation reports, aiding physicians in the detection and assessment of various lesions, including inflammation, bleeding, or tumors in the small intestine. The Endo-Capsule 10 features a highly sensitive camera capable of automatically adjusting lighting within the patient's body, ensuring high-quality images under any condition. Furthermore, the system provides an intuitive data management platform, supporting efficient review and analysis of image data by physicians, thereby further optimising the diagnostic process.

Volume Laser Endoscopy (VLE) offers wide-field imaging and can scan the esophageal wall layers up to 3 mm in depth at near-microscopic resolution, commonly used to detect and evaluate Barrett's Oesophagus (BE) (Swager et al. 2017). VLE requires endoscopists to examine a large volume of image data in a short time (1,200 images from a 6 cm segment of the oesophagus in 90 seconds), which can be challenging. To address this, Redmon et al. (2016) proposed a miniaturised YOLO architecture based CAD method. Using partial residual networks achieved a detection accuracy of 98.23% on resource-limited devices and reduced inference speed by a factor of 10 compared to standard YOLO (Selmanaj et al. 2021).

Furthermore, Le et al. (2022) explored the potential of CNNs in classifying anatomical landmarks in upper gastrointestinal endoscopic images on edge devices. By employing quantisation techniques, they managed to reduce the memory and computational demands of the model, facilitating real-time diagnostics. Also, as the model allows processing more images per second, it potentially enables increasing the diagnostic accuracy. This work underscores the advantages of adopting lightweight techniques on edge devices to aid in medical diagnosis and treatment planning in gastroenterology.

Kara et al. (2023) developed an intelligent handheld edge device equipped with a unique tactile sensing module and a dual-stage learning algorithm. This device is specifically designed for on-site diagnosis and histopathological assessment of resected colorectal cancer polyps, offering immediate insights into their texture and hardness. The practicality of this device demonstrates the transformative impact of edge computing in bridging the gap between data collection and interpretation in gastrointestinal diagnostics.

### 6.2 Pulmonary and thoracic diagnosis

Diagnosis of lung and chest diseases are typically based on the analysis of chest X-ray (CXR) and CT scans. CXR imaging is the method of choice for screening of pneumonia, tuberculosis, emphysema, rib fractures, and cardiac abnormalities. Chest CT scans provide more detailed imaging for complex or difficult-to-diagnose cases, such as early-stage lung cancer, small lung nodules, pulmonary embolism, lung infections, and abnormalities of the thoracic organs. Local CAD systems not only enhance the confidence and accuracy of image interpretation, but also reduce the time taken for reading images. Moreover, in situations with limited resources or the need for rapid response, edge computing technology demonstrated advantages in quickly identifying diseases and promoting timely intervention.

Pneumonia, an acute disease, often requires rapid diagnosis and intervention. Paluru et al. (2021) proposed Anam-Net, a lightweight CNN to segment anomalies in COVID-19 chest CT images, which can be deployed on embedded systems such as Raspberry Pi 4, NVIDIA Jetson Xavier, and Android applications for mobile devices. Hou et al. (2020) developed a method that employs subspace learning techniques for analysing CXR scans in edge devices, showing a significant reduction in computational and memory demands.

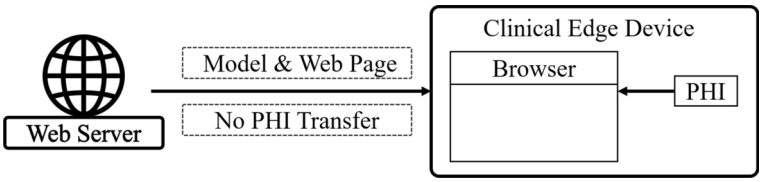


**Fig. 12** Browser-based pulmonary and thoracic diagnostics on clinical edge devices with secure patient health information (PHI) handling. The figure is adapted from the methodology presented in Dong et al. (2023)

This underscores the potential of subspace learning to facilitate efficient image analysis for pulmonary diagnosis on edge devices.

Besides, Abubeker and Baskar (2023) introduced B2-Net, a model designed to differentiate among normal pneumonia, bacterial pneumonia, and viral pneumonia in chest X-ray images. This approach leverages multiple ImageNet pretrained models integrated through Depth-Wise Convolution (DWC) and Squeeze-and-Excitation techniques, employing transfer learning for fine-tuning the models. The architecture of B2-Net is optimised for deployment on the Jetson Nano, offering rapid and accurate diagnostic support.

In addition, Sait et al. (2019) proposed a smartphone-based app for the preliminary detection of pneumonia using CXR images. The app uses a MobileNets model trained on a large dataset of CXR images from known pneumonia cases for detection. The app can quickly conduct a preliminary analysis of CXR images in conjunction with the camera of smartphone. In addition, the app includes an electronic diagnostic feature, allowing users to seek advice from qualified physicians based on the results. It also incorporates a respiratory pattern recorder module to enhance the app's predictive accuracy.

Furthermore, EdgeMedNet proposed by Liu et al. (2023b) is a lightweight U-Net architecture designed for efficient segmentation of medical images on edge devices. Tested on various thoracic imaging datasets, the framework demonstrated compelling performance in segmentation and classification tasks for pulmonary and thoracic diagnostics. Furthermore, to address clinical data security constraints, Dong et al. (2023) integrated serverless edge computing with a browser-based medical imaging application (Fig. 12). This integration aims to facilitate real-time image analysis while ensuring data security, which is crucial to the advancement of efficient and secure pulmonary and thoracic diagnostic methodologies on edge devices. France and Newman (2020) introduced a method employing cluster neural networks to harness the computational capabilities of edge devices to analyse thoracic imaging data. The method showcased an advancement in edge intelligence, enabling more sophisticated image analysis for pulmonary and thoracic diagnostics.

Regular low-dose CT screening for lung cancer can reduce the mortality rate by 20% in high-risk groups (National Lung Screening Trial Research Team 2011; Kauczor et al. 2015). However, the number of CT scans can overwhelm radiologists, raising the false-positive rate. Thus, DL-based CAD may provide an alternative. (Masud et al. 2020) proposed a lightweight CNN for mobile devices which achieved an impressive accuracy rate of 97.9%. Additionally, Raghavan et al. (2020) introduced a mobile low-dose CT screening device, aimed at improving opportunities for lung cancer screening, especially for uninsured and underinsured patients. The device, equipped with high-quality CT scanners and wireless internet connectivity, achieved lung cancer detection rates comparable to established trials

such as the National Lung Screening Trial (NLST), even for patients not eligible for medical insurance. The study demonstrated its efficacy in reaching socio-economically disadvantaged groups, offering opportunities to detect early-stage lung cancer and increase survival rates at a lower cost per case. This mobile device represents a significant advancement in healthcare accessibility and has the potential to bring changes in lung cancer screening policies.

Several commercial edge devices have been developed specifically for pulmonary and thoracic diagnosis. GE Healthcare leveraged the Intel Distribution of the OpenVINO toolkit[42] to improve the performance of pneumothorax detection algorithms in CXRs.[43] GE Healthcare also developed the portable X-ray device Optima XR240amx,[44] which performs real-time diagnostics using the multimodal DL-based Critical Care Suite.[45] This suite comprises multiple DL models that work collaboratively to automatically detect and prioritise critical conditions such as pneumothorax. Specifically, the detection models focus on identifying and marking key abnormalities in the images, such as pneumothorax and lung nodules, while the segmentation models further isolate these abnormal regions, generating precise contours and area information. Moreover, the suite provides confidence scores for the detection and segmentation results, aiding clinicians in assessing the reliability of the diagnostic outcomes.

In addition, Butterfly Network[46] introduced a handheld ultrasound device that, through DL algorithms, offers various scanning modes and provides real-time image analysis and interpretation, enabling nonexpert users to perform ultrasound examinations.

### 6.3 Dermatological diagnosis

The skin, as the largest organ of the body, serves as a vital barrier against environmental hazards like microbes, viruses, and pollutants. Skin diseases affect individuals across all age groups, stemming from various causes such as hereditary factors, lifestyle choices, and exposure to environmental elements. Among the prevalent skin disorders are acne, skin cancer, seborrheic keratosis, psoriasis, melanoma, and vitiligo. Given the ongoing and prevalent nature of skin diseases, their negative effects can significantly impair both the physical and psychological well-being of those afflicted.

Additionally, diseases such as melanoma can lead to severe consequences if not treated promptly. Diagnosing skin diseases from clinical images is particularly challenging due to the complexity, diversity, and similarity of these conditions. Moreover, manual diagnosis by medical experts is both time-consuming and subjective. Thus, accurate and timely diagnosis is essential for effective treatment and health management. The integration of edge computing in dermatological diagnostics has ushered in a new era of efficient and real-time analysis

[42] OpenVINO™ Toolkit. https://www.intel.com/content/www/us/en/developer/tools/openvino-toolkit/overview.html.

[43] Intel and GE Healthcare Partner to Advance AI in Medical Imaging. https://www.intel.com/content/www/us/en/customer-spotlight/stories/ge-healthcare-medical-imaging.html.

[44] Optima XR240amx. https://www.gehealthcare.com/products/radiography/mobile-xray-systems/optima-xr240.

[45] Critical Care Suit.https://www.gehealthcare.com/en-ph/products/radiography/mobile-xray-systems/critical-care-suite.

[46] Butterfly iQ3. https://www.butterflynetwork.com.

of skin images (Göçeri 2020; Goceri 2021a). The ability to process and analyse data on edge devices close to the data source facilitates timely diagnostic feedback, thereby improving patient outcomes and optimising treatment strategies. We delve into several methods and frameworks that leverage edge computing for dermatological diagnosis.

Goceri (2021b) developed a lightweight network based on an improved MobileNet architecture and a hybrid loss function, specifically designed for efficient image classification on resource constrained devices such as smartphones. It enables mobile applications to diagnose five common skin diseases, including seborrheic dermatitis, rosacea, hemangioma, psoriasis, and acne vulgaris, with high accuracy. Additionally, Mieras et al. (2018) addressed the high prevalence of skin diseases in resource-poor areas by developing SkinApp, a mobile app intended to assist healthcare workers in diagnosing and managing these diseases, including neglected tropical diseases (NTDs) with skin manifestations. Their research highlights the lack of healthcare personnel trained in dermatology in these areas and the potential of mobile health technology to bridge this gap.

In addition, Shrivastava et al. (2023) proposed a lightweight model for distinguishing between benign and malignant skin lesions, tailored for the Jetson Nano platform. Utilising transfer learning from pretrained networks such as ResNet50 and MobileNet, this model achieves a classification accuracy of up to 93.3% on the PH2 skin lesion dataset (Senan et al. 2021). This approach demonstrates the integration of lightweight models with compact hardware platforms to provide portable and efficient healthcare solutions.

Dai et al. (2019) presented an app for mobile devices for rapid detection and diagnosis of skin diseases. Furthermore, they highlighted the limitations of cloud-based ML, such as privacy concerns and latency. The authors deployed a pretrained neural network on mobile devices to ensure that all computations are performed locally, thereby reducing latency, saving bandwidth, and enhancing privacy. This method underscores the advantages of mobile edge computing in the diagnosis of skin diseases.

To facilitate early skin cancer detection in resource limited rural areas of developing countries, Ngeh et al. (2020) developed a low-cost, user-friendly, and internet-independent prescreening device capable of classifying skin abnormalities and performing region segmentation. The device, powered by a Raspberry Pi and a CNN trained on the Skin Cancer MNIST dataset (Tschandl et al. 2018), offers a practical solution for remote skin cancer assessment. Providing prescreening to identify high-risk individuals, it optimises the allocation of medical resources and improves early detection in underserved areas.

Transitioning to more advanced frameworks, Shi et al. (2023) explored a federated contrastive learning framework for automatic skin lesion diagnosis. The proposed federated learning approach adeptly navigates the challenges of data silos while enhancing the model's diagnostic generalisability to unseen data. This work stands as a testament to the potential of federated learning in conjunction with edge computing to deliver robust and privacy-preserving dermatological diagnostic solutions.

The VISIA Complexion Analysis system[47] by Canfield Scientific integrates various imaging and DL technologies to provide a comprehensive analysis of skin health. Through RBX technology,[48] VISIA conducts in-depth inspections of wrinkles, spots, pores, textures, and subdermal issues that are invisible to the naked eye. The system not only aids in

[47] VISIA. https://www.canfieldsci.com/imaging-systems/visia-complexion-analysis.

[48] Canfield Scientific RBX® Technology Overview. https://www.canfieldsci.com/research/stories/white-paper-rbx-technology-overview.

detecting and quantifying various skin conditions to develop targeted treatment plans but also enhances patient engagement through personalised reports and progress tracking. As a modular and adaptable platform, VISIA allows for ongoing updates and feature integrations, making it a crucial tool for comprehensive dermatological diagnosis and personalised skincare solutions.

### 6.4 Pathological images analysis

Microscopic examination of tissue slides (histopathology) is considered the gold standard for cancer diagnosis and prognosis. This process requires pathologists to identify subtle histopathological patterns within highly complex tissue images. It is time-consuming, subjective, and prone to considerable inter-observer and intra-observer variability. Hematoxylin and eosin (H &E) staining is the most popular method for tissue staining. With the advent of whole-slide imaging (WSI) technology, a large number of H &E stained tissues can be scanned, digitally represented, and archived. The analysis of WSIs in pathology using CAD systems is becoming a routine clinical practice. We focus on lightweight yet powerful models and introduce groundbreaking methodologies that have successfully harnessed edge computing for efficient and accurate analysis of histopathological images.

Auguste and Palsana (2015) developed an economical method for pathology in resource limited settings. Their method uses a standard optical microscope, an iPhone 5s, and a custom 3D-printed adapter to capture high-quality WSIs. Compatible with various stains including H &E, this method is particularly cost-effective and portable, making it a viable option for medical facilities in developing countries where traditional WSI systems are prohibitively expensive. This innovation has significant potential to enhance diagnostic capabilities and improve healthcare outcomes in underserved areas.

Similarly, Ramey et al. (2011) evaluated the feasibility of interpreting WSIs in pathology, particularly frozen sections (FS), using a mobile viewing device (MVD). Their study involved scanning FS samples and assessing them on an iPad using the Interpath app. The study found a high concordance of 89% between the initial FS diagnosis and the diagnosis made using the iPad, with the latter taking an average time of less than 2 min per slide. Set against the backdrop of increasing digitisation in the medical field, this study highlights the potential of using MVDs like iPads for the analysis of WSI images.

For the classification of histopathological images, Datta Gupta et al. (2023) introduced ReducedFireNet. This lightweight model stands out for its compact size (merely 0.391 MB) and low computational demand (0.201 GFLOPS), making it exceptionally suitable for edge devices with limited processing capabilities. Despite its lightweight nature, the model achieved an impressive average accuracy rate of 96.88% and an F1 score of 0.968 on the Malignant Lymphoma datasets, as per the settings of reference (Orlov et al. 2010). This dataset comprises 374 histopathological images, including types such as Lymphocytic Leukemia (CLL), Mantle Cell Lymphoma (MCL), Follicular Lymphoma (FL) stained with hematoxylin and eosin (H &E). The success of ReducedFireNet demonstrates its potential in facilitating timely and accurate disease diagnoses, especially in scenarios where rapid and efficient image analysis is critical.

To enable the detection of cervical cancer cells on embedded devices, Wang et al. (2019b) proposed a lightweight network incorporating optimised convolution operations, model parameter compression, and enhanced feature expression depth in the network struc-

ture design. It achieves a 94.1% accuracy rate on the NVIDIA Jetson TK1 embedded device, with fewer model parameters compared to ResNet18 and MobileNet.

There exist several edge devices employing DL for histopathological and cytopathological analysis. Cell Metric X,[49] a high-resolution imager driven by DL technology, enhances the efficiency and accuracy of cell line development processes. Maestro TrayZ's real-time cell impedance monitoring technology[50] provides critical data for cell health and function, and offers real-time, noninvasive monitoring, widely used in cytotoxicity and cell growth studies. Digital pathology tools have radically changed nephropathology by improving the accuracy of disease detection and classification. zenCELL owl[51] is a compact automated microscope for monitoring cell cultures, essential for accurate and reliable cell analysis. And CytoPAN,[52] a portable image cytometer, shows exceptional diagnostic precision in identifying breast cancer subtypes, especially ER/PR and HER2 (Min et al. 2020).

### 6.5 Telemedicine

With the rapid advancements in internet-of-things (IoT) and information and communication technologies (ICT), telemedicine is emerging as a promising healthcare service model, gradually becoming a vital component of the global healthcare sector (Azimi et al. 2018; Ye et al. 2023). Telemedicine, powered by edge computing, presents an alternative to traditional diagnostic and preventive measures. In recent years, the evolution of 5G technology and mobile health applications has unveiled the potential of telemedicine in chronic disease management and remote diagnostics. We survey how edge computing empowers telemedicine, enhancing the accessibility and efficiency of healthcare services.

Van Netten et al. (2017) embarked on an innovative exploration of telemedicine capabilities, particularly in the context of managing diabetic foot ulcers (DFUs). They used images captured by iPhones for remote assessment of DFUs, carefully evaluating the accuracy and reliability of such digital diagnoses compared to traditional on-site clinical assessments conducted by experienced podiatrists. The study integrated a comprehensive analysis of clinical features and treatment decisions, and examined the key limitations of mobile image-based DFU assessments. Subsequently, Yap et al. (2018) introduced an iPad-based DFU diagnostic app, FootSnap, and tested its reliability across different operators and patients. The results indicated high intra-operator and inter-operator reliability, with Jaccard similarity index values ranging from 0.89 to 0.94, demonstrating that FootSnap can effectively monitor the condition of diabetic feet longitudinally.

Furthermore, Goyal et al. (2018) proposed a DFU detection and localisation method based on Fast R-CNN, compiling a dataset of 1,775 DFU images annotated by medical experts and optimising the model through transfer learning. The authors showed that the model can achieve 91.8% average precision on an NVIDIA Jetson TX2 with an inference speed of 48 ms per image. This demonstrates the effectiveness and practical potential of DL approaches for DFU diagnosis on medium-range devices.

[49] Cell Metric® X. https://www.aicompanies.com/cell-line-development/cell-metric-x.

[50] Maestro TrayZ. https://www.axionbiosystems.com/products/cell-analysis/maestro-trayz.

[51] zenCellowl. https://zencellowl.com.

[52] CytoPAN. https://getzpharma.com/product/cytopan.

To support remote medical practices in managing chronic skin ulcers, Cazzolato et al. (2021) launched the UTrack mobile app, which facilitates remote wound image capture, segmentation, measurement, and monitoring. It leverages an innovative unsupervised segmentation method, UTrack-Seg, using mobile device cameras to accurately segment and measure ulcers while storing data locally to protect privacy. The app demonstrates superiority in accuracy and speed, providing patients and healthcare providers with an effective tool for monitoring the progress of ulcer healing.

Focusing on cloud and edge computing for fall detection, Mrozek et al. (2020) discussed a scalable telemedicine system for remote monitoring of the elderly. It achieves efficient monitoring and alert mechanisms through mobile IoT devices and shows how edge-based processing can reduce data transmission needs and storage consumption, offering practical solutions for monitoring indoor and outdoor activities of the elderly. This ensures timely assistance in the case of falls, a major health risk for the elderly, contributing to improved care and self-sufficiency.

For gait analysis, Martini et al. (2022) introduced a telemedicine app based on 3D HPE using the NVIDIA Jetson Xavier. It employs DL for markerless motion capture, balancing accuracy, portability, and privacy. The app is also suitable for monitoring the elderly, providing real-time and high-precision capabilities. It showcases the potential for remote gait analysis, meeting the demand for scalable, efficient telemedicine solutions.

Han and Lv (2022) introduced a method named SR-Telemedicine to enhance video quality for telemedicine. It focuses on upgrading very low-resolution video chunks to high resolutions such as 720p or 1080p via a DL-based super-resolution model. The method aims to provide high-quality videos for physicians without requiring extensive network bandwidth. The approach leverages a scalable neural network model and a double deep Q-Network (DDQN) algorithm to dynamically adjust the video resolution and model scale based on network and computational capabilities, significantly improving the user's quality of experience by 17-79% over baseline methods.

TytoCare[53] is a versatile handheld examination device capable of capturing a variety of medical data, including heart and lung sounds, throat images, skin conditions, ear canal images, and body temperature. The device utilises advanced computer vision and DL technology to analyse collected data, ensuring diagnoses are comparable in accuracy to in-person consultations. Seamless integration with smartphones and tablets facilitates real-time communication between patients and healthcare providers, enabling immediate medical consultations and diagnoses. The platform supports wireless updates, allowing continuous improvements to its diagnostic algorithms and functionalities to meet evolving healthcare needs.

Proximie[54] is a pioneering platform in the field of remote surgical collaboration and education, designed to transcend geographical boundaries, enabling surgeons to virtually participate in operating rooms across distances and offer their expertise. The platform employs augmented reality (AR), computer vision, and DL algorithms to enhance the visualisation of surgical procedures, facilitating precise guidance, mentoring, and decision making. Proximie's capability to record and analyse surgeries creates a rich database for educational purposes and continuous learning.

[53] TytoCare. https://www.tytocare.com.

[54] Proximie. https://www.proximie.com.

Furthermore, Philips Lumify[55] offers a remote ultrasound solution that, beyond offering portable handheld ultrasound acquisition, optimises video management, transmission, and communication to deliver superior and efficient remote diagnostics and collaboration. Its intuitive, user-friendly interface simplifies the diagnostic process, enabling a broader range of healthcare professionals to utilise advanced ultrasound technology.

## 7 Future directions

Notwithstanding many impressive recent advances surveyed above, there is much room for further development of Edge DL in computer vision and medical applications. Complementing our discussion of the state of the art in this nascent field, we finally discuss potential future directions in improving critical aspects of Edge DL, including privacy and security, energy efficiency and performance optimisation, adaptive edge computing architectures, multimodal edge computing fusion techniques, efficient edge inference in resource-constrained environments, explainable AI in edge models, quantum-inspired edge computing, and human-centric edge applications.

### 7.1 Privacy and security

Edge computing introduces both opportunities and challenges in ensuring data privacy and security. One major area of focus is defending against adversarial attacks that exploit small perturbations in input data to mislead models, leading to dangerous misdiagnoses in medical contexts. To address this challenge, researchers are developing new adversarial training techniques that enhance model robustness by incorporating adversarial samples during training (Isakov et al. 2019). Additionally, there is ongoing exploration into novel regularisation techniques and network architectures to increase model resilience against input disturbances (Jha et al. 2021; Vairo et al. 2023).

Another research direction involves utilising privacy-preserving technologies, such as homomorphic encryption and secure multiparty computation (SMPC), to protect training data and model weights from leaking sensitive information (Giannopoulos and Mouris 2018), which is particularly crucial in healthcare where patient data confidentiality is paramount. Ensuring the ethical handling of patient data is critical, as breaches can lead to severe consequences for patient trust and safety. Moreover, federated learning (Zhang et al. 2021b) enables model training across multiple edge devices without centralising sensitive data, ensuring that patient information remains local. This approach minimises privacy risks while enabling collaborative learning between institutions. Lastly, advances in trusted execution environments (TEE) (Muñoz et al. 2023) at the hardware level offer promising security features by isolating sensitive computations, further reducing vulnerabilities on edge devices.

[55] Philips Lumify. https://www.philips.com.au/healthcare/sites/lumify-handheld-ultrasound/products/reacts-tele-ultrasound.

### 7.2 Energy efficiency and performance optimisation

Energy efficiency and performance optimisation in edge computing require innovations at both hardware and software levels. In medical applications, such optimisations are critical not only for continuous patient monitoring and portable medical devices but also for enabling accurate real-time diagnostics in resource-constrained environments. For instance, devices such as wearable electrocardiogram monitors or portable ultrasound scanners rely on energy-efficient edge processors to provide immediate diagnostic insights without relying on cloud infrastructure. On the hardware front, research is focused on developing processors and storage solutions specifically designed for edge computing, such as low-power microcontrollers, ASICs, and dedicated neural network accelerators (Li and Liewig 2020; Reuther et al. 2021).

Software-level optimisations encompass more efficient data compression algorithms and lightweight neural network architectures, designed to minimise computational and storage requirements while sustaining high performance. Additionally, energy-aware scheduling algorithms, which dynamically optimise task allocation and execution based on the device's energy state and computational demands, are a current research hotspot (Chang et al. 2019). Techniques such as adaptive voltage scaling and power gating are also being integrated at the software level to further reduce energy usage during idle periods. These techniques are particularly useful in scenarios like wearable health monitoring devices, where the system spends a significant portion of time in low-power states but must remain ready for immediate activation in response to critical events, such as detecting arrhythmias (Sarma and Biswas 2020; Yousri et al. 2023).

### 7.3 Adaptive edge computing architectures

The key to adaptive edge computing architectures lies in using advanced algorithms to automatically adjust networks to the ever-changing environment and application requirements. NAS techniques show great potential in this area, automatically optimising the network design to achieve the best performance within given resource constraints (Wistuba et al. 2019). In particular, hardware-aware NAS (Benmeziane et al. 2021) has emerged as an essential tool, considering device-specific metrics like latency, memory usage, and power consumption during the search process. For instance, ProxylessNAS (Cai et al. 2018) is a prominent example of hardware-aware NAS applied to edge computing. It introduces a latency constraint directly into the search process, optimising the architecture for specific hardware, such as mobile devices. By conducting the search directly on target hardware instead of relying on proxy tasks, ProxylessNAS can efficiently generate models with a good balance of latency and accuracy.

In addition, dynamic solutions such as Slimmable Neural Networks (Yu et al. 2019) allow the width of the network to be dynamically adjusted during inference to suit the available computational resources. This method allows a single model to run at different scales, from full network capacity to a more lightweight version, depending on real-time resource availability (Han et al. 2021).

Another practical approach is the use of resource-aware dynamic pruning strategies (Liu and Deng 2019), where unimportant channels are pruned during inference based on input data. This enables the model to adaptively reduce its computational complexity when pro-

cessing less demanding tasks, optimising both power consumption and performance for edge devices. These specific methods highlight the growing integration of adaptive mechanisms within edge computing architectures, offering tailored, resource-efficient solutions for various tasks and conditions, ensuring flexibility and responsiveness across different device and workload scenarios.

### 7.4 Multimodal edge computing fusion techniques

Multimodal edge computing aims to process and analyse heterogeneous data from various sensors and sources. In healthcare, this involves integrating data from modalities such as medical imaging, electronic health records, and wearable sensors to provide comprehensive patient assessments (Calisto 2017). A key challenge lies in developing effective multimodal fusion algorithms that can integrate data from diverse modalities like vision, audio, and text, extracting useful information to support complex decision-making processes. For example, DL and data fusion techniques are being employed to integrate data from multiple devices for more accurate health monitoring and event response. In this domain, researchers are also exploring how to maintain data processing efficiency while ensuring the generalisability and accuracy of algorithms (Acosta et al. 2022), which is vital for reliable and timely medical diagnoses.

Recent advances in transformer models, particularly the Vision-Language Transformer (ViLT) (Fields and Kennington 2023), have been instrumental in fusing visual and textual data with minimal preprocessing. ViLT uses a joint embedding space to represent both image and text modalities, allowing it to perform multimodal tasks efficiently, such as medical report generation or cross-modal retrieval (Delbrouck et al. 2022). The simplification of cross-modal interactions in ViLT demonstrates how transformer-based architectures can potentially be adapted to handle multimodal tasks with reduced computational overhead, which is a key consideration in edge environments. Another emerging approach is the use of contrastive learning for multimodal fusion, where models are trained to maximise agreement between different modalities while maintaining modality-specific representations. For instance, the CLIP model (Radford et al. 2021) has been adapted for medical tasks to align radiology images with text-based reports, making it a powerful tool for image retrieval and diagnostic support (Zhao et al. 2023).

### 7.5 Efficient edge inference in resource-constrained environments

In resource-constrained environments, efficient edge inference relies not only on model optimisation but also involves intelligent energy management and task scheduling strategies. These strategies maximise energy efficiency by optimising the allocation of computational tasks across different types of processing units, such as CPUs, GPUs, and FPGAs, as well as facilitating collaborative work between edge and cloud computing resources (Krzywda et al. 2018). Additionally, efficient processing of real-time data streams is a key component of achieving efficient edge inference. These strategies and optimisations can significantly enhance the performance and energy efficiency of edge computing devices in resource-limited settings (Zhou et al. 2019; Mao et al. 2017; Yu et al. 2017).

One prominent example of optimising edge inference speed is SparseNN (Zhu et al. 2018), which focuses on sparsity-aware execution. SparseNN leverages the inherent spar-

sity in neural networks by skipping computations involving zero-value activations and weights, significantly reducing the overall computational load. Similarly, ShiftAddNet (You et al. 2020a) replaces computationally expensive multiplication operations in convolutional layers with lightweight shift and add operations. ShiftAddNet maintains competitive performance while reducing latency and resource usage, demonstrating a practical balance between accuracy and efficiency in edge computing environments.

### 7.6 Explainable AI in edge models

Explainable AI (XAI) has emerged as a key aspect of deploying AI systems. The demand for transparency, trust, and compliance with regulatory requirements has driven the development of explainability, which is crucial for sensitive applications in healthcare and autonomous driving. Current integrated learning approaches, such as LIME (Ribeiro et al. 2016) and SHAP (Lundberg and Lee 2017), provide visual explanations that elucidate how various features influence model decisions. Recent research has focused on developing lightweight and real-time explainability frameworks suitable for edge devices (Gilpin et al. 2018; Li et al. 2023; Huang and Gao 2022; Saini et al. 2023). While progress has been made in integrating XAI into edge models, challenges remain, such as the trade-off between model performance and explainability, and the need for standardised frameworks for model explainability.

In the context of healthcare, explainability becomes even more critical. Medical professionals require clear, understandable insights into how AI models arrive at specific diagnoses or treatment recommendations. This transparency not only aids in clinical decision-making but also fosters trust among patients and healthcare providers. In addition, edge-based diagnostic tools that use XAI can help doctors understand the reasoning behind AI-driven imaging analyses, making it easier to justify and trust AI recommendations in a clinical setting. Therefore, enhancing the explainability of edge models in medical applications is not just a technical challenge but a necessity for broader acceptance and effective use in healthcare environments.

### 7.7 Quantum-inspired edge computing

Quantum-inspired edge computing leverages the principles of quantum computing to enhance traditional computational methods for edge applications. In medical diagnostics, quantum-inspired techniques can potentially revolutionise data processing capabilities and optimise resource allocation. Researchers are exploring the use of quantum ML to enhance efficiency (Biamonte et al. 2017). Quantum teleportation methods may enhance secure communications in edge computing (Pirandola et al. 2015; Liu 2020) and quantum optimisation methods such as quantum annealing may help solve complex optimisation problems and improve performance (Das and Chakrabarti 2008). The integration of quantum-inspired technologies in edge computing is still in its early stages, but it offers broad prospects for overcoming the limitations of classical computing in edge environments. Future research may focus on creating more powerful quantum algorithms, developing quantum-resistant security protocols for edge networks, and exploring the potential of hybrid classical and quantum systems to improve diagnostic accuracy and speed in medical applications.

### 7.8 Human-centric edge applications

Human-centric edge applications emphasise technology design and services centred around human needs and experiences. In health monitoring, human-centric edge applications can utilise portable edge devices for instantaneous analysis and detection of health data to provide customised services, personalised feedback, and health advice. This enables users to better understand their health condition, potentially identifying health issues early and promoting proactive health management (García et al. 2017; Abrantes et al. 2023). In smart homes, human-centric edge applications leverage various edge devices for user interaction and environmental monitoring, automating home appliances and systems for optimal comfort and energy efficiency (Patchava et al. 2015; Schneider and Banerjee 2018). An example is the use of thermal imaging to detect human body temperature and adjust room temperature accordingly (Vázquez and Kastner 2012). While human-centric edge applications offer many benefits, they also present challenges related to privacy, data security, and computational limitations. Protecting user data while providing personalised services requires robust security protocols and careful handling of sensitive information, which are key topics of future research (Calisto et al. 2022; Lakshminarayanan et al. 2023).

### 7.9 Integration of advanced learning-based methods

Recent years have also witnessed significant advancements in learning-based diagnostic methods for medical images, including the use of transformers (You et al. 2022), contrastive learning (You et al. 2022, 2023a, c), few-shot learning (You et al. 2022, 2024), transfer learning (You et al. 2021, 2022), domain adaptation (You et al. 2020b) and generative models (You et al. 2018, 2023b). Transformers, with their self-attention mechanisms, have shown a remarkable ability to capture intricate patterns and long-range dependencies in medical images, significantly improving diagnostic accuracy. Contrastive learning leverages unlabeled data to learn robust feature representations, reducing the reliance on large labeled datasets and enhancing model generalisation. Few-shot learning aims to train models with a very small amount of labeled data, which is particularly valuable in medical diagnostics where labeled data is often scarce. Transfer learning allows for fine-tuning or distillation of models pre-trained on large datasets, enabling them to effectively handle specific medical image diagnostic tasks. Domain adaptation techniques address the challenge of varying imaging conditions and equipment by aligning feature distributions between different domains, thereby improving model generalisation. Integrating these advanced learning-based technologies with edge computing can enhance the accuracy and efficiency of medical diagnostics, contributing to more reliable analysis and support in clinical settings.

## 8 Conclusion

The integration of edge computing and deep learning (Edge DL) offers unprecedented opportunities for real-time processing and interpretation of data close to the source in resource-constrained settings. Focusing on applications in computer vision and medical diagnostics, we have explored the fundamental concepts and technical merits of Edge DL, underlining its critical role in the evolution of modern computational paradigms. Specifi-

cally, we have delved into model compression and the design of lightweight models suitable for operation on edge devices. Reducing latency, conserving bandwidth, and bolstering data privacy, Edge DL holds the promise of revolutionising various industries, from autonomous vehicles to smart healthcare systems. In surveying the current landscape and potential future developments of Edge DL, we have highlighted its burgeoning possibilities and challenges. No doubt, continuing advances in hardware and software for Edge DL will have an increasingly transformative impact in a growing range of applications. It is our hope that Edge DL will contribute to improve our daily lives and bring us closer to the objective of universal healthcare accessibility.

**Acknowledgements** The authors would like to express their gratitude to Matthew Sheedy, George Bou-Rizk, Tim Kannegieter, and Geoff Sizer from Genesys Electronics Design, Sydney, Australia, for engaging and insightful discussions that have been invaluable in refining our understanding and conceptual framing of the discussed themes.

**Author contributions** Yiwen Xu and Dr. Tariq M. Khan both contributed to drafting the initial manuscript. Dr. Tariq M. Khan also designed the overall framework of the paper. A/Prof. Yang Song critically reviewed of the manuscript, ensuring the scientific rigor and clarity of the content. Prof. Erik Meijering extensively revised the manuscript for intellectual content, refining the arguments and improving the overall narrative flow. Prof. Erik Meijering also supervised the entire project, ensuring the integrity and accuracy of the work, and provided final approval of the version to be published.

**Funding** This work was supported by the ARC Research Hub for Connected Sensors for Health (Grant number IH210100040). Partial financial support was received from Genesys.

**Data availability** No datasets were generated or analysed during the current study.

## Declarations

**Competing interests** The authors have no relevant financial or non-financial interests to disclose.



## References

Abbas N, Zhang Y, Taherkordi A, Skeie T (2017) Mobile edge computing: a survey. IEEE Internet Things J 5(1):450–465

Abdellatif AA, Samara L, Mohamed A, Erbad A, Chiasserini CF, Guizani M, O'Connor MD, Laughton J (2021) MEdge-Chain: leveraging edge computing and blockchain for efficient medical data exchange. IEEE Internet Things J 8(21):15762–15775

Abrantes J, Silva MJ, Meneses J, Oliveira C, Calisto FM, Filice R (2023) External validation of a deep learning model for breast density classification. ESR-European Society of Radiology, Vienna

Abreha HG, Hayajneh M, Serhani MA (2022) Federated learning in edge computing: a systematic survey. Sensors 22(2):450
Abubeker K, Baskar S (2023) B2-net: an artificial intelligence powered machine learning framework for the classification of pneumonia in chest x-ray images. Machine Learning: Science and Technology 4(1):015036
Achakir F, Mohtaram N, Escartin A (2023) An automated AI-based solution for out-of-stock detection in retail environments. In: International conference on electrical. computer, communications and mechatronics engineering (ICECCME), pp 1–6
Acosta JN, Falcone GJ, Rajpurkar P, Topol EJ (2022) Multimodal biomedical AI. Nat Med 28(9):1773–1784
Albattah W, Masood M, Javed A, Nawaz M, Albahli S (2023) Custom CornerNet: a drone-based improved deep learning technique for large-scale multiclass pest localization and classification. Complex Intell Syst 9(2):1299–1316
Albuquerque CK, Polimante S, Torre-Neto A, Prati RC (2020) Water spray detection for smart irrigation systems with mask R-CNN and UAV footage. In: IEEE International workshop on metrology for agriculture and forestry (MetroAgriFor), pp 236–240
Alexey G, Klyachin V, Eldar K, Driaba A (2021) Autonomous mobile robot with AI based on Jetson Nano. In: Future technologies conference (FTC), pp 190–204
Ali B, Gregory MA, Li S (2021) Multi-access edge computing architecture, data security and privacy: a review. IEEE Access 9:18706–18721
Alwakeel AM (2021) An overview of fog computing and edge computing security and privacy issues. Sensors 21(24):8226
Ampatzidis Y, Partel V (2019) UAV-based high throughput phenotyping in citrus utilizing multispectral imaging and artificial intelligence. Remote Sensing 11(4):410
Angus A, Duan Z, Zussman G, Kostić Z (2022) Real-time video anonymization in smart city intersections. In: IEEE international conference on mobile ad hoc and smart systems (MASS), pp. 514–522
Arshad B, Barthelemy J, Pilton E, Perez P (2020) Where is my deer? Wildlife tracking and counting via edge computing and deep learning. In: IEEE SENSORS, pp 1–4
ASGE Technology Committee ASGE, Wang A, Banerjee S, Barth BA, Bhat YM, Chauhan S, Gottlieb KT, Konda V, Maple JT, Murad F, Pfau PR, Pleskow DK, Siddiqui UD, Tokar JL, Rodriguez SA (2013) Wireless capsule endoscopy. Gastrointest Endosc 78(6):805–815
Auguste L, Palsana D (2015) Mobile Whole Slide Imaging (mWSI): a low resource acquisition and transport technique for microscopic pathological specimens. BMJ Innov 1(3):137–143
Avvenuti M, Bongiovanni M, Ciampi L, Falchi F, Gennaro C, Messina N (2022) A spatio-temporal attentive network for video-based crowd counting. In: IEEE symposium on computers and communications (ISCC), pp 1–6
Azimi I, Takalo-Mattila J, Anzanpour A, Rahmani AM, Soininen J-P, Liljeberg P (2018) Empowering healthcare iot systems with hierarchical edge-based deep learning. In: Proceedings of the 2018 IEEE/ACM international conference on connected health: applications, systems and engineering technologies, pp 63–68
Babar M, Khan MS, Ali F, Imran M, Shoaib M (2021) Cloudlet computing: recent advances, taxonomy, and challenges. IEEE Access 9:29609–29622
Banbury C, Zhou C, Fedorov I, Navarro RM, Thakker U, Gope D, Reddi VJ, Mattina M, Whatmough PN (2021) MicroNets: neural network architectures for deploying TinyML applications on commodity microcontrollers. In: Machine learning and systems (MLSys), pp 1–16
Barisoni L, Lafata KJ, Hewitt SM, Madabhushi A, Balis UG (2020) Digital pathology and computational image analysis in nephropathology. Nat Rev Nephrol 16(11):669–685
Baumgartner T, Klatt S (2023) Monocular 3D human pose estimation for sports broadcasts using partial sports field registration. In: IEEE/CVF conference on computer vision and pattern recognition (CVPR), pp 5108–5117
Benmeziane H, Maghraoui KE, Ouarnoughi H, Niar S, Wistuba M, Wang N (2021) A comprehensive survey on hardware-aware neural architecture search. arXiv:2101.09336
Bhardwaj K, Diffenderfer J, Kailkhura B, Gokhale M (2022) Unsupervised test-time adaptation of deep neural networks at the edge: a case study. In: Design, automation & test in Europe conference & exhibition (DATE), pp 412–417
Biamonte J, Wittek P, Pancotti N, Rebentrost P, Wiebe N, Lloyd S (2017) Quantum machine learning. Nature 549(7671):195–202
Bochkovskiy A, Wang C-Y, Liao H-YM (2020) YOLOv4: Optimal speed and accuracy of object detection. arXiv:2004.10934
Bonam J, Kondapalli SS, Prasad L, Marlapalli K et al (2023) Lightweight cnn models for product defect detection with edge computing in manufacturing industries. J Sci Ind Res 82(04):418–425

Bonomi F, Milito R, Zhu J, Addepalli S (2012) Fog computing and its role in the Internet of Things. In: Workshop on mobile cloud computing (MCC), pp 13–16
Bottou L (2010) Large-scale machine learning with stochastic gradient descent. In: International conference on computational statistics (COMPSTAT), pp 177–186
Brecko A, Kajati E, Koziorek J, Zolotova I (2022) Federated learning for edge computing: a survey. Appl Sci 12(18):9124
Cai H, Zhu L, Han S (2018) Proxylessnas: Direct neural architecture search on target task and hardware. arXiv:1812.00332
Calisto FM (2017) Medical imaging multimodality breast cancer diagnosis user interface. Master's thesis. Instituto Superior Técnico, Pais 1
Calisto FM, Nunes N, Nascimento JC (2022) Modeling adoption of intelligent agents in medical imaging. Int J Hum Comput Stud 168:102922
Cao K, Liu Y, Meng G, Sun Q (2020) An overview on edge computing research. IEEE Access 8:85714–85728
Cao L, Song P, Wang Y, Yang Y, Peng B (2023) An improved lightweight real-time detection algorithm based on the edge computing platform for UAV images. Electronics 12(10):2274
Cao W, Shen W, Zhang Z, Qin J (2023) Privacy-preserving healthcare monitoring for IoT devices under edge computing. Comput Secur 134:103464
Cass S (2020) Nvidia makes it easy to embed AI: the Jetson Nano packs a lot of machine-learning power into DIY projects. IEEE Spectr 57(7):14–16
Cazzolato MT, Ramos JS, Rodrigues LS, Scabora LC, Chino DY, Jorge AE, Azevedo-Marques PM, Traina C Jr, Traina AJ (2021) The UTrack framework for segmenting and measuring dermatological ulcers through telemedicine. Comput Biol Med 134:104489
Chang H-Y, Narayanan P, Lewis SC, Farinha NC, Hosokawa K, Mackin C, Tsai H, Ambrogio S, Chen A, Burr GW (2019) AI hardware acceleration with analog memory: microarccroarchitectures for low energy at high speed. IBM J Res Dev 63(6):8–1814
Chang R, Jie W, Thakur N, Zhao Z, Pahwa RS, Yang X (2024) A unified and adaptive continual learning method for feature segmentation of buried packages in 3d XRM images. In: 2024 IEEE 74th electronic components and technology conference (ECTC), pp 1872–1879. IEEE
Chavan S, Ford J, Yu X, Saniie J (2021) Plant species image recognition using artificial intelligence on Jetson Nano computational platform. In: IEEE international conference on electro information technology (EIT), pp 350–354
Chen J, Ran X (2019) Deep learning with edge computing: a review. Proc IEEE 107(8):1655–1674
Chen M, Zhang X (2023) Structured pruning for deep neural networks. J Artif Intell Res 68:77–102
Chen X-L, Zhao H-M, Li P-X, Yin Z-Y (2006) Remote sensing image-based analysis of the relationship between urban heat island and land use/cover changes. Remote Sens Environ 104(2):133–146
Chen Y, Zhao Q, Hu X, Hu B (2019) Multi-resolution parallel magnetic resonance image reconstruction in mobile computing-based IoT. IEEE Access 7:15623–15633
Chen X, He Y, Li Y, Shi J (2021) Leveraging pruning and quantization for efficient neural network inference. IEEE Trans Neural Netw Learn Syst 32(8):3373–3384
Chen J, Li X, Zhang Y (2021) Mobile transformer for face recognition on low-power edge devices. IEEE Trans Mob Comput 20(5):1058–1070
Chen B, Bakhshi A, Batista G, Ng B, Chin T-J (2022a) Update compression for deep neural networks on the edge. In: IEEE/CVF conference on computer vision and pattern recognition (CVPR), pp 3076–3086
Chen L, Zhou Y, Zhou H, Zu J (2022b) Detection of polarizer surface defects based on an improved lightweight YOLOv3 model. In: International conference on intelligent control, measurement and signal processing (ICMSP), pp 138–142
Chino DY, Scabora LC, Cazzolato MT, Jorge AE, Traina-Jr C, Traina AJ (2020) Segmenting skin ulcers and measuring the wound area using deep convolutional networks. Comput Methods Programs Biomed 191:105376
Chung C-C, Chen W-T, Chang Y-C (2020) Using quantization-aware training technique with post-training fine-tuning quantization to implement a MOBILENET hardware accelerator. In: 2020 Indo—Taiwan 2nd international conference on computing, analytics and networks (Indo-Taiwan ICAN), pp. 28–32
Cioppa A, Deliege A, Giancola S, Ghanem B, Droogenbroeck MV, Gade R, Moeslund TB (2020) A context-aware loss function for action spotting in soccer videos. In: IEEE/CVF conference on computer vision and pattern recognition (CVPR), pp 13126–13136
Corley DA, Jensen CD, Marks AR, Zhao WK, Lee JK, Doubeni CA, Zauber AG, Boer J, Fireman BH, Schottinger JE, Quinn VP, Ghai NR, Levin TR, Quesenberry CP (2014) Adenoma detection rate and risk of colorectal cancer and death. N Engl J Med 370(14):1298–1306
Cui X, Hu R (2022) Application of intelligent edge computing technology for video surveillance in human movement recognition and Taekwondo training. Alex Eng J 61(4):2899–2908

Cum F (2022) A neural network application for impedance-based plant monitoring: from a development framework towards edge computing. PhD Thesis, Politecnico di Torino
D'Agostino D, Morganti L, Corni E, Cesini D, Merelli I (2019) Combining edge and cloud computing for low-power, cost-effective metagenomics analysis. Futur Gener Comput Syst 90:79–85
Dai W, Mujeeb A, Erdt M, Sourin A (2020) Soldering defect detection in automatic optical inspection. Adv Eng Inform 43:101004
Dai X, Spasić I, Meyer B, Chapman S, Andres F (2019) Machine learning on mobile: an on-device inference app for skin cancer detection. In: International conference on fog and mobile edge computing (FMEC), pp 301–305
Dang LM, Hassan SI, Suhyeon I, Sangaiah A, Mehmood I, Rho S, Seo S, Moon H (2020) UAV based wilt detection system via convolutional neural networks. Sustain Comput Inf Syst 28:100250
Das A, Chakrabarti BK (2008) Colloquium: quantum annealing and analog quantum computation. Rev Mod Phys 80(3):1061
Datta Gupta K, Sharma DK, Ahmed S, Gupta H, Gupta D, Hsu C-H (2023) A novel lightweight deep learning-based histopathological image classification model for IoMT. Neural Process Lett 55(1):205–228
Dave R, Seliya N, Siddiqui N (2021) The benefits of edge computing in healthcare, smart cities, and IoT. arXiv:2112.01250
De Simone G, Foggia P, Saggese A, Vento M (2023) Autonomous mobile robot for automatic out of stock detection in a supermarket. In: Proceedings of the IEEE/CVF international conference on computer vision, pp 1829–1838
Dekhovich A, Bessa MA (2024) Continual learning for surface defect segmentation by subnetworkcreation and selection. J Intell Manuf. https://doi.org/10.1007/s10845-024-02393-4
Delbrouck J-b, Saab K, Varma M, Eyuboglu S, Chambon P, Dunnmon J, Zambrano J, Chaudhari A, Langlotz C (2022) Vilmedic: a framework for research at the intersection of vision and language in medical ai. In: Proceedings of the 60th annual meeting of the association for computational linguistics: system demonstrations, pp 23–34
Dell: Edge Intelligence Trends in the Retail Industry. https://infohub.delltechnologies.com/zh-cn/p/edge-intelligence-trends-in-the-retail-industry/. Accessed 26 July 2024
Dell: Precision 3660 Tower Workstation. https://www.dell.com/en-au/shop/dell-desktop-computers/precision-3660-tower-workstation/spd/precision-3660-workstation (Accessed 28 January 2024)
Deng Z-Y, Chiang H-H, Kang L-W, Li H-C (2023) A lightweight deep learning model for real-time face recognition. IET Image Proc 17(13):3869–3883
Di Benedetto M, Carrara F, Ciampi L, Falchi F, Gennaro C, Amato G (2022) An embedded toolset for human activity monitoring in critical environments. Expert Syst Appl 199:117125
Dong P, Ning Z, Obaidat MS, Jiang X, Guo Y, Hu X, Hu B, Sadoun B (2020) Edge computing based healthcare systems: enabling decentralized health monitoring in internet of medical things. IEEE Network 34(5):254–261
Dong S, Wang P, Abbas K (2021) A survey on deep learning and its applications. Comput Sci Rev 40:100379
Dong C, Li TZ, Xu K, Wang Z, Maldonado F, Sandler K, Landman BA, Huo Y (2023) Characterizing browser-based medical imaging AI with serverless edge computing: towards addressing clinical data security constraints. SPIE Med Imaging 12469:10–1117122653626
Dong Z, He Q, Chen F, Jin H, Gu T, Yang Y (2023) EdgeMove: pipelining device-edge model training for mobile intelligence. In: ACM web conference, pp 3142–3153
Dutta L, Bharali S (2021) TinyML meets IoT: a comprehensive survey. Internet Things 16:100461
Elsken T, Metzen JH, Hutter F (2019) Neural architecture search: a survey. J Mach Learn Res 20(55):1–21
Esteva A, Chou K, Yeung S, Naik N, Madani A, Mottaghi A, Liu Y, Topol E, Dean J, Socher R (2021) Deep learning-enabled medical computer vision. NPJ Digit Med 4(1):5
Fan X, Yan Y, Yang P, Han F (2021) CMSS: use low-power IoT cameras to monitor store shelves. In: International conference on big data computing and communications (BigCom), pp 309–315
Farooq H, Zafar Z, Saadat A, Khan TM, Iqbal S, Razzak I (2024) LSSF-NET: lightweight segmentation with self-awareness, spatial attention, and focal modulation. arXiv:2409.01572
Feng H, Mu G, Zhong S, Zhang P, Yuan T (2022) Benchmark analysis of yolo performance on edge intelligence devices. Cryptography 6(2):16
Fernández J, Cañas JM, Fernández V, Paniego S (2021) Robust real-time traffic surveillance with deep learning. Comput Intell Neurosci 2021:4632353
Fields C, Kennington C (2023) Vision language transformers: a survey. arXiv:2307.03254
France KK, Newman ZA (2020) Cluster neural networks for edge intelligence in medical imaging. ResearchGate
García CG, Meana-Llorián D, G-Bustelo BCP, Lovelle JMC, Garcia-Fernandez N (2017) MIDGAR: detection of people through computer vision in the Internet of Things scenarios to improve the security in smart cities, smart towns, and smart homes. Fut Gen Comput Syst 76: 301–313 (2017)

Geiger A, Lenz P, Stiller C, Urtasun R (2013) Vision meets robotics: the KITTI dataset. Int J Robot Res 32(11):1231–1237
Gholami A, Kwon K, Wu B, Tai Z, Yue X, Jin P, Zhao S, Keutzer K (2018) SqueezeNext: hardware-aware neural network design. arXiv:1803.10615
Giannopoulos AG, Mouris DI (2018) Privacy preserving medical data analytics using secure multi party computation: an end-to-end use case. Master Thesis, National and Kapodistrian University of Athens
Gilpin LH, Bau D, Yuan BZ, Bajwa A, Specter M, Kagal L (2018) Explaining explanations: an overview of interpretability of machine learning. In: IEEE international conference on data science and advanced analytics (DSAA), pp 80–89
Girshick R (2015) Fast R-CNN. In: IEEE international conference on computer vision (ICCV), pp 1440–1448
Göçeri E (2020) Impact of deep learning and smartphone technologies in dermatology: automated diagnosis. In: 2020 10th international conference on image processing theory, tools and applications (IPTA), pp 1–6. IEEE
Goceri E (2021a) Automated skin cancer detection: where we are and the way to the future. In: 2021 44th International conference on telecommunications and signal processing (TSP), pp 48–51. IEEE
Goceri E (2021b) Diagnosis of skin diseases in the era of deep learning and mobile technology. Comput Biol Med 134:104458
Goceri E (2024) Polyp segmentation using a hybrid vision transformer and a hybrid loss function. J Imaging Inf Med 37(2):851–863
Gonzalez-Huitron V, León-Borges JA, Rodriguez-Mata A, Amabilis-Sosa LE, Ramírez-Pereda B, Rodriguez H (2021) Disease detection in tomato leaves via CNN with lightweight architectures implemented in Raspberry Pi 4. Comput Electron Agric 181:105951
Gotthard R, Broström M (2023) Edge machine learning for wildlife conservation: a part of the ngulia project. Master Thesis, Linköping University
Goyal M, Reeves ND, Rajbhandari S, Yap MH (2018) Robust methods for real-time diabetic foot ulcer detection and localization on mobile devices. IEEE J Biomed Health Inform 23(4):1730–1741
Greco L, Percannella G, Ritrovato P, Tortorella F, Vento M (2020) Trends in IoT based solutions for health care: moving AI to the edge. Pattern Recogn Lett 135:346–353
Gtifa W, Sakly A (2023) Integrating xilinx fpga and intelligent techniques for improved precision in 3D brain tumor segmentation in medical imaging. J Real-Time Image Proc 20(6):115
Gu L, Mukherjee M, Guo M, Lloret J, Matam R (2022) Low-cost assistive body temperature screening system to combat communicable infectious diseases leveraging edge computing and long-range and low-power wireless networks. IEEE Internet Things J 10(5):4174–4183
Gupta S, Mohan N, Nayak P, Nagaraju KC, Karanam M (2022) Deep vision-based surveillance system to prevent train-elephant collisions. Soft Comput 26:4005–4018
Han H, Lv J (2022) Super-resolution-empowered adaptive medical video streaming in telemedicine systems. Electronics 11(18):2944
Han Y, Huang G, Song S, Yang L, Wang H, Wang Y (2021) Dynamic neural networks: a survey. IEEE Trans Pattern Anal Mach Intell 44(11):7436–7456
Hang J, Sun H, Yu X, Rodríguez-Andina JJ, Yang X (2022) Surface defect detection in sanitary ceramics based on lightweight object detection network. IEEE Open J Ind Electron Soc 3:473–483
Han S, Mao H, Dally WJ (2015) Deep compression: compressing deep neural networks with pruning, trained quantization and huffman coding. arXiv:1510.00149
Han K, Wang Y, Tian Q, Guo J, Xu C, Xu C (2020) GhostNet: more features from cheap operations. In: IEEE/CVF conference on computer vision and pattern recognition (CVPR), pp 1580–1589
He K, Gkioxari G, Dollár P, Girshick R (2017a) Mask R-CNN. In: IEEE international conference on computer vision (ICCV), pp 2961–2969
He Y, Zhang X, Sun J (2017b) Channel pruning for accelerating very deep neural networks. In: IEEE international conference on computer vision (ICCV), pp 1389–1397
He Y, Lin J, Liu Z, Wang H, Li L-J (2018) AMC: AutoML for model compression and acceleration on mobile devices. In: European conference on computer vision (ECCV), pp 784–800
He Y, Zhang X, Sun J (2019) Soft filter pruning for accelerating deep convolutional neural networks. IEEE International Conference on Computer Vision (ICCV), 2234–2243
He Y, Kang G, Dong X, Fu Y, Yang Y (2020) Learning filter pruning criteria for deep convolutional neural networks. IEEE Trans Pattern Anal Mach Intell 42(10):2535–2544
He Q, Wang X, Zhang L (2021) Neural architecture search for wearable healthcare devices: efficient and lightweight models for real-time monitoring. IEEE Trans Mob Comput 20(6):1492–1504
Heo J, Kim G, Park J, Kim Y, Cho S-S, Lee CW, Kang S-J (2020) Lightweight deep neural network-based real-time pose estimation on embedded systems. In: IEEE intelligent vehicles symposium (IV), pp 1066–1071

Ho J, Jain A, Abbeel P (2020) Denoising diffusion probabilistic models. Adv Neural Inf Process Syst (NeurIPS) 33:6840–6851
Hou D, Hou MR, Hou J (2020) On-device subspace learning chest X-ray screening. In: IEEE international conference on consumer electronics (ICCE), pp 1–5
Howard AG, Zhu M, Chen B, Kalenichenko D, Wang W, Weyand T, Andreetto M, Adam H (2017) MobileNets: efficient convolutional neural networks for mobile vision applications. arXiv:1704.04861
Howard A, Sandler M, Chu G, Chen L-C, Chen B, Tan M, Wang W, Zhu Y, Pang R, Vasudevan V, Le QV, Adam H (2019) Searching for MobileNetV3. In: IEEE/CVF International conference on computer vision (ICCV), pp 1314–1324
Hu H, Peng R, Tai Y-W, Tang C-K (2016) Network trimming: a data-driven neuron pruning approach towards efficient deep architectures. arXiv:1607.03250
Hu J, Shen L, Sun G (2018) Squeeze-and-excitation networks. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 7132–7141
Hu H, Wang D, Wu C (2020) Distributed machine learning through heterogeneous edge systems. In: AAAI conference on artificial intelligence, vol 34, pp 7179–7186
Huang K, Gao W (2022) Real-time neural network inference on extremely weak devices: agile offloading with explainable AI. In: Annual international conference on mobile computing and networking (MobiCom), pp 200–213
Huang G, Liu Z, Van Der Maaten L, Weinberger KQ (2017) Densely connected convolutional networks. In: IEEE conference on computer vision and pattern recognition (CVPR), pp 4700–4708
Huang G, Liu S, Maaten L, Weinberger KQ (2018) CondenseNet: an efficient DenseNet using learned group convolutions. In: IEEE conference on computer vision and pattern recognition (CVPR), pp 2752–2761
Huang Q, Han Y, Zhang X, Sheng J, Zhang Y, Xie H (2023a) FFKD-CGhostNet: a novel lightweight network for fault diagnosis in edge computing scenarios. IEEE Trans Instrum Meas 72:3536410
Huang X, Liu Z, Liu S-Y, Cheng K-T (2023b) Efficient quantization-aware training with adaptive coreset selection. In: ICLR 2024 conference
Huang W, Qin H, Liu Y, Liang J, Zhang Y, Li Y, Liu X (2024) OHQ: on-chip hardware-aware quantization. arXiv preprint. arXiv:2309.01945
Iandola FN, Han S, Moskewicz MW, Ashraf K, Dally WJ, Keutzer K (2016) SqueezeNet: AlexNet-level accuracy with 50× fewer parameters and < 0.5MB model size. arXiv:1602.07360
Idlahcen F, Idri A, Goceri E (2024) Exploring data mining and machine learning in gynecologic oncology. Artif Intell Rev 57(2):20
Imran HA, Mujahid U, Wazir S, Latif U, Mehmood K. (2020) Embedded development boards for edge-AI: a comprehensive report. arXiv:2009.00803
Intel: how edge computing is driving advancements in healthcare analytics. https://www.intel.com/content/www/us/en/healthcare-it/edge-analytics.html. Accessed 25 July 2024
Iqbal S, Khan TM, Naqvi SS, Naveed A, Usman M, Khan HA, Razzak I (2023) LDMRes-Net: a lightweight neural network for efficient medical image segmentation on iot and edge devices. IEEE J Biomed Health Inf 28(7):3860–3871
Isakov M, Gadepally V, Gettings KM, Kinsy MA (2019) Survey of attacks and defenses on edge-deployed neural networks. In: IEEE high performance extreme computing conference (HPEC). IEEE, pp 1–8
Javed S, Khan T.M, Qayyum A, Sowmya A, Razzak I (2024) Advancing medical image segmentation with Mini-Net: a lightweight solution tailored for efficient segmentation of medical images. arXiv:2405.17520
Jebadurai J, Jebadurai IJ, Paulraj GJL, Joseph BRC (2021) Green IoT-low cost device for the detection of deep vein thrombosis using edge computing. J Green Eng 11:1266–1276
Jha S, Jalaian B, Roy A, Verma G (2021) Trinity: trust resilience and interpretability of machine learning models. In: Game theory and machine learning for cyber security. IEEE, pp 317–333
Jiang Z, Chen T, Li M (2018) Efficient deep learning inference on edge devices. In: International conference on systems and machine learning (SysML), pp 1–3
Jiang Y, Zhang L, Wu Z (2020) Efficient neural architecture search for autonomous drone navigation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops, pp. 158–165. IEEE
Jiang X, Hu Z, Wang S, Zhang Y (2023) Deep learning for medical image-based cancer diagnosis. Cancers 15(14):3608
Jocher G (2020) YOLOv5 by Ultralytics. https://github.com/ultralytics/yolov5
Jolles JW (2021) Broad-scale applications of the Raspberry Pi: a review and guide for biologists. Methods Ecol Evol 12(9):1562–1579
Kang D, Kang D, Kang J, Yoo S, Ha S (2018) Joint optimization of speed, accuracy, and energy for embedded image recognition systems. In: Design, automation & test in Europe conference & exhibition (DATE), pp 715–720

Kanjula KR, Reddy VV, Jnanesh KP, Abraham JS, Tanuja K (2022) People counting system for retail analytics using edge AI. arXiv:2205.13020
Kara OC, Xue J, Venkatayogi N, Mohanraj TG, Hirata Y, Ikoma N, Atashzar SF, Alambeigi F (2023) A smart handheld edge device for on-site diagnosis and classification of texture and stiffness of excised colorectal cancer polyps. arXiv:2309.09642
Karaman O, Alhudhaif A, Polat K (2021) Development of smart camera systems based on artificial intelligence network for social distance detection to fight against COVID-19. Appl Soft Comput 110:107610
Kauczor H-U, Bonomo L, Gaga M, Nackaerts K, Peled N, Prokop M, Remy-Jardin M, Von Stackelberg O, Sculier J-P (2015) European Society of Radiology (ESR), European Respiratory Society (ERS): ESR/ERS white paper on lung cancer screening. Eur Radiol 25:2519–2531
Kaymak C, Aysegul U (2018) Implementation of object detection and recognition algorithms on a robotic arm platform using Raspberry Pi. In: International Conference on artificial intelligence and data processing (IDAP), pp 1–8
Kellermayr-Scheucher M, Hörandner L, Brandtner P (2022) Digitalization at the point-of-sale in grocery retail-state of the art of smart shelf technology and application scenarios. Procedia Comput Sci 196:77–84
Khan TM, Arsalan M, Robles-Kelly A, Meijering E (2022a) MKIS-Net: a light-weight multi-kernel network for medical image segmentation. In: International conference on digital image computing: techniques and applications (DICTA), pp 1–8. https://doi.org/10.1109/DICTA56598.2022.10034573
Khan TM, Naqvi SS, Robles-Kelly A, Meijering E (2022b) Neural network compression by joint sparsity promotion and redundancy reduction. In: International conference on neural information processing. Springer, Cham, pp 612–623
Khan TM, Robles-Kelly A, Naqvi, SS (2022c) T-net: A resource-constrained tiny convolutional neural network for medical image segmentation. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, pp. 644–653
Khan TM, Iqbal S, Naqvi SS, Razzak I, Meijering E (2024a) LMBF-Net: a lightweight multipath bidirectional focal attention network for multifeatures segmentation. arXiv:2407.02871
Khan TM, Naqvi SS, Meijering E (2024b) ESDMR-Net: a lightweight network with expand-squeeze and dual multiscale residual connections for medical image segmentation. Eng Appl Artif Intell 133:107995
Kim J, Lee M, Cho Y (2022) Efficient object tracking for augmented reality with mobile transformers. In: Proceedings of the 2022 IEEE conference on virtual reality and 3D user interfaces. IEEE, pp 652–661
Kirillov A, Mintun E, Ravi N, Mao H, Rolland C, Gustafson L, Xiao T, Whitehead S, Berg AC, Lo W-Y, Dollár P, Girshick R (2023) Segment anything. arXiv:2304.02643
Koonce B, Koonce B (2021) MobileNetV3. Convolutional neural networks with swift fortensorflow: image recognition and dataset categorization, pp 125–144
Kortli Y, Gabsi S, Voon LFLY, Jridi M, Merzougui M, Atri M (2022) Deep embedded hybrid CNN-LSTM network for lane detection on NVIDIA Jetson Xavier NX. Knowl Based Syst 240:107941
Krizhevsky A, Sutskever I, Hinton GE (2012) ImageNet classification with deep convolutional neural networks. In: Advances in neural information processing systems (NeurIPS), pp 1–9
Krupnik O, Shafer E, Jurgenson T, Tamar A (2023) Fine-tuning generative models as an inference method for robotic tasks. In: Conference on robot learning (CoRL), pp 866–886
Krzywda J, Ali-Eldin A, Carlson TE, Östberg P-O, Elmroth E (2018) Power-performance tradeoffs in data center servers: DVFS, CPU pinning, horizontal, and vertical scaling. Futur Gener Comput Syst 81:114–128
Kumar S, Thosani N, Ladabaum U, Friedland S, Chen AM, Kochar R, Banerjee S (2017) Adenoma miss rates associated with a 3-minute versus 6-minute colonoscopy withdrawal time: a prospective, randomized trial. Gastrointest Endosc 85(6):1273–1280
Kumar A, Sharma A, Bharti V, Singh AK, Singh SK, Saxena S (2021) MobiHisNet: a lightweight CNN in mobile edge computing for histopathological image classification. IEEE Internet Things J 8(24):17778–17789
Kyrkou C (2020) YOLOpeds: efficient real-time single-shot pedestrian detection for smart camera applications. IET Comput Vis 14(7):417–425
Lachhab W (2023) Deep learning for efficient retail shelf stock monitoring and analysis. Master Thesis, Aalto University, pp 1–48
Lakshminarayanan V, Ravikumar A, Sriraman H, Alla S, Chattu VK (2023) Health care equity through intelligent edge computing and augmented reality/virtual reality: a systematic review. J Multidisc Healthc 16:2839–2859
Law H, Deng J (2018) CornerNet: detecting objects as paired keypoints. In: European conference on computer vision (ECCV), pp 734–750

Le MQ, Nguyen QT, Dao VH, Tran T-H (2022) CNN quantization for anatomical landmarks classification from upper gastrointestinal endoscopic images on edge devices. In: IEEE international conference on communications and electronics (ICCE), pp 389–394
LeCun Y, Bengio Y, Hinton G (2015) Deep learning. Nature 521(7553):436–444
Lee J, Kim S (2024) Dynamic learning of quantisation intervals for efficient neural networks. IEEE Trans Pattern Anal Mach Intell 46(1):58–72
Leenhardt R, Li C, Mouel J-PL, Rahmi G, Saurin JC, Cholet F, Boureille A, Amiot X, Delvaux M, Duburque C, Leandri C, Gérard R, Lecleire S, Mesli F, Nion-Larmurier I, Romain O, Sacher-Huvelin S, Simon-Shane C, Vanbiervliet G, Marteau P, Histace A, Dray X (2020) CAD-CAP: a 25,000-image database serving the development of artificial intelligence for capsule endoscopy. Endosc Int Open 8(3):415–420
Lertsinsrubtavee A, Ali A, Molina-Jimenez C, Sathiaseelan A, Crowcroft J (2017) Picasso: a lightweight edge computing platform. In: IEEE international conference on cloud networking (CloudNet), pp 1–7
Lestariningati SI (2018) Mobile point of sale design and implementation. IOP Conf Ser Mater Sci Eng 407:012094
Leufkens A, Van Oijen M, Vleggaar F, Siersema P (2012) Factors influencing the miss rate of polyps in a back-to-back colonoscopy study. Endoscopy 44(5):470–475
Li W, Liewig M (2020) A survey of ai accelerators for edge environment. In: Trends and innovations in information systems and technologies: vol 28. Springer, Cham, pp 35–44
Li M, Wong A (2022) Genetic algorithm based filter pruning for deep convolutional neural networks. Neural Netw 144:732–744
Li Y, Huang H, Xie Q, Yao L, Chen Q (2018) Research on a surface defect detection algorithm based on MobileNet-SSD. Appl Sci 8(9):1678
Li H, Zhu H, Liu P, Liu J (2020) EagleEye: fast sub-net evaluation for efficient neural network pruning. In: Proceedings of the European conference on computer vision (ECCV), pp 689–704
Li X, Yang Z, Zhang S (2021a) Efficientnet for real-time object detection in autonomous vehicles. In: Proceedings of the IEEE/CVF international conference on computer vision workshops
Li S, Zhang X, Sun J (2021b) Learning to prune: exploring the future of neural network compression. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR), pp 7437–7446
Li C, Li L, Jiang H, Weng K, Geng Y, Li L, Ke Z, Li Q, Cheng M, Nie W et al (2022) Yolov6: a single-stage object detection framework for industrial applications. arXiv:2209.02976
Li Z, Zhang Y, Ai J, Zhao Y, Yu Y, Dong Y (2023) A lightweight and explainable data-driven scheme for fault detection of aerospace sensors. IEEE Trans Aerosp Electron Syst 59(6):8392–8410
Lin W, Zhang S, Liu F (2022) Efficient object detection for cashierless checkout using edge-based transformers. Comput Ind 141:103432
Lingappa E, Parvathy LR (2022) Active contour neural network identifying MRI image edge computing methods deep learning bone cancer detection. In: International conference on advance computing and innovative technologies in engineering (ICACITE), pp 830–834
Liu T (2020) The applications and challenges of quantum teleportation. J Phys Conf Ser 1634:012089
Liu L, Deng J (2019) Dynamic deep neural networks: optimizing accuracy-efficiency trade-offs by selective execution. In: Proceedings of the AAAI conference on artificial intelligence, vol 32
Liu F, Sharma A (2024) Theoretical insights into sparsity and generalisation in neural networks. Mach Learn 105(1):159–189
Liu W, Anguelov D, Erhan D, Szegedy C, Reed S, Fu C-Y, Berg AC (2016) SSD: single shot multibox detector. In: European conference on computer vision (ECCV), pp 21–37
Liu C, Zoph B, Neumann M, Shlens J, Hua W, Li L-J, Fei-Fei L, Yuille A, Huang J, Murphy K (2018) Progressive neural architecture search. In: European conference on computer vision (ECCV), pp 19–34
Liu Z, Chen H, Zhang J (2021) Neural architecture search in automotive edge computing: Optimizing models for adas and autonomous driving. IEEE Trans Intell Veh 6(3):370–382
Liu D, Kong H, Luo X, Liu W, Subramaniam R (2022) Bringing AI to edge: from deep learning's perspective. Neurocomputing 485:297–320
Liu H, Wu C, Wang H (2023a) Real time object detection using lidar and camera fusion for autonomous driving. Sci Rep 13(1):8056
Liu Q, Zhou S, Lai J (2023b) EdgeMedNet: lightweight and accurate U-Net for implementing efficient medical image segmentation on edge devices. IEEE Trans Circuits Syst II Express Briefs 70(12):4329–4333
Lundberg SM, Lee S-I (2017) A unified approach to interpreting model predictions. In: Advances in Neural Information Processing Systems (NIPS), pp 1–10
Ma J, He Y, Li F, Han L, You C, Wang B (2024) Segment anything in medical images. Nat Commun 15(1):654
Mach P, Becvar Z (2017) Mobile edge computing: a survey on architecture and computation offloading. IEEE Commun Surv Tutor 19(3):1628–1656

Mahenge MP, Li C, Sanga CA (2019) Mobile edge computing: Cost-efficient content delivery in resource-constrained mobile computing environment. Int J Mobile Comput Multimedia Commun 10(3):23–46
Makoviychuk V, Wawrzyniak L, Guo Y, Lu M, Storey K, Macklin M, Hoeller D, Rudin N, Allshire A, Handa A, State G (2021) Isaac Gym: high performance GPU-based physics simulation for robot learning. arXiv:2108.10470
Mao Y, You C, Zhang J, Huang K, Letaief KB (2017) A survey on mobile edge computing: the communication perspective. IEEE IEEE Commun Surv Tutor 19(4):2322–2358
Markets (2024) Markets: edge computing in healthcare market size, share and trend. https://www.marketsandmarkets.com/Market-Reports/edge-computing-in-healthcare-market-133588379.html. Accessed 25 July 2024
Martini E, Boldo M, Aldegheri S, Valè N, Filippetti M, Smania N, Bertucco M, Picelli A, Bombieri N (2022) Enabling gait analysis in the telemedicine practice through portable and accurate 3D human pose estimation. Comput Methods Programs Biomed 225:107016
Masud M, Muhammad G, Hossain MS, Alhumyani H, Alshamrani SS, Cheikhrouhou O, Ibrahim S (2020) Light deep model for pulmonary nodule detection from CT scan images for mobile devices. Wirel Commun Mob Comput 2020:8893494
Mathe SE, Pamarthy AC, Kondaveeti HK, Vappangi S (2022) A review on Raspberry Pi and its robotic applications. In: International conference on artificial intelligence and signal processing (AISP), pp 1–6
Matloob Abbasi M, Iqbal S, Aurangzeb K, Alhussein M, Khan TM (2024) LMBiS-Net: a lightweight bidirectional skip connection based multipath cnn for retinal blood vessel segmentation. Sci Rep 14(1):15219
Matsubara Y, Yang R, Levorato M, Mandt S (2022) Supervised compression for resource-constrained edge computing systems. In: IEEE/CVF winter conference on applications of computer vision (WACV), pp 2685–2695
Mauri A, Khemmar R, Decoux B, Haddad M, Boutteau R (2022) Lightweight convolutional neural network for real-time 3D object detection in road and railway environments. J Real-Time Image Proc 19(3):499–516
McEnroe P, Wang S, Liyanage M (2022) A survey on the convergence of edge computing and ai for UAVs: opportunities and challenges. IEEE Internet Things J 9(17):15435–15459
Mehta S, Rastegari M (2022) MobileViT: light-weight, general-purpose, and mobile-friendly vision transformer. arXiv:2110.02178
Mehta S, Hajishirzi H, Rastegari M (2022) DiCENet: dimension-wise convolutions for efficient networks. IEEE Trans Pattern Anal Mach Intell 44(5):2416–2425
Mendez J, Bierzynski K, Cuéllar M, Morales DP (2022) Edge intelligence: concepts, architectures, applications, and future directions. ACM Trans Embedded Comput Syst 21(5):1–41
Mieras LF, Taal AT, Post EB, Ndeve AG, Van Hees CL (2018) The development of a mobile application to support peripheral health workers to diagnose and treat people with skin diseases in resource-poor settings. Trop Med Infect Dis 3(3):102
Min J, Chin LK, Oh J, Landeros C, Vinegoni C, Lee J, Lee SJ, Park JY, Liu A-Q, Castro CM, Lee H, Im H, Weissleder R (2020) CytoPAN—portable cellular analyses for rapid point-of-care cancer diagnosis. Sci Transl Med 12(555):9746
Miori L, Sanin J, Helmer S (2017) A platform for edge computing based on Raspberry Pi clusters. In: British international conference on databases (BICOD), pp 153–159
Mittal S (2019) A survey on optimized implementation of deep learning models on the NVIDIA Jetson platform. J Syst Architect 97:428–442
Mittapalli PS, Tagore M, Reddy PA, Kande GB, Reddy YM (2023) Deep learning based real-time object detection on Jetson Nano embedded GPU. In: Microelectronics, circuits and systems: select proceedings of micro 2021, pp 511–521
Mohanty SP, Hughes DP, Salathé M (2020) Image-based plant disease detection using deep learning. Front Plant Sci 7:1419
Momin MA, Junos MH, Mohd Khairuddin AS, Abu Talip MS (2023) Lightweight CNN model: automated vehicle detection in aerial images. SIViP 17(4):1209–1217
Mrozek D, Koczur A, Małysiak-Mrozek B (2020) Fall detection in older adults with mobile IoT devices and machine learning in the cloud and on the edge. Inf Sci 537:132–147
Muhammad K, Ahmad J, Baik SW (2018a) Early fire detection using convolutional neural networks during surveillance for effective disaster management. Neurocomputing 288:30–42
Muhammad K, Ahmad J, Mehmood I, Rho S, Baik SW (2018b) Convolutional neural networks based fire detection in surveillance videos. IEEE Access 6:18174–18183
Muñoz A, Rios R, Román R, López J (2023) A survey on the (in) security of trusted execution environments. Comput Secur 129:103180
Murshed MS, Murphy C, Hou D, Khan N, Ananthanarayanan G, Hussain F (2021) Machine learning at the network edge: a survey. ACM Comput Surv 54(8):1–37

Mustafa A, Sethi I (2005) Detecting retail events using moving edges. In: IEEE conference on advanced video and signal based surveillance (AVSS), pp 626–631
Mwansa PL, Alshaigy AO, Almaeeni DSM, Qasem KGH, Rego L, Nair P, Baniyas HAS (2022) Augmented reality delivers differential value in safety assurance on rigs onshore Abu Dhabi during Covid-19 pandemic courtesy of the wearable camera. In: Abu Dhabi international petroleum exhibition and conference, pp 011–002003
Nain G, Pattanaik K, Sharma G (2022) Towards edge computing in intelligent manufacturing: past, present and future. J Manuf Syst 62:588–611
Naqvi SS, Langah ZA, Khan HA, Khan MI, Bashir T, Razzak MI, Khan TM (2023) GLAN: GAN assisted lightweight attention network for biomedical imaging based diagnostics. Cogn Comput 15(3):932–942
Natarajan S, Chattopadhyay A, Seth S (2021) Mobilenetv3 for pneumonia detection on chest X-rays. Int J Mach Learn Comput 11(3):224–228
National Lung Screening Trial Research Team (2011) Reduced lung-cancer mortality with low-dose computed tomographic screening. N Engl J Med 365(5):395–409
Nayak S, Patgiri R, Waikhom L, Ahmed A (2022) A review on edge analytics: Issues, challenges, opportunities, promises, future directions, and applications. Digit Commun Netw 10:783–804
Nazeer M, Qayyum M, Ahad A (2022) Real time object detection and recognition in machine learning using Jetson Nano. Int J Innov Eng Manag Res 11(10):118–124
Ngeh CJ, Ma C, Ho TK-W, Wang Y, Raiti J (2020) Deep learning on edge device for early prescreening of skin cancers in rural communities. In: IEEE global humanitarian technology conference (GHTC), pp 1–4
Nguyen T, Pham Q (2024) Quantisation strategies for transformer-based language models. In: Proceedings of the annual meeting of the association for computational linguistics
Nguyen H, Pham L, Le Q (2024) Quantization strategies for transformer models in natural language processing. In: Proceedings of the ACL conference, pp 789–798
Nunez-Yanez J, Howard N (2021) Energy-efficient neural networks with near-threshold processors and hardware accelerators. J Syst Architect 116:102062
Nvidia: how edge computing is transforming healthcare. https://resources.nvidia.com/en-us-healthcare-and-edge-ai/healthcare-at-the-edge?ncid=no-ncid. Accessed 25 July 2024
Orlov NV, Chen WW, Eckley DM, Macura TJ, Shamir L, Jaffe ES, Goldberg IG (2010) Automatic classification of lymphoma images with transform-based global features. IEEE Trans Inf Technol Biomed 14(4):1003–1013
Oro D, Fernández C, Saeta JR, Martorell X, Hernando J (2011) Real-time GPU-based face detection in HD video sequences. In: IEEE international conference on computer vision (ICCV) workshops, pp 530–537
Ou S, Gao Y, Zhang Z, Shi C (2021) Polyp-YOLOv5-Tiny: a lightweight model for real-time polyp detection. In: 2021 IEEE 2nd international conference on information technology, big data and artificial intelligence (ICIBA), vol 2. IEEE, pp 1106–1111
Oueida S, Kotb Y, Aloqaily M, Jararweh Y, Baker T (2018) An edge computing based smart healthcare framework for resource management. Sensors 18(12):4307
Paluru N, Dayal A, Jenssen HB, Sakinis T, Cenkeramaddi LR, Prakash J, Yalavarthy PK (2021) Anam-Net: anamorphic depth embedding-based lightweight CNN for segmentation of anomalies in COVID-19 chest CT images. IEEE Trans Neural Netw Learn Syst 32(3):932–946
Pan J, Bulat A, Tan F, Zhu X, Dudziak L, Li H, Tzimiropoulos G, Martinez B (2022) EdgeViTs: competing light-weight CNNs on mobile devices with vision transformers. arXiv:2205.03436
Pandey A, Kumar R (2023) Practical approaches to mixed-precision quantisation for deep neural networks. J Mach Learn Res 24(3):123–145
Patchava V, Kandala HB, Babu PR (2015) A smart home automation technique with Raspberry Pi using IoT. In: International conference on smart sensors and systems (IC-SSS), pp 1–4
Pirandola S, Eisert J, Weedbrook C, Furusawa A, Braunstein SL (2015) Advances in quantum teleportation. Nat Photon 9(10):641–652
Pogorelov K, Randel KR, Griwodz C, Eskeland SL, Lange T, Johansen D, Spampinato C, Dang-Nguyen D-T, Lux M, Schmidt PT et al (2017) Kvasir: a multi-class image dataset for computer aided gastrointestinal disease detection. In: Proceedings of the 8th ACM on multimedia systems conference, pp 164–169
Radford A, Kim JW, Hallacy C, Ramesh A, Goh G, Agarwal S, Sastry G, Askell A, Mishkin P, Clark J et al (2021) Learning transferable visual models from natural language supervision. In: International conference on machine learning. PMLR, pp 8748–8763
Raghavan D, Wheeler M, Doege D, Doty JD, Levy H, Dungan KA, Davis LM, Robinson JM, Kim ES, Mileham KF, Oliver J, Carrizosa D (2020) Initial results from mobile low-dose computerized tomographic lung cancer screening unit: Improved outcomes for underserved populations. Oncologist 25(5):777–781
Raj S, Padhi S, Simmhan Y (2023) Ocularone: Exploring drones-based assistive technologies for the visually impaired. In: Extended abstracts of the chi conference on human factors in computing systems, p 220

Rajpurkar P, Irvin J, Zhu K, Yang B, Mehta H, Duan T, Ding D, Bagul A, Langlotz C, Shpanskaya K et al (2019) Chexnet: radiologist-level pneumonia detection on chest X-rays with deep learning. PLoS Med 14(11):1002686
Ramey J, Fung KM, Hassell LA (2011) Use of mobile high-resolution device for remote frozen section evaluation of whole slide images. J Pathol Inf 2(1):41
Ranaweera P, Jurcut AD, Liyanage M (2021) Survey on multi-access edge computing security and privacy. IEEE Commun Surv Tutor 23(2):1078–1124
Ray PP (2022) A review on TinyML: state-of-the-art and prospects. J King Saud Univ Comput Inf Sci 34(4):1595–1623
Real E, Aggarwal A, Huang Y, Le QV (2019a) AmoebaNet: evolutionary neural architecture search. In: IEEE conference on computer vision and pattern recognition (CVPR), pp 7749–7757
Real E, Aggarwal A, Huang Y, Le QV (2019b) Regularized evolution for image classifier architecture search. In: Proceedings of the AAAI conference on artificial intelligence, vol 33, pp 4780–4789
Redmon J, Farhadi A (2018) Yolov3: an incremental improvement. arXiv:1804.02767
Redmon J, Divvala S, Girshick R, Farhadi A (2016) You only look once: Unified, real-time object detection. In: IEEE conference on computer vision and pattern recognition (CVPR), pp 779–788
Research PM Edge Computing In Healthcare Market. https://www.polarismarketresearch.com/industry-analysis/edge-computing-in-healthcare-market. Accessed 25 July 2024
Reuther A, Michaleas P, Jones M, Gadepally V, Samsi S, Kepner J (2021) AI accelerator survey and trends. In: 2021 IEEE high performance extreme computing conference (HPEC). IEEE, pp 1–9
Ribeiro MT, Singh S, Guestrin C (2016) "Why should I trust you?" Explaining the predictions of any classifier. In: ACM SIGKDD international conference on knowledge discovery and data mining (KDD), pp 1135–1144
Sadique KM (2013) Secure mobile POS system: a point of sale application for secure financial transitions in a mobile business enviroment. Master Thesis, KTH Royal Institute of Technology, pp 1–56
Sahafi A, Wang Y, Rasmussen C, Bollen P, Baatrup G, Blanes-Vidal V, Herp J, Nadimi E (2022) Edge artificial intelligence wireless video capsule endoscopy. Sci Rep 12(1):13723
Saini N, Chattopadhyay C, Das D (2023) E2AlertNet: an explainable, efficient, and lightweight model for emergency alert from aerial imagery. Remote Sens Appl Soc Environ 29:100896
Sait U, Shivakumar S, KV GL, Kumar T, Ravishankar VD, Bhalla K (2019) A mobile application for early diagnosis of pneumonia in the rural context. In: IEEE global humanitarian technology conference (GHTC), pp 1–5
Sandler M, Howard A, Zhu M, Zhmoginov A, Chen L (2018) Mobilenetv2: inverted residuals and linear bottlenecks. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 4510–4520
Sarma J, Biswas R (2020) Vlsi based adaptive power management architecture for ecg monitoring in WBAN. In: 2020 33rd International conference on VLSI design and 2020 19th international conference on embedded systems (VLSID). IEEE, pp 113–118
Sati V, Sánchez SM, Shoeibi N, Arora A, Corchado JM (2021) Face detection and recognition, face emotion recognition through NVIDIA Jetson Nano. In: International symposium on ambient intelligence (ISAmI), pp 177–185
Schizas N, Karras A, Karras C, Sioutas S (2022) TinyML for ultra-low power AI and large scale IoT deployments: a systematic review. Future Internet 14(12):363
Schneider B, Banerjee T (2018) Activity recognition using imagery for smart home monitoring. In: Advances in soft computing and machine learning in image processing, pp 355–371
Seguí S, Drozdzal M, Pascual G, Radeva P, Malagelada C, Azpiroz F, Vitrià J (2016) Generic feature learning for wireless capsule endoscopy analysis. Comput Biol Med 79:163–172
Selmanaj E, Sommen F, Okel SE, Putten J, Struyvenberg MR, Bergman JJGHM, With PHN (2021) Fast tissue detection in volumetric laser endomicroscopy using convolutional neural networks: an object-detection approach. SPIE Med Imaging:Image Process 11596:854–860
Senan EM, Jadhav ME, Kadam A (2021) Classification of PH2 images for early detection of skin diseases. In: 2021 6th international conference for convergence in technology (I2CT). IEEE, pp 1–7
Shahzadi S, Iqbal M, Dagiuklas T, Qayyum ZU (2017) Multi-access edge computing: open issues, challenges and future perspectives. J Cloud Comput 6:1–13
Shen H, Chen L, Jin Y, Zhao L, Kong B, Philipose M, Krishnamurthy A, Sundaram R (2019) Nexus: a GPU cluster engine for accelerating DNN-based video analysis. In: ACM symposium on operating systems principles (SOSP), pp 322–337
Shen M, Liang F, Gong R, Li Y, Li C, Lin C, Yu F, Yan J, Ouyang W (2021) Once quantization-aware training: high performance extremely low-bit architecture search. In: International conference on computer vision (ICCV 2021), pp 5340–5349

Shen Z, Howard N, Nunez-Yanez J (2022) Big-little adaptive neural networks on low-power near-subthreshold processors. J Low Power Electron Appl 12(2):28
Shi W, Cao J, Zhang Q, Li Y, Xu L (2016) Edge computing: vision and challenges. IEEE Internet Things J 3(5):637–646
Shi Y, Li X, Chen S (2023) Skin lesion intelligent diagnosis in edge computing networks: A federated contrastive learning approach. ACM Transactions on Intelligent Systems and Technology 14(4):69
Shrivastava VK et al (2023) Skin disease classification using deep convolutional neural network on jetson nano developer kit. In: 2023 IEEE 3rd International conference on applied electromagnetics, signal processing, & communication (AESPC). IEEE, pp 1–4
Shuvo MMH, Islam SK, Cheng J, Morshed BI (2022) Efficient acceleration of deep learning inference on resource-constrained edge devices: a review. Proc IEEE 111(1):42–91
Simões F, Bouveyron C, Precioso F (2023) Deepwild: wildlife identification, localisation and estimation on camera trap videos using deep learning. Eco Inform 75:102095
Simonyan K, Zisserman A (2014) Very deep convolutional networks for large-scale image recognition. arXiv:1409.1556
Singh A, Chatterjee K (2021) Securing smart healthcare system with edge computing. Comput Secur 108:102353
Sipola T, Alatalo J, Kokkonen T, Rantonen M (2022) Artificial intelligence in the IoT era: a review of edge AI hardware and software. In: Conference of open innovations association (FRUCT), pp 320–331
Smith J, Gupta A (2023) Integrated framework for network compression via pruning, quantisation, and knowledge distillation. In: Proceedings of the IEEE conference on computer vision and pattern recognition
Solanki N, Patel C, Tailor N, Pathan N (2021) Performance analysis of SOC and hardware design flow in medical image processing using Xilinx ZedBoard FPGA. In: Proceedings of 2nd international conference on computing, communications, and cyber-security: IC4S 2020. Springer, pp 945–966
Srivastava G, K DR, Yenduri G, Hegde P, Gadekallu TR, Maddikunta PKR, Bhattacharya S (2023) Federated learning enabled edge computing security for Internet of medical things: concepts, challenges and open issues. In: Security and risk analysis for intelligent edge computing, pp 67–89
Su L, Xu W, Li P, Zeng X (2020) Real-time ECG classification on edge devices with a lightweight neural network model. J Healthc Eng 2020:1–12
Su J, Zhu X, Li S, Chen W-H (2023) AI meets UAVs: a survey on ai empowered UAV perception systems for precision agriculture. Neurocomputing 518:242–270
Sun L, Jiang X, Ren H, Guo Y (2020) Edge-cloud computing and artificial intelligence in internet of medical things: Architecture, technology and application. IEEE Access 8:101079–101092
Swager A-F, Sommen F, Klomp SR, Zinger S, Meijer SL, Schoon EJ, Bergman JJ, With PH, Curvers WL (2017) Computer-aided detection of early Barrett's neoplasia using volumetric laser endomicroscopy. Gastrointest Endosc 86(5):839–846
Tabassum N, Islam SMR, Bulbul F (2023) Brain tumor detection from brain mri using soft IP core on FPGA. Circuits Syst Signal Process 42(2):724–747
Tan M, Le QV (2019) EfficientNet: Rethinking model scaling for convolutional neural networks. arXiv:1905.11946
Tan M, Le QV (2020a) Efficientnet-lite: Improved accuracy and efficiency with optimized mobile models. In: arXiv:2004.02984
Tan M, Le QV (2020b) Efficientnet: Rethinking model scaling for convolutional neural networks. In: Proceedings of the International Conference on Machine Learning, pp. 6105–6114
Tan M, Le QV (2021) Efficientnetv2: Smaller models and faster training. In: Proceedings of the 38th International Conference on Machine Learning, pp. 10096–10106
Tan M, Yu R (2019) MixConv: Mixed depthwise convolutional kernels. In: IEEE/CVF International Conference on Computer Vision (ICCV), pp. 5547–5555
Tan M, Chen B, Pang R, Vasudevan V, Sandler M, Howard A, Le QV (2019) MnasNet: platform-aware neural architecture search for mobile. In: IEEE conference on computer vision and pattern recognition (CVPR), pp 2820–2828
Tang J, Ren Y, Liu S (2017) Real-time robot localization, vision, and speech recognition on Nvidia Jetson TX1. arXiv:1705.10945
Tang M, Xin Y (2023) Efficient energy consumption optimization for wireless sensor health monitoring system in mobile edge computing. IEEE Internet Things J 11(5):7948–7955
Tschandl P, Rosendahl C, Kittler H (2018) The HAM10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions. Scientific Data 5:180161
Tulasi D, Granados A, Gunawardane P, Kashyap A, McDonald Z, Thulasidasan S (2023) Smart camera traps: enabling energy-efficient edge-AI for remote monitoring of wildlife. In: ACM SIGSPATIAL international workshop on AI-driven spatio-temporal data analysis for wildlife conservation (GeoWildLife), pp 9–16

Ullah I, Khan MA, Alkhalifah A, Nordin R, Alsharif MH, Alghtani AH, Aly AA (2021) A multi-message multi-receiver signcryption scheme with edge computing for secure and reliable wireless Internet of medical things communications. Sustainability 13(23):13184
Urban G, Tripathi P, Alkayali T, Mittal M, Jalali F, Karnes W, Baldi P (2018) Deep learning localizes and identifies polyps in real time with 96% accuracy in screening colonoscopy. Gastroenterology 155(4):1069–1078
Vairo T, Pettinato M, Reverberi AP, Milazzo MF, Fabiano B (2023) An approach towards the implementation of a reliable resilience model based on machine learning. Process Saf Environ Prot 172:632–641
Van Hooren B, Pecasse N, Meijer K, Essers JMN (2023) The accuracy of markerless motion capture combined with computer vision techniques for measuring running kinematics. Scandinavian Journal of Medicine & Science in Sports 33(6):966–978
Van Netten JJ, Clark D, Lazzarini PA, Janda M, Reed LF (2017) The validity and reliability of remote diabetic foot ulcer assessment using mobile phone images. Sci Rep 7:9480
Varghese B, Wang N, Bermbach D, Hong C.-H, Lara E, Shi W, Stewart C (2020) A survey on edge benchmarking. arXiv:2004.11725
Vázquez FI, Kastner W (2012) Thermal comfort support application for smart home control. In: International symposium on ambient intelligence (ISAmI), pp 109–118
Vázquez D, Bernal J, Sánchez F.J, Fernández-Esparrach G, López A.M, Romero A, Drozdzal M, Courville A (2017) A benchmark for endoluminal scene segmentation of colonoscopy images. J Healthc Eng 2017:4037190
Verma GK, Gupta P (2018) Wild animal detection using deep convolutional neural network. In: International conference on computer vision & image processing (CVIP), pp 327–338
Voulodimos A, Doulamis N, Doulamis A, Protopapadakis E (2018) Deep learning for computer vision: a brief review. Comput Intell Neurosci 2018:7068349
Wan S, Ding S, Chen C (2022) Edge computing enabled video segmentation for real-time traffic monitoring in internet of vehicles. Pattern Recogn 121:108146
Wang B, Huang F (2021) A lightweight deep network for defect detection of insert molding based on X-ray imaging. Sensors 21(16):5612
Wang E, Lee D (2024) Hybrid compression: Integrating sparsity and quantisation for efficient neural networks. Neural Netw 78:34–50
Wang L, Li Q (2021) Energy-aware pruning for deep convolutional neural networks. IEEE Trans Sustain Comput 6(1):112–124
Wang P, Berzin TM, Brown JRG, Bharadwaj S, Becq A, Xiao X, Liu P, Li L, Song Y, Zhang D, Li Y, Xu G, Tu M, Liu X (2019a) Real-time automatic detection system increases colonoscopic polyp and adenoma detection rates: a prospective randomised controlled study. Gut 68(10):1813–1819
Wang H, Jiang C, Bao K, Xu C (2019b) Recognition and clinical diagnosis of cervical cancer cells based on our improved lightweight deep network for pathological image. J Med Syst 43:301
Wang X, Han Y, Leung VC, Niyato D, Yan X, Chen X (2020a) Convergence of edge computing and deep learning: a comprehensive survey. IEEE Commun Surv Tutor 22(2):869–904
Wang P, Liu P, Brown JRG, Berzin TM, Zhou G, Lei S, Liu X, Li L, Xiao X (2020b) Lower adenoma miss rate of computer-aided detection-assisted colonoscopy vs routine white-light colonoscopy in a prospective tandem study. Gastroenterology 159(4):1252–1261
Wang F, Zhang M, Wang X, Ma X, Liu J (2020c) Deep learning for edge computing applications: a state-of-the-art survey. IEEE Access 8:58322–58336
Wang T, Hu Y, Liu G, Cao L (2020d) Efficientnet for diabetic retinopathy detection. IEEE Journal of Biomedical and Health Informatics
Wang L, Xiang L, Xu J, Chen J, Zhao X, Yao D, Wang X, Li B (2020e) Context-aware deep model compression for edge cloud computing. In: IEEE international conference on distributed computing systems (ICDCS), pp 787–797
Wang R, Wang Z, Xu Z, Wang C, Li Q, Zhang Y, Li H (2021a) A real-time object detector for autonomous vehicles based on YOLOv4. Comput Intell Neurosci 2021:9218137
Wang Z, Zhang Z, Cui L (2021b) Real-time object detection on mobile devices for precision agriculture using deep learning. IEEE Access 9:36602–36612
Wang R, Zhang J, Chen J, Xu Y, Li P, Liu T, Wang H (2023) DexGraspNet: a large-scale robotic dexterous grasp dataset for general objects based on simulation. In: IEEE International conference on robotics and automation (ICRA), pp 11359–11366
Wen H, Li Y, Zhang Z, Jiang S, Ye X, Ouyang Y, Zhang Y, Liu Y (2023) AdaptiveNet: post-deployment neural architecture adaptation for diverse edge environments. In: ACM annual international conference on mobile computing and networking (MobiCom), pp 1–17

Winzig J, Almanza JCA, Mendoza MG, Schumann T (2022) Edge AI—use case on Google Coral Dev Board Mini. In: IET international conference on engineering technologies and applications (IET-ICETA), pp 1–2
Wistuba M, Rawat A, Pedapati T (2019) A survey on neural architecture search. arXiv:1905.01392
Wu B, Dai X, Zhang P, Wang Y, Sun F, Wu Y, Tian Y, Vajda P, Jia Y (2019a) FBNet: Hardware-aware efficient ConvNet design via differentiable neural architecture search. In: IEEE conference on computer vision and pattern recognition (CVPR), pp 10734–10742
Wu X, Zhan C, Lai Y-K, Cheng M-M, Yang J (2019b) IP102: a large-scale benchmark dataset for insect pest recognition. In: IEEE/CVF conference on computer vision and pattern recognition (CVPR), pp 8787–8796
Wu B, Dai X, Wan A, Zhang P, Wu P, Keutzer K (2021) Tinynas: a framework for fast, automated, and efficient neural architecture search. arXiv:2012.11281
Wu H, Zhao Z, Zhong J, Wang W, Wen Z, Qin J (2022) Polypseg+: a lightweight context-aware network for real-time polyp segmentation. IEEE Trans Cybern 53(4):2610–2621
Xia B, Cao J, Wang C (2019) SSIM-NET: real-time PCB defect detection based on SSIM and MobileNet-V3. In: World conference on mechanical engineering and intelligent manufacturing (WCMEIM), pp 756–759
Xie Y, Hu Y, Chen Y, Liu Y, Shou G (2018) A video analytics-based intelligent indoor positioning system using edge computing for IoT. In: International conference on cyber-enabled distributed computing and knowledge discovery (CyberC), pp 118–125
Xu J, Hu Z, Zou Z, Zou J, Hu X, Liu L, Zheng L (2020) Design of smart unstaffed retail shop based on iot and artificial intelligence. IEEE Access 8:147728–147737
Xu L, Wang H, Zheng Y (2021) Real-time medical image analysis on edge devices using mobile transformer networks. J Med Syst 45(3):1–12
Yao J, Zhang S, Yao Y, Wang F, Ma J, Zhang J, Chu Y, Ji L, Jia K, Shen T, Wu A, Zhang F, Tan Z, Kuang K, Wu C, Wu F, Zhou J, Yang H (2022) Edge-cloud polarization and collaboration: a comprehensive survey for AI. IEEE Trans Knowl Data Eng 35(7):6866–6886
Yap MH, Chatwin KE, Ng C-C, Abbott CA, Bowling FL, Rajbhandari S, Boulton AJ, Reeves ND (2018) A new mobile application for standardizing diabetic foot images. J Diabetes Sci Technol 12(1):169–173
Ye J, He L, Beestrum M (2023) Implications for implementation and adoption of telehealth in developing countries: a systematic review of china's practices and experiences. NPJ Digit Med 6(1):174
Yi X, Peng C, Zhang Z, Xiao L (2022) The defect detection for X-ray images based on a new lightweight semantic segmentation network. Math Biosci Eng 19(4):4178–4195
You C, Yang Q, Shan H, Gjesteby L, Li G, Ju S, Zhang Z, Zhao Z, Zhang Y, Cong W et al (2018) Structurally-sensitive multi-scale deep neural network for low-dose ct denoising. IEEE Access 6:41839–41855
You H, Chen X, Zhang Y, Li C, Li S, Liu Z, Wang Z, Lin Y (2020a) ShiftAddNet: a hardware-inspired deep network. Adv Neural Inf Process Syst 33:2771–2783
You C, Yang J, Chapiro J, Duncan JS (2020b) Unsupervised wasserstein distance guided domain adaptation for 3d multi-domain liver segmentation. In: Interpretable and Annotation-Efficient Learning for Medical Image Computing: Third International Workshop, iMIMIC 2020, Second International Workshop, MIL3ID 2020, and 5th International Workshop, LABELS 2020, Held in Conjunction with MICCAI 2020, Lima, Peru, October 4–8, 2020, Proceedings 3, pp. 155–163. Springer
You C, Xiang J, Su K, Zhang X, Dong S, Onofrey J, Staib L, Duncan JS (2021) Incremental learning meets transfer learning: application to multi-site prostate mri segmentation. arXiv:2206.01369
You C, Zhao R, Liu F, Dong S, Chinchali S, Topcu U, Staib L, Duncan J (2022a) Class-aware adversarial transformers for medical image segmentation. Adv Neural Inf Process Syst 35:29582–29596
You C, Zhou Y, Zhao R, Staib L, Duncan JS (2022b) Simcvd: Simple contrastive voxel-wise representation distillation for semi-supervised medical image segmentation. IEEE Trans Med Imaging 41(9):2228–2237
You C, Dai W, Liu F, Min Y, Su H, Zhang X, Li X, Clifton DA, Staib L, Duncan JS (2022c) Mine your own anatomy: revisiting medical image segmentation with extremely limited labels. arXiv:2209.13476
You C, Zhao R, Staib LH, Duncan JS (2022d) Momentum contrastive voxel-wise representation learning for semi-supervised volumetric medical image segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention, pp. 639–652. Springer
You C, Dai W, Min Y, Staib L, Duncan JS (2023a) Bootstrapping semi-supervised medical image segmentation with anatomical-aware contrastive distillation. In: International conference on information processing in medical imaging. Springer, pp 641–653
You C, Dai W, Min Y, Staib L, Duncan JS (2023b) Implicit anatomical rendering for medical image segmentation with stochastic experts. In: International conference on medical image computing and computer-assisted intervention. Springer, pp 561–571

You C, Dai W, Min Y, Staib L, Sekhon J, Duncan JS (2023c) Action++: improving semi-supervised medical image segmentation with adaptive anatomical contrast. In: International conference on medical image computing and computer-assisted intervention. Springer, pp 194–205
You C, Dai W, Min Y, Liu F, Clifton D, Zhou S.K, Staib L, Duncan J (2024) Rethinking semi-supervised medical image segmentation: a variance-reduction perspective. In: Advances in neural information processing systems, vol 36
Yousri R, Elbayoumi M, Soltan A, Darweesh MS (2023) A power-aware task scheduler for energy harvesting-based wearable biomedical systems using snake optimizer. Analog Integr Circ Sig Process 115(2):183–194
Yu W, Liang F, He X, Hatcher WG, Lu C, Lin J, Yang X (2017) A survey on the edge computing for the internet of things. IEEE Access 6:6900–6919
Yu J, Yang L, Xu N, Yang J, Huang T (2019) Slimmable neural networks. arXiv:1812.08928
Yu F, Chen H, Wang X, Xian W, Chen Y, Liu F, Madhavan V, Darrell T (2020) BDD100k: a diverse driving video database with scalable annotation tooling. In: IEEE/CVF conference on computer vision and pattern recognition (CVPR), pp 2636–2645
Zhang N, Izquierdo E (2023) A four-point camera calibration method for sport videos. IEEE Trans Circuits Syst Video Technol 33(8):3811–3821
Zhang H, Qie Y (2023) Applying deep learning to medical imaging: a review. Appl Sci 13(18):10521
Zhang J, Chen B, Zhao Y, Cheng X, Hu F (2018a) Data security and privacy-preserving in edge computing paradigm: Survey and open issues. IEEE Access 6:18209–18237
Zhang X, Zhou X, Lin M, Sun J (2018b) ShuffleNet: an extremely efficient convolutional neural network for mobile devices. In: IEEE conference on computer vision and pattern recognition (CVPR), pp 6848–6856
Zhang Y-M, Lee C-C, Hsieh J-W, Fan K-C (2021a) CSL-YOLO: a new lightweight object detection system for edge computing. arXiv:2107.04829
Zhang C, Xie Y, Bai H, Yu B, Li W, Gao Y (2021b) A survey on federated learning. Knowl-Based Syst 216:106775
Zhang J, Guo D, Wu Y, Xu X, Liu H (2023) Toward lifelong learning for industrial defect classification: a proposed framework. IEEE IEEE Robot Autom Mag 30(2):10–21
Zhao W, Liu Y (2023) Bayesian sparsity pruning for deep neural networks. J Mach Learn Res 24(1):557–579
Zhao CW, Jegatheesan J, Loon SC (2015) Exploring IoT application using Raspberry Pi. Int J Comput Netw Appl 2(1):27–34
Zhao Z, Jiang Z, Ling N, Shuai X, Xing G (2018) ECRT: An edge computing system for real-time image-based object tracking. In: ACM Conference on Embedded Networked Sensor Systems (SenSys), pp. 394–395
Zhao Y, Chen J, Xu Z (2020) Real-time object recognition on mobile robots using mobilenet and deep learning. In: Proceedings of the 2020 International Conference on Robotics and Automation (ICRA), pp. 5636–5641. IEEE
Zhao Z, Liu Y, Wu H, Wang M, Li Y, Wang S, Teng L, Liu D, Cui Z, Wang Q et al (2023) Clip in medical imaging: A comprehensive survey. arXiv:2312.07353
Zheng X, Shah SBH, Ren X, Li F, Nawaf L, Chakraborty C, Fayaz M (2021) Mobile edge computing enabled efficient communication based on federated learning in internet of medical things. Wirel Commun Mob Comput 2021:1–10
Zhou M, Zhang J (2023) Adaptive quantisation of neural networks using data-driven bit-width allocation. Neural Process Lett 47(1):201–216
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
Zhu J, Jiang J, Chen X, Tsui C-Y (2018) Sparsenn: an energy-efficient neural network accelerator exploiting input and output sparsity. In: 2018 design, automation & test in Europe conference & exhibition (DATE). IEEE, pp 241–244
Zou Z, Zhang R, Shen S, Pandey G, Chakravarty P, Parchami A, Liu HX (2022) Real-time full-stack traffic scene perception for autonomous driving with roadside cameras. In: International conference on robotics and automation (ICRA), pp 890–896
Zúñiga Espinoza C, Khot LR, Sankaran S, Jacoby PW (2017) High resolution multispectral and thermal remote sensing-based water stress assessment in subsurface irrigated grapevines. Remot Sens 9(9):961

## Authors and Affiliations

**Yiwen Xu[1] · Tariq M. Khan[1] · Yang Song[1] · Erik Meijering[1]**

✉ Yiwen Xu
yiwen.xu1@unsw.edu.au

Tariq M. Khan
tariq.khan@unsw.edu.au

Yang Song
yang.song1@unsw.edu.au

Erik Meijering
erik.meijering@unsw.edu.au

[1] School of Computer Science and Engineering, University of New South Wales, Sydney, NSW 2052, Australia