Internet of Things 36 (2026) 101876

Contents lists available at ScienceDirect

Internet of Things

journal homepage: www.elsevier.com/locate/iot

Review article

Creation of AI-driven smart spaces for enhanced indoor environments – A survey

Aygün Varol [a,*], Naser Hossein Motlagh [b,*], Mirka Leino [c], Sasu Tarkoma [b], Johanna Virkki [a]

[a] *Faculty of Information Technology and Communication Sciences, Tampere University, Tampere, Finland*
[b] *Department of Computer Science, University of Helsinki,*
[c] *Satakunta University of Applied Sciences,*

ARTICLE INFO

Keywords:
Smart spaces
Internet of things
Sensor networks
Indoor environments
Large language models
Transformer-based networks

ABSTRACT

Smart spaces are ubiquitous computing environments that integrate diverse sensing and communication technologies to enhance functionality, optimize energy utilization, and improve user comfort and well-being. The adoption of emerging artificial intelligence (AI) methodologies has led to the development of AI-driven smart spaces, further expanding capabilities through applications such as personalized comfort settings, interactive living spaces, and automation of space systems. These advancements collectively elevate the quality of indoor experiences for users. To systematically examine these developments, we present a comprehensive survey of the foundational components of AI-driven smart spaces, including sensor technologies, data communication protocols, network management and maintenance strategies, and data collection, processing, and analytics. We investigate both traditional machine learning (ML) methods, such as deep learning (DL), and emerging approaches, including transformer networks and large language models (LLMs), highlighting their contributions and potential. We also showcase real-world applications of these technologies and provide insights to guide their continued development. Each section details relevant technologies and methodologies and concludes with an analysis of challenges and limitations, identifying directions for future research.

1. Introduction

Smart spaces are defined as ubiquitous computing environments that enhance indoor settings by seamlessly integrating sensors, actuators, and communication technologies into the physical environment to deliver intelligent services [1]. These systems leverage sensing mechanisms, communication protocols, and pervasive computing approaches to automate various indoor functions, including lighting and Heating, Ventilation, and Air Conditioning (HVAC) systems. Based on the literature, such environments also support real-time data processing and analytics to create intelligent, adaptive living spaces [2].

While existing smart spaces provide significant automation benefits, their designs often prioritize sensor-driven responses over user-centric needs. These systems typically rely on rule-based automation and manual configuration-such as turning lighting systems on or off based on predefined thresholds-without adapting dynamically to user preferences or behaviors [3,4]. Consequently, they fall short in utilizing the extensive data generated by the diverse array of sensors deployed in these environments. The lack of robust

[*] Corresponding authors.
E-mail addresses: aygun.varol@tuni.fi (A. Varol), naser.motlagh@helsinki.fi (N. Hossein Motlagh), mirka.leino@samk.fi (M. Leino), sasu.tarkoma@helsinki.fi (S. Tarkoma), johanna.virkki@tuni.fi (J. Virkki).

https://doi.org/10.1016/j.iot.2026.101876
Received 3 September 2025; Received in revised form 30 December 2025; Accepted 15 January 2026
Available online 18 January 2026

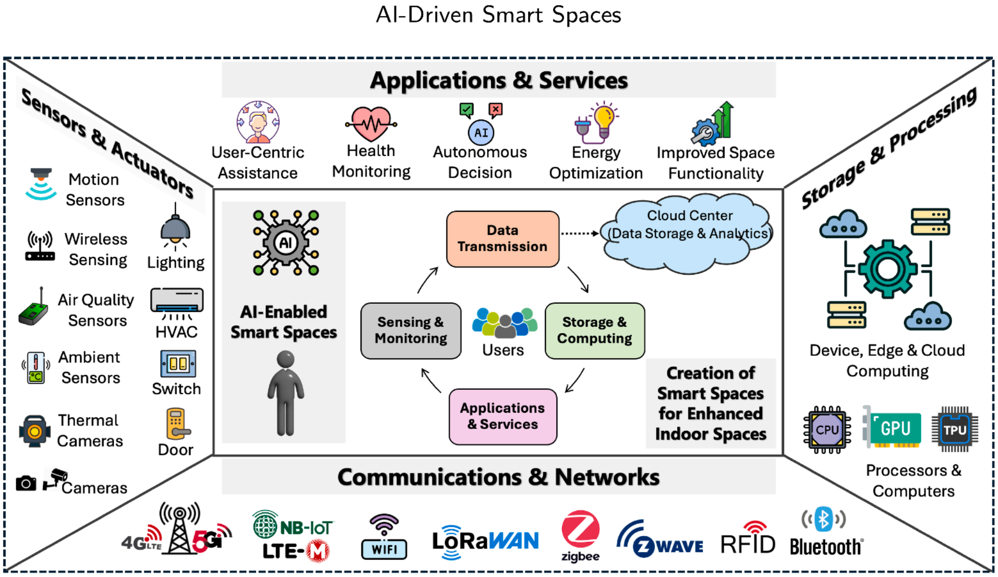

Fig. 1. The necessary components needed for creation of AI-driven smart spaces.

mechanisms for interpreting data and generating actionable insights limits their capacity to transcend basic automated actions and evolve into truly adaptive systems.

The integration of Artificial Intelligence (AI) into smart spaces holds the potential to revolutionize these environments, enabling them to transition into fully intelligent ecosystems. AI-driven smart spaces can support real-time decision-making, predict and respond to user needs, and enhance personalization. By leveraging AI, these spaces can achieve higher efficiency, functionality, and user-centric adaptability, leading to improved safety, well-being, and comfort for residents.

AI-driven smart spaces (depicted in Fig. 1) enable diverse applications, including activity monitoring, anomaly detection, energy optimization [5–9], occupancy detection [10–12], user behavior prediction [13], interactive environment creation [14], personalized recommendation systems [15], advanced care solutions [16], and user-centric adaptive frameworks [17]. They build upon existing smart space technologies and integrate emerging methodologies, such as Internet of Things (IoT) frameworks, machine learning (ML), deep learning (DL), transformer networks, and large language models (LLMs). These advancements enable continuous learning from user behaviors to optimize operations, fostering functional, interactive, healthy, and sustainable indoor environments.

Despite their promise, developing AI-driven smart spaces poses several challenges alongside significant opportunities. This article provides a comprehensive review of components and technologies, such as sensor systems and communication protocols, with an emphasis on emerging AI methodologies, such as transformer networks and LLMs, essential for building AI-driven smart spaces. We explore current research, analyze existing methodologies and solutions, and identify emerging trends alongside the challenges and limitations of each technology. To advance the development of AI-enabled smart spaces, we also propose future research directions and present a roadmap for the path forward.

2. Scope of survey

2.1. Related surveys

Due to the growing popularity and importance of smart spaces, numerous studies have been conducted in this domain. Table 1 provides a summary of existing works that intersect with our study.

State-of-the-art research often emphasizes survey studies focusing on practical machine learning (ML) and deep learning (DL) applications, addressing the need for diverse sensing systems to enable smart space applications [18]. Some studies specifically survey ML techniques for smart lighting applications [20] or introduce sensor technologies and deployment strategies leveraging AI methodologies [19]. Other research concentrates on human activity recognition (HAR) within smart spaces, exploring public datasets, sensor requirements, and traditional ML models [21,22]. Surveys also explore HAR applications in resource-constrained smart living environments, focusing on sensing technologies and sensor communication frameworks [23–25]. In addition, several studies examine the application of AI methodologies such as deep reinforcement learning (DRL) for energy management in smart buildings, addressing challenges such as energy efficiency, carbon emission reduction, and user comfort [26–30]. Other research investigates indoor air quality, correlating it with energy consumption and demonstrating how AI enhances air quality while optimizing energy utilization in smart spaces [31–33].

While the existing literature offers valuable insights, it has notable limitations. First, most surveys narrow their scope to specific applications, such as human activity recognition or energy optimization. Second, they often overlook essential components such as

Table 1
Existing related surveys in smart spaces.

Category	Scope	Coverage Gap
Sensing Applications	DL for behavior detection [18] Occupancy recognition [19]	Limited AI diversity Missing LLM
Smart Lighting	ML-based comfort control [20]	Single AI technology
Human Activity Recognition	Conventional ML models [21,22] Resource-constrained IoT environments [23–25]	Missing transformers Single AI application
Energy Management	DRL for HVAC control and residential optimization [26–30]	Single AI technology
Indoor Air Quality	AI-driven IAQ monitoring [31] ML/DL optimization [32,33]	Missing SOTA AI Single AI Application

communication technologies, protocols, and data collection and processing strategies, which are critical for developing AI-powered smart spaces. Third, the intricacies of advanced AI models, such as Transformer networks, are frequently neglected, despite their pivotal role in transforming smart spaces into fully intelligent and interactive spaces.

In contrast, our work takes a broader perspective, encompassing diverse applications of smart spaces. We introduce the foundational components required to build AI-driven smart environments and highlight the significance of both conventional ML approaches and emerging technologies, such as Transformer networks and large language models (LLMs). By providing a comprehensive overview, our study underscores the deployment of cutting-edge technologies in creating intelligent, interactive, and efficient smart spaces.

2.2. *Selection of articles*

To provide a comprehensive overview of AI-driven smart spaces, we adopted a systematic methodology that ensured both breadth and depth in literature coverage. The process encompassed our selection and categorization of articles as well as the formulation of a structured survey framework. Our article selection procedure (summarized in Table 2) involved querying major scholarly databases-Google Scholar, IEEE Xplore, ACM Digital Library, and ScienceDirect-for studies published between January 2010 and September 2025. We combined keywords related to AI and ML techniques (e.g., artificial intelligence, machine learning, deep learning, transformer networks, generative pre-trained transformer networks, large language models, long short-term memory, recurrent neural networks, natural language processing) with terms associated with smart indoor environments (e.g., smart home, smart building, smart office). For instance, we used queries such as “LLMs in smart homes” and “Transformer networks in indoor environments.”

After aggregating results from all databases, we removed duplicates and included only peer-reviewed journal and conference papers directly addressing AI or ML applications in indoor smart spaces, excluding preprints and studies lacking direct relevance. However, given the rapidly evolving nature of AI technologies-particularly LLMs and generative AI-we selectively incorporated preprints and white papers from reputable sources (e.g., arXiv) only as supplementary references to capture the latest advancements in areas where peer-reviewed literature has not yet been established. To ensure comprehensive and current coverage, we conducted subsequent searches through September 2025, prioritizing recent and highly cited literature. We examined each selected article in detail and categorized them into eight thematic areas, including Human Activity Recognition, Energy Optimization, Air Quality Monitoring, and Generative AI for Smart Spaces. This process resulted in a curated corpus of 298 articles, which forms the foundation of our survey.

2.3. *Survey structure*

The creation of AI-driven smart spaces requires the integration of sensing technologies and the establishment of communication networks within these spaces. Consequently, we structured this survey to align with the common requirements of IoT architecture. The core components, including the technologies and methodologies that underpin AI-driven smart spaces, are outlined in Fig. 2. Each section presents these components, addressing the relevant technologies and methodologies in detail. To show the significance, we conclude each section with a dedicated subsection summarizing the associated challenges and limitations of the existing approaches.

In Section 3, we discuss sensor technologies, communication technologies and protocols, sensor network management, and the processes of collecting and processing data, which form the foundational components of smart spaces. To emphasize the role of AI in enhancing AI-driven smart spaces, we first present the capabilities of conventional ML and DL models in enabling various applications. Following this, we introduce transformer networks and highlight LLMs as emerging learning techniques with transformative potential for developing AI-driven smart spaces. We address AI applications in two sections. In Section 4, we provide a survey of smart space applications powered by ML, DL, and transformer networks. Then, Section 5 introduces the LLM families and models, emphasizing their effectiveness in advancing AI-based applications within smart spaces. Finally, Section 6 concludes the paper.

Table 2
Systematic literature review methodology and selection framework.

Search Configuration	
Study Period	January 2010 – September 2025 (Initial: Jan 2010 – May 2024; Updates: Jun – Sep 2025)
Databases	Google Scholar, IEEE Xplore, ACM Digital Library, ScienceDirect
Search Keywords	AI/ML Methods: artificial intelligence, machine learning, deep learning, transformers, GPT, LLMs, RNNs, LSTM, NLP Application Domains: smart space, smart home, smart building, smart house, smart office, indoor environment Technologies: environmental sensors (air quality, temperature, etc.), cameras, wireless sensing, wearables, smart meters, PIR, thermal arrays, wireless communication (BLE, RFID, Wi-Fi, etc.), communication protocols (REST, QUIC, etc.)
Query Examples	“LLMs in smart homes”, “Transformer networks in indoor environments”, “AI energy optimization in buildings”
Criteria	
Inclusion	• Peer-reviewed journal articles and conference papers • Focus on AI/ML applications in indoor/smart environments • Empirical results, theoretical frameworks, or system architectures
Exclusion	• Non-peer-reviewed sources as primary references, though preprints and white papers were included as supplementary material for latest fields where peer-reviewed literature is not yet established (e.g., latest LLM applications) • Outdoor-only IoT applications without indoor relevance • Duplicate entries across databases • Papers not specifically addressing AI applications within indoor environments
Selection Process	
Workflow	Database query → Duplicate removal → Title/abstract screening → Full-text assessment → Categorization → Subsequent search → Creation of sections with most cited articles → Citation tracking
Total Articles	298 conference and journal articles
Categories	(1) Sensing & Communication Technologies; (2) Machine Learning & Deep Learning Applications; (3) Human Activity & Occupancy Recognition; (4) Energy & Environmental Optimization; (5) Edge, Cloud & Federated AI Systems; (6) Security, Privacy & Data Management; (7) Large Language Models & Generative AI; (8) Surveys & Review Works

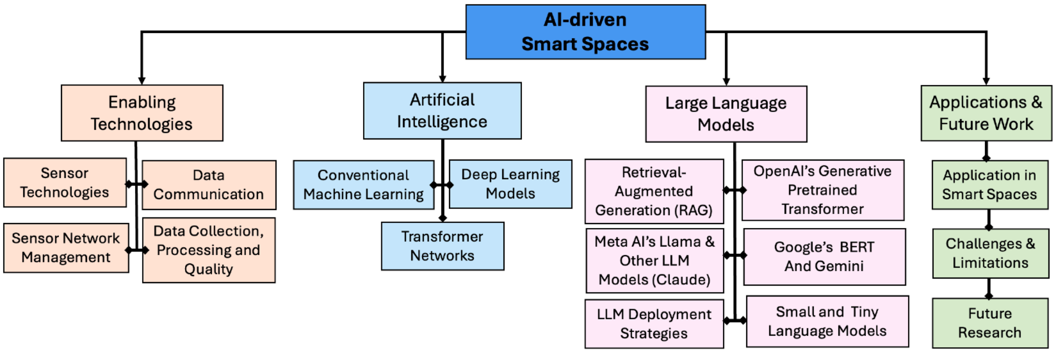
AI-driven Smart Spaces
Enabling Technologies
Sensor Technologies
Data Communication
Sensor Network Management
Data Collection, Processing and Quality
Artificial Intelligence
Conventional Machine Learning
Deep Learning Models
Transformer Networks
Large Language Models
Retrieval-Augmented Generation (RAG)
OpenAI's Generative Pretrained Transformer
Meta AI's Llama & Other LLM Models (Claude)
Google's BERT And Gemini
LLM Deployment Strategies
Small and Tiny Language Models
Applications & Future Work
Application in Smart Spaces
Challenges & Limitations
Future Research

Fig. 2. The scope and structure of the survey.

3. Enabling technologies

3.1. Sensor technologies

Sensors are essential components in the development of smart spaces. They can collect various types of data from environments, providing insights into the state of the surroundings. Examples of such data include light, noise, temperature, humidity, pressure, motion, images, and other environmental factors [34]. Generally, sensors are equipped with wired or wireless communication protocols such as Bluetooth, Wi-Fi, or LoRaWAN, enabling them to collect and transmit data to servers for storage and processing. In this subsection, we present sensor technologies that are suitable for creating smart spaces. Representative sensors and the services they enable are summarized in Table 3.

3.1.1. Environmental sensors

Environmental sensors are essential technologies used indoors to monitor and understand various conditions and aspects of the environment. Below are examples of sensors commonly utilized in indoor settings.

Air quality sensors: These sensors are vital technologies that monitor and provide essential data on various air pollutants. Examples of these pollutants include Carbon Monoxide (CO), Carbon Dioxide (CO_2), Nitrogen Dioxide (NO_2), Ozone (O_3), Formaldehyde

Table 3
Common sensor types in smart spaces.

Sensor Type	Typical Measurands	Applications	Limitations
Environmental (gas, particulate matter, Temp, RH, light, noise)	CO, CO_2, NO_2, O_3, VOCs, $PM_{2.5}$, temperature, relative humidity, illuminance (lux), acoustic levels (dB)	IAQ monitoring, demand-controlled ventilation, thermal comfort, adaptive lighting	Sensor drift and aging require in-situ or opportunistic calibration [79]
Infrared motion (PIR, 8 × 8 thermal arrays)	Human presence, motion patterns	Lighting and HVAC control, privacy-preserving occupancy detection	Inability to detect stationary occupants, coarse spatial resolution [38,55]
Cameras (RGB, thermal)	Visual imagery, thermal signatures, facial features, gesture patterns	Security surveillance, fine-grained activity and gesture recognition	High bandwidth and power requirements, privacy and GDPR compliance [60]
Vibration sensing (geophones, accelerometers)	RF channel perturbations, signal strength variations, ground vibrations, structural movements	Device-free occupant tracking, structural health monitoring	Sensitive to ambient building noise, requires professional installation [69]
Wireless sensing (Wi-Fi CSI, BLE RSSI, RFID)	RF channel perturbations	Device-free localization, gesture and vital-sign detection	Multi path sensitivity, site-specific calibration required [70]
Smart meters and power monitoring	Energy (kWh), gas, water	Non-intrusive load monitoring, appliance disaggregation, leak detection, demand response	Single point of failure, limited granularity [66]
Wearables and mobile devices	IMU, GPS, BLE beacons, physiological signals	Personalized comfort control, health analytics, indoor positioning	Requires user compliance, battery-life constraints [76,77]

(HCHO), Volatile Organic Compounds (VOC), and particulate matter like $PM_{2.5}$ [35–38]. In smart spaces, these sensors enable real-time air quality monitoring, ensuring a clean, healthy, and comfortable indoor atmosphere. They can also trigger early warnings when air quality deviates from safe levels. The data gathered by these sensors can be used to intelligently control ventilation systems, adjusting airflow rates when pollution levels exceed safety thresholds [39].

Temperature and humidity sensors: Temperature sensors play a crucial role in maintaining optimal thermal conditions. In indoor environments, they are used to assess thermal comfort and provide insights into personalized comfort experiences [40–43]. By integrating temperature sensors with physiological data, occupant-responsive models [44] can be developed to capture personalized thermal preferences and predict individual comfort levels more effectively. These temperature measurements enable HVAC systems to precisely adjust heating and cooling, ensuring occupant comfort while enhancing thermal satisfaction, e.g., through the integration of temperature and humidity sensing with digital twins to offer personalized thermal comforts [45]. Furthermore, automating HVAC systems based on real-time data from temperature sensors reduces excessive energy consumption, optimizing energy use within the space. Alongside temperature sensing, humidity sensors are also important sensing technology that enables maintaining proper indoor conditions by regulating relative humidity. These sensors help prevent mold growth, material damage, and discomfort caused by excess moisture or dryness, thereby creating a healthier and more comfortable indoor environment [36]. Consequently, the use of temperature and humidity sensors is essential in smart spaces to enhance personalized occupant experiences, improve well-being, and improve energy efficiency.

Light and noise sensors: The use of ambient light sensors (also known as photometric sensors) and noise sensors is crucial in indoor environments, as they help monitor and regulate light and noise levels, ensuring both visual and auditory comfort [46,47]. Indoor illumination is closely tied to the activities of the occupants [48]. By sensing indoor light conditions, lighting systems can intelligently adjust and dim light levels to meet the visual comfort needs of individuals, optimizing energy use while maintaining comfort [49,50]. Similarly, noise sensors are vital for detecting sound levels, which can be used to monitor occupancy or identify specific events occurring within indoor spaces [51]. These technologies contribute to creating responsive and adaptive environments that enhance the overall well-being of occupants.

3.1.2. Infrared sensors

Infrared sensors are among the most commonly used sensors in buildings, employed to detect the presence of people and monitor space occupancy, identify occupant activities and space utilization, and support energy management systems [19].

Passive infrared (PIR) sensor: This sensing technology, commonly known as a PIR (Passive Infrared) sensor, is widely used indoors to detect motion by passively sensing infrared (heat) radiation emitted by objects within its field of view [38]. PIR sensors operate in three modes: idle mode, where no motion is detected (i.e., no infrared radiation) and the sensor remains inactive; detection mode, which occurs when motion is detected; and triggering mode, when the sensor activates an event, such as triggering an alarm system to detect unauthorized entry. In smart environments, PIR sensors provide valuable data that can be used to enhance various comfort, health, and safety applications [52]. For instance, they enable motion-activated lighting, where sensors automatically turn lights on or off, and automated appliance control, where sensors manage electrical systems such as fans, smart curtains, heaters, or HVAC systems. By adjusting or turning off electrical systems when no motion is detected, these PIR-enabled applications contribute to significant energy savings [53].

Ultralow-resolution infrared (IR) sensors: This sensor technology is a type of thermal sensor that captures infrared (IR) radiation at a very low resolution. These IR sensors detect radiation emitted by objects and generate information using a limited number of pixels, restricting their ability to provide high accuracy or detailed outputs, such as precise temperature mapping [54]. However, these limitations make IR sensors cost-effective, energy-efficient, and privacy-preserving, making them well-suited for use in various smart environment applications. These applications typically involve basic infrared detection tasks, such as motion and presence detection, occupancy monitoring, proximity sensing, and simple gesture recognition [55–57].

3.1.3. *Cameras*

Cameras are commonly used technologies in smart spaces. They offer a rich source of information from both users and systems, enabling the development of various applications such as security, access control, and activity recognition.

Surveillance cameras: These technologies also known as video surveillance systems, have become ubiquitous and widely accessible. Depending on the specific application and the desired image quality, commercially available cameras vary in frame rates, resolutions, and fields of view [58]. In smart environments, these cameras can be integrated with other sensor technologies and actuators, facilitating advanced applications such as facial recognition, gesture detection, augmented reality, and occupancy monitoring [59]. However, despite their capabilities, surveillance cameras pose significant challenges related to privacy [60], as well as data storage and processing [61]. Hence, careful consideration is required when designing and deploying these systems in smart environments to mitigate potential risks.

Thermal cameras: These technologies utilize a passive sensing approach, capturing infrared radiation (heat) emitted by objects [62]. While thermal cameras typically have lower resolution compared to surveillance cameras, they are well-suited for various privacy-preserving applications, such as people detection, counting, and tracking [63]. In smart environments, thermal cameras can be employed to monitor energy consumption of systems such as TVs, computers, lighting, and HVAC systems, helping optimize their operation. Additionally, these technologies can be used indoors to monitor occupancy in a privacy-conscious manner [64].

3.1.4. *Other sensing technologies*

A wide variety of sensor technologies are available today, each tailored to measure specific variables and serve numerous applications [65]. In this subsection, we present a few common sensor technologies that can be used to create smart spaces.

Smart meters: These devices are designed to measure, record, and monitor the consumption of resources such as electricity, gas, and water. In buildings, smart meters can be used to detect leaks or abnormal resource usage [66,67]. Therefore, in smart environments, they can play an important role in collecting consumption data, enabling a better understanding of space utilization. This data can be used to identify the energy consumed by specific systems (e.g., lighting, TVs, HVAC), and to develop models and forecasting methods for scheduling and optimizing resource utilization [68].

Geophone sensors: These highly sensitive devices can detect ground vibrations with high precision, making them widely used in seismic monitoring. Their accuracy and reliability also make them valuable for applications such as monitoring structural vibrations and movements in buildings [69]. In smart environments, geophone sensors can be utilized for several purposes: (i) detecting occupancy to improve energy management, security, and space optimization, (ii) tracking occupant positions, providing insights into movement patterns, (iii) identifying footsteps and estimating direction, and enabling personalized tracking and behavior analysis, and (iv) analyzing space utilization to enhance interior design by adjusting layouts or seating arrangements.

Wireless sensing: This sensing technique leverages the channel state information of wireless signals from communication technologies such as IEEE 802.11 (Wi-Fi) or cellular systems. By analyzing fluctuations in the radio waves transmitted between devices, this technique enables detecting the activities and events. Wireless sensing enables estimation of persons' presence, activity, location, motion, and their gestures [70–72]. In smart environments, common devices such as wireless access points, TVs, mobile phones, laptops, and other smart devices-equipped with technologies like Wi-Fi, Bluetooth, or cellular networks-can facilitate the implementation of wireless sensing. Smart environments can harness wireless sensing for a wide range of applications, including security, health and wellness monitoring, emergency response, and intelligent interactions [73,74].

Actuators: These technologies are designed to convert energy-whether electrical, hydraulic, or pneumatic-into mechanical motion. Actuators function as devices or processes that translate control signals into physical movement. They are employed in a wide range of applications, from industrial automation to consumer electronics, enabling the operation of various systems [75]. In smart environments, actuators play a crucial role in optimizing energy efficiency. For instance, they control the operation of HVAC and lighting systems, regulating airflow or adjusting lighting levels based on environmental conditions. They can also be integrated with smart curtains to modulate light exposure, contributing to energy conservation and enhanced user comfort.

Smartphones and wearables: Equipped with advanced sensor technologies, smartphones and wearables generate valuable data that can seamlessly integrates into smart spaces. Smartphones, for instance, provide GPS, accelerometer, and Bluetooth data, enabling precise tracking, indoor positioning, and an understanding of the services users require within the space [76]. Wearables, with their physiological and biomedical sensors, allow continuous monitoring of users' health, safety, and activities, such as tracking health metrics within smart spaces [77]. Additionally, these devices can support privacy-aware solutions, ensuring that smart spaces respect user confidentiality while leveraging their capabilities [78].

3.2. *Data communications*

In this subsection, we present the most common wireless communication technologies and protocols that can be used to establish smart spaces.

3.2.1. *Communication technologies*

The diverse range of technologies used to establish smart spaces-such as sensors, actuators, smart phones, smart wearables, and smart TVs-each have unique communication requirements that vary based on factors like deployment location, lifespan constraints, and data traffic patterns. Therefore, selecting the right communication technology for a smart space is a critical design decision, as it can greatly influence the performance, power consumption, and overall functionality of the connected devices, ultimately affecting the efficiency of the smart environment. This section provides an overview of the wireless technologies suitable for smart spaces. Key characteristics of communication technologies that can be used in smart spaces are compared in Table 3.2.1.

Bluetooth Low Energy (BLE) is a short-range wireless communication technology specifically designed for Internet of Things (IoT) applications, serving as an extension of classic Bluetooth. BLE is cost-effective to implement and supports a typical range of up to 100 m. Operating in the 2.4 GHz ISM band, BLE can achieve data rates of up to 2 Mbps, especially with BLE v5.0, which offers high-throughput communication. Optimized for low power consumption, BLE is ideal for battery-operated devices [80]. It is a highly versatile and widely adopted technology, commonly used in IoT-enabled devices such as consumer electronics. BLE provides reliable solutions for indoor localization, proximity detection, and energy-efficient operation [81], making it a an efficient technology to be used for creation of smart environments.

Wi-Fi, also known as WLAN, includes various standards designed for short- and extended-range scenarios, each suited to specific use cases [82]. Among these, 802.11ac (Wi-Fi 5), 802.11ax (Wi-Fi 6), and 802.11ah (HaLoW) are particularly relevant for smart indoor environments. The 802.11ac standard offers a theoretical throughput of up to 7 Gbps, with coverage of approximately 35 m indoors and 100 m outdoors, making it suitable for high-bandwidth applications [83,84]. 802.11ax provides an indoor range of about 30 m and an outdoor range of 120 m, with throughput up to 10 Gbps, and can support large-scale sensor deployments and big data applications [85]. Operating in sub-1 GHz bands, 802.11ah delivers extended coverage up to 1 km, but at lower data rates (150 Kbps to 78 Mbps), making it ideal for IoT applications requiring broad coverage in outdoor environments [86,87]. Given its support for high-data-rate applications, Wi-Fi typically consumes significant power, particularly in continuous or always-on scenarios. Therefore, in smart environments, it is best suited for applications with high bandwidth and sufficient power availability, such as video streaming [88].

Long Range Wide Area Network (LoRaWAN), commonly referred to as LoRa, is a low-cost, low-power technology designed to support a maximum data rate of 50 kbps. This technology targets a battery life of up to 10 years, making it particularly well-suited for resource-constrained IoT devices. LoRaWAN operates on a star-of-stars topology and does not facilitate device-to-device communication. The technology employs an adaptive data rate technique, optimizing transmission parameters to enhance both energy efficiency and network longevity [89]. With minimal energy consumption, LoRaWAN can maintain reliable connectivity over long distances while providing high accuracy. This capability enables effective remote monitoring and control of IoT devices. These features, combined with its low-power requirements, make LoRaWAN exceptionally suitable for smart space applications [90].

Low Power Wireless Personal Area Networks (6LoWPAN) is a low-power, short-range communication technology specifically designed for transmitting IPv6 packets. It offers a maximum data transfer rate of 250 kbps and supports a range of up to 200 m. 6LoWPAN facilitates the deployment of mesh network topologies and implements the IEEE 802.15.4 standard for device connectivity [91]. In smart environments, the features of 6LoWPAN make it an ideal technology for interconnecting various smart devices, including sensors, actuators, and home appliances [92].

ZigBee is a low-power, cost-effective, and low-data-rate wireless technology based on the IEEE 802.15.4 standard, designed for short-range communication. Operating in a mesh network configuration, ZigBee supports up to 65,536 devices and offers a maximum data transfer rate of 250 kbps [93]. This technology allows devices in close proximity to communicate directly without the need for routers or access points. With a battery life of up to 20 years, ZigBee devices typically have a communication range of 10 to 100 m, depending on environmental factors and device configurations. Due to its energy efficiency and flexible network scalability, ZigBee is widely used in smart environments, particularly in home automation systems, such as smart sensors, lighting, door locks, and home controllers [94].

Z-Wave is a wireless technology designed for low-power, low-data-rate applications, with a range of up to 100 m in open air and up to 20 m indoors. It supports multiple data rates, including 100 kbps for maximum throughput, and 9.6 kbps or 40 kbps for more energy-efficient communication. Z-Wave allows bidirectional communication, facilitating a robust mesh network topology. In theory, a Z-Wave network can accommodate up to 232 devices, though for optimal performance and reliability, a practical limit of around 50 connections is recommended [95]. Due to its reliability and ease of integration, Z-Wave is widely adopted in smart home automation and security systems [96].

Thread is a versatile wireless technology built on the IEEE 802.15.4 standard, enabling IPv6-based wireless mesh networking. It is optimized for low-power operation, making it ideal for battery-powered devices. Thread supports a range of up to 30 m in open air and up to 15 m indoors [97]. Initially, this technology is intended for smart home environments, enabling various applications such as appliances, access control, climate control, smart thermostats, lighting, as well as safety and security systems, including cameras and alarms [98].

Radio Frequency Identification (RFID) is a communication technology used for identifying and tracking objects. The ISO/IEC 18,000 series establishes standards for RFID, including information management protocols. For example, ISO/IEC 18000-6B allows simultaneous reading of up to 10 tags with a large data storage capacity, while ISO/IEC 18000-6C enables the reading of hundreds of tags at once, although with lower individual data storage per tag. RFID is a cost-effective solution, offering a typical range from a few centimeters to several meters. This makes it ideal for efficient tracking systems, especially in indoor positioning and localization scenarios. RFID-based indoor localization is particularly useful in smart environments for monitoring daily activities and tracking objects or visitors [99].

Table 4
Comparison of common IoT communication technologies in smart spaces.

Technology	Range	Data Rate	Power	Topology	Security	Applications	Limitations
RFID [99]						Asset tracking	Very short range
NFC [100,101]	0.02–3 m	≤1.7 Mb/s	Passive	Reader–tag	EPC Gen2	payments	privacy
Thread [97]	15–30 m	250 kb/s	Ultra-low	Mesh (IPv6)	128-bit AES	Smart-home automation	Needs IPv6 infrastructure
BLE 5.0 [80]	≤100 m	≤2 Mb/s	Ultra-low	Star/mesh	AES-CCM	Wearables proximity beacons	Connection overhead 2.4 GHz interference
ZigBee 3.0 [94]	10–100 m	250 kb/s	Ultra-low	Mesh	128-bit AES	Smart lighting HVAC	Co-channel Wi-Fi interference
Z-Wave [95]	30–100 m	100 kb/s	Low	Mesh	S2 security	Retrofit home-automation	Proprietary sub-1 GHz only
Wi-Fi 6 [82]	45–70 m	≤10 Gb/s	High	Star (BSS)	WPA3	Video streaming edge analytics	High power AP required
6LoWPAN [91]	10–100 m	250 kb/s	Ultra-low	Mesh	IPsec/DTLS	IPv6 sensor networks	Header-compression overhead
LoRaWAN [89]	0.5–15 km	0.3–50 kb/s	Ultra-low	Star-of-stars	AES-128	Environmental monitoring	Duty-cycle limits
NB-IoT [102]						Remote monitoring	Carrier dependency
EC-GSM-IoT [103]						Asset tracking	Subscription cost
LTE-M [104]	>1 km	0.35–1 Mb/s	Low	Cellular	3GPP security	HVAC	No voice

Near Field Communication (NFC) is a subset of RFID technology, primarily used for close-range applications. It enables communication within a range of up to 2 cm and supports data transfer rates from 46 kbps to 1.7 Mbps [100]. NFC can enhance indoor navigation systems by allowing users to update their location by simply tapping NFC tags placed throughout an environment, effectively addressing the limitations of GPS in indoor settings [101]. The short operational range of NFC, combined with its ease of use and security features, makes it an ideal choice for secure identification systems. For instance, it can be used for access control in smart environments, enabling users to authenticate themselves when entering restricted areas.

Extended Coverage - GSM - Internet of Things (EC-GSM-IoT) is a cellular wireless technology developed specifically to support IoT applications that require low data rates and extended coverage. By utilizing existing 2G infrastructure, EC-GSM-IoT enables efficient and reliable IoT connectivity over a broader GSM coverage area, making it ideal for smart environments where numerous devices require secure, energy-efficient connections [102]. This low-power technology is designed to achieve up to 10 years of battery life using just a 5 Wh battery. While EC-GSM-IoT does not support voice communication, it offers flexible data rates ranging from 350 bps to 70 kbps, adjusting based on coverage and transmission requirements. These characteristics make it an excellent choice for automating and controlling key home systems, such as HVAC, lighting, and surveillance cameras [103].

Long Term Evolution for Machines (LTE-M) is an open-standard cellular IoT technology designed to support a wide range of low to mid-range IoT applications. LTE-M offers several advantages, including low device costs, extended coverage, and long battery life, while maintaining the capacity to support a high density of devices per cell. With data rates reaching up to 384 kbps, LTE-M is well-suited for applications requiring higher data throughput and low latency [102]. A key benefit of LTE-M is its support for mobility, voice, and SMS, which makes it highly adaptable across various smart home applications. In smart environments, LTE-M can power numerous systems such as home security, closed-circuit television (CCTV), and fire alarm networks, where reliable connectivity and real-time responsiveness are essential [104].

Narrowband Internet of Things (NB-IoT) is a low-power cellular IoT standard designed to provide a 20 dB gain over GPRS, making it particularly effective for indoor applications. With a target battery life of up to 10 years and a data rate of up to 100 kbps, NB-IoT can support up to 50,000 IoT devices per cell [102]. These features make it an ideal choice for smart environments. In such settings, NB-IoT can be used for various applications, including smart lighting systems, indoor environment monitoring, home automation and control, as well as enabling efficient, accurate, and low-latency indoor positioning [105].

3.2.2. Communication protocols

The communication protocols are a set of rules that govern the transmission and exchange of data across sensor networks. These protocols typically follow either the publish-subscribe model, the request-response model, or capable to implement both of the models.

Publish-Subscribe and Request-Response Methods: In the publish-subscribe model, communication between data producers (publishers) and consumers (subscribers) is decoupled. Publishers send data to a central broker, which then distributes it to subscribers based on predefined topics or criteria. This model is particularly effective for systems requiring many-to-many communication, as it reduces the overhead of direct communication between devices. In contrast, the request-response model involves direct communication between devices. A client sends a request to a server, and the server responds with the required information. This model is well-suited for scenarios that demand immediate, one-to-one interactions [106]. Both models offer distinct advantages. In the publish-subscribe model, publishers and subscribers do not need to be aware of each other's presence. The complexity of connections is reduced by eliminating point-to-point communication, allowing a single subscriber to receive data from multiple publishers, and a single publisher to broadcast data to multiple subscribers. Additionally, publishers and subscribers do not need to be active simultaneously, as the broker can queue and store messages, forwarding them when clients become active. On the other hand, the request-response model has the advantage in providing reliable, real-time interactions, particularly when the server can manage high traffic volumes and meet client demands effectively [107]. When designing smart environments, the choice between the communication models depends on the infrastructure, type of sensors, scale of sensor deployment, and the computational capabilities of the system.

Message Queuing Telemetry Transport (MQTT) is a lightweight messaging protocol supported by the Advancement of Structured Information Standards (OASIS), a nonprofit standardization consortium. It operates on a publish-subscribe model and, due to its minimal header design, is highly efficient for devices with limited resources, such as low-power sensors. This makes it an ideal choice for sensor networks with low bandwidth and high latency. MQTT relies on the TCP transport protocol, ensuring reliable communication. Since MQTT is designed to be lightweight, data is exchanged in plain text by default, therefore to secure data transmission, it uses TLS/SSL encryption [108]. In MQTT, two key entities are involved in communication: clients and brokers. Clients can either publish messages or subscribe to receive messages. The broker, acting as the central component, receives messages from publishing clients, filters them by topic, and forwards them to the appropriate subscribers. The broker can also store messages for new subscribers or discard them if no subscribers exist for a given topic [109].

Constrained Application Protocol (CoAP) is a lightweight messaging protocol supported by the Internet Engineering Task Force (IETF), an open standards organization [110]. CoAP follows a request-response model and runs over the UDP transport protocol, offering minimal overhead, making it ideal for communication in constrained environments. The protocol architecture consists of two layers: a messaging layer and a request-response layer. The messaging layer manages message transmission and ensures reliability, while the request-response layer handles the actual exchange of data [111]. When a CoAP client sends a confirmable (CON) request to a CoAP server, the server generates an acknowledgment (ACK) in response. The response data is embedded within the ACK message, a process known as a piggybacked response. To secure communication, CoAP uses Datagram Transport Layer Security (DTLS), running over UDP. Additionally, CoAP includes a feature that allows clients to continuously receive updates on a requested resource from the server, enhancing the traditional request-response model. The IETF further extends CoAP's capabilities to support a publish-subscribe model, enabling even more flexible communication approach [112].

Advanced Message Queuing Protocol (AMQP) is an open-standard communication protocol supported by the OASIS standard. AMQP is designed to ensure interoperability across a broad spectrum of devices, enabling efficient data exchange between platforms developed in different programming languages, particularly suited for heterogeneous systems. The protocol operates using both the publish-subscribe and request-response models and runs on top of the TCP transport protocol. AMQP incorporates a store-and-forward mechanism to enhance message reliability. AMQP also implements TLS/SSL protocols to ensure the confidentiality and integrity of data, meeting the security requirements [113]. Moreover, while less commonly used compared to MQTT and CoAP, AMQP offers robust messaging capabilities at the price of higher resource demand [109]. AMQP is typically employed for communication between software applications, servers, and hosts in smart home environments [114].

Data Distribution Service (DDS) is a protocol based on the decentralized publish-subscribe model that is standardized by the Object Management Group (OMG) [115]. It operates without the need for a broker component, enabling asynchronous data exchange between publishers and subscribers via a data bus. By eliminating the need for message brokers, DDS enables devices to join or leave the network at any time. DDS is particularly well-suited for real-time systems, utilizing both UDP and TCP transport protocols to provide extensive QoS capabilities. For security, DDS employs TLS, DTLS, and DDS Security (DDS-Sec) protocols, ensuring robust data protection. To address the high overhead of TLS and DTLS protocols in constrained environments, the OMG DDS security specification introduces a comprehensive security model, with a service plugin interface (SPI) architecture, allowing security implementations in IoT systems [116].

HTTP is a widely used application-layer protocol governed by IETF standards. Over the years, different versions of HTTP have been developed, each designed to enhance the protocol's performance. HTTP follows a request-response model and relies on TCP as its transport protocol, ensuring reliable data transmission. It commonly uses JSON for data exchange, and for security purposes, it employs TLS encryption, which is often referred to as HTTPS [117]. Despite its widespread use, HTTP faces limitations in IoT environments due to its complexity, large header sizes, and high power consumption. Additionally, HTTP is often associated with REST, a set of guidelines that prescribes using specific HTTP methods (e.g., GET, POST, PUT, DELETE). RESTful HTTP (or REST over HTTP) supports CRUD (Create, Retrieve, Update, Delete) operations, enabling the development of web applications and simplifying state management for IoT nodes [118]. In environments where communication and power efficiency are less critical, such as edge and cloud computing systems, RESTful HTTP remains a viable option [106].

Extensible Messaging and Presence Protocol (XMPP) is a communication protocol defined by the IETF which is designed for near-real-time data exchange between network entities. It supports various communication models, including request-response, publish-subscribe, end-to-end, and multicast interactions. XMPP is built on Extensible Markup Language (XML) and operates within a distributed client-server architecture. Data is exchanged asynchronously using XML fragments, enabling functionalities like messaging, presence updates, and request-response interactions [119]. For security, XMPP employs Transport Layer Security (TLS) for authentication and uses the Simple Authentication and Security Layer (SASL) for encryption. Its scalable design allows for custom protocol extensions, making it adaptable to a wide range of applications, including instant messaging, IoT deployments, and multi-user conferencing.

QUIC is a recently developed protocol which is standardized by the IETF, designed to improve upon traditional TCP-based protocols including HTTP/1.1 and HTTP/2, particularly the latency, connection establishment, and transport layer security. Unlike TCP, QUIC operates over UDP and follows a request-response architecture similar to HTTP [120]. QUIC enhances web performance by integrating key features such as multiplexing, encryption, congestion control, and connection migration directly into the transport layer. These optimizations make it especially effective in mobile and unreliable network environments. Additionally, QUIC improves energy efficiency by enabling faster data transmission, reducing the time devices spend actively communicating, which in turn lowers power consumption. Its advanced features make QUIC an ideal choice for latency-sensitive applications, real-time services, and IoT deployments, where reliable and efficient data transmission is essential [121].

3.3. Sensor network management

Creating smart environments requires the effective management of IoT networks to efficiently control their operations. This involves meeting several key requirements, where we present in this subsection.

3.3.1. Sensor network energy management

Efficient energy management is crucial for the design and deployment of sensors in smart spaces. Since many IoT devices are battery-powered and often installed in hard-to-reach locations, such as ceilings, it is vital to implement strategies that minimize energy consumption while maintaining long-term operational efficiency [122]. Sensors equipped with high-power wireless communication modules are particularly significant energy consumers, which highlights the need for optimization techniques to extend their operational lifespan. For battery-powered sensors that require regular recharging or replacement, frequent maintenance is often impractical and costly [123]. Moreover, scaling the number of sensors increases the complexity of the network, demanding more efficient coordination and dynamic energy optimization techniques. For instance, sensor signals are prone to interference, as many devices operate on the same frequency bands. As the network scales, additional energy may be required for signal re-transmission due to lost or corrupted data. To ensure stable and optimal energy consumption and performance as the network expands, current state-of-the-art research introduces scalable procedures to address these challenges [124].

Energy Harvesting: Sensor devices that frequently transmit small amounts of data can utilize energy harvesting techniques-by converting ambient energy sources such as solar, vibration, or thermal energy into electrical power-to reduce dependence on batteries or grid power. For instance, research has proposed battery-free, light-based sensor systems powered by lighting, enabling energy harvesting even in indoor environments [125]. Similarly, other studies have introduced dual-use energy harvesting systems that leverage both light and radio frequency technologies to power data transmission from environmental sensors [126,127]. Implementing energy harvesting methodologies in smart spaces is therefore a sustainable solution, significantly reducing energy requirements and operational costs within the sensor network. By minimizing the need for battery replacements or recharging, energy harvesting enhances the long-term efficiency and scalability of IoT deployments, supporting resilient and cost-effective smart environments.

Sensor Energy Management Techniques: Several research studies have proposed various energy management strategies to optimize sensor power usage. One commonly employed technique is *duty cycling*, which conserves energy by periodically switching sensors on and off during periods of inactivity. Many sensor systems utilize heavy duty cycling, with over 90% of their operational time spent in low-power sleep modes, interrupted by brief bursts of activity [123]. The ML can further enhance duty cycling, improving energy efficiency by up to 30% without compromising sensing accuracy [128]. Another approach is *power gating*, which selectively powers down individual components or modules of the sensors when they are not in use. This method is particularly effective for sensors with multiple functional modules, as these often do not need to operate simultaneously, leading to significant energy savings [123]. *Sleep scheduling* is another energy-saving technique that puts sensors into low-power modes during inactivity and activates them only when necessary. This is especially useful in periodic data collection scenarios where continuous sensor operation is not required [129]. *Dynamic Voltage Scaling (DVS)* is a technique that dynamically adjusts the voltage supplied to sensors based on real-time demands, significantly extending sensor node lifespan. DVS also adapts to changes in workload and network conditions, ensuring sensors only consume the required amount of power. A task-driven feedback DVS algorithm can further optimize energy usage by dynamically scaling voltage while using feedback loops to correct errors [130]. In addition to these methods, a variety of other strategies can extend sensor battery life. These include selecting low-energy components, optimizing hardware design, and employing techniques such as “race to sleep” and “think before you talk” [129].

Energy-Aware Communication Protocols: In sensor networks, energy-efficient communication protocols are crucial for minimizing power consumption while ensuring reliable data transmission. The protocols, Low-Power Wireless Personal Area Networks (LoWPANs) and Low-Power Short-Area Networks (LPSANs)-such as RFID, NFC, Zigbee, Bluetooth, Z-Wave, and 6LoWPAN-are specifically designed for short-range communication with minimal energy consumption. These energy-aware protocols achieve efficiency by optimizing transmission frequency, allowing sensors to preserve data integrity while minimizing power usage [122]. Additionally, Low-Power Wide-Area Networks (LPWANs) are designed for long-range communication with broad coverage, maintaining low power consumption. Examples of LPWAN technologies include LoRa, Sigfox, NB-IoT, LTE-M, and EC-GSM-IoT, which are ideal for applications that require extensive network reach with minimal energy use [131]. In large-scale sensor deployments, LPWANs enable extended sensing and monitoring over vast areas while ensuring energy efficiency.

Other Approaches: Several additional techniques in the literature focus on optimizing energy consumption in IoT sensor networks. Among the most effective are *ML-based energy management* approaches, which enhance sensor energy efficiency by dynamically optimizing sampling rates. These techniques can predict the energy consumption of various sensors in smart environments and adjust their sampling rates in real time, based on factors like occupancy patterns. By leveraging historical data and ML algorithms, these systems enable more intelligent decision-making. For example, in a smart space, during periods of low occupancy, ML models can instruct sensors to reduce their sampling rates, conserving energy without compromising sensing accuracy [132]. Another approach is *sensor fusion*, which integrates data from multiple sensors to improve accuracy while reducing the energy consumption of individual devices. By combining information from various types of sensors-such as temperature, humidity, and motion-sensor fusion allows for more efficient data collection and processing, ultimately conserving power of sensor devices [133].

3.3.2. Sensors operations and maintenance

Effective sensor operations and management are crucial for the seamless functioning of smart spaces. To ensure accurate and reliable data collection, the following key aspects must be addressed:

Maintenance and Real-Time Monitoring: In smart spaces, regular maintenance-such as cleaning or replacing sensors-is crucial to ensure sensor longevity and prevent performance degradation. Real-time monitoring and condition-based maintenance of sensors and networking infrastructure enable the identification of inefficiencies and the adjustment of sensor operations only when necessary, ensuring optimal performance while minimizing the costs of physical inspections [134]. Real-time monitoring in smart spaces often involves occupancy detection sensors, such as Passive Infrared (PIR) motion sensors, which help differentiate between high- and low-occupancy zones [135]. The combined use of PIR motion detectors and CO_2 sensors can further estimate occupancy levels and analyze patterns within a space, providing valuable insights for improved environmental control [12]. Additionally, digital twins play a key role in sensor deployment and management by creating virtual replicas of physical spaces. These digital models allow for the simulation of real-world conditions, offering insights through descriptive, diagnostic, predictive, and prescriptive analytics, enhancing decision-making and sensor maintenance [2].

Sensor Fault Detection: Sensors are susceptible to issues such as drift, bias, or complete failure, and early detection of these faults is critical for maintaining reliable sensor performance [136]. Research has proposed various distributed fault detection methods to identify both permanent and intermittent sensor faults. For example, establishing trust relationships between sensors in smart spaces can effectively detect sensor faults while using minimal computational resources [137]. To ensure the overall reliability of a sensor network, it is essential to regularly assess the performance of sensors to verify that they produce accurate measurements. This can be achieved through frequent testing, statistical comparisons of measured variables, and correlation analyses between sensor data and that of an accurate portable sensor within the smart space. Sensors that exhibit drift and produce anomalous data patterns are referred to as “anomaly sensors”. These anomalies can be identified through outlier analysis and by calibrating the sensor’s measurements against accurate reference sensor. Sensor failure-where a sensor stops transmitting data-is a common issue, often caused by faults in the power unit, sensing components, or communication modules. Continuous real-time monitoring and regular inspections enables preventing sensor failures [138].

3.4. Data collection, processing, and quality

3.4.1. Data collection and processing

The creation of smart environments relies on deploying heterogeneous sensor technologies, generating data with differing volumes, velocities, varieties, veracity, value, and vulnerabilities [139]. For instance, an environmental sensor designed to measure temperature and air quality information may transmit data every few seconds. In contrast, cameras-whether infrared or surveillance-may operate continuously, streaming high volumes of image data at a rapid pace [139]. As a result, deploying diverse sensor technologies in smart spaces leads to the generation of complex datasets with varying time-stamped data. In addition, the sensor heterogeneity and varying standards often result in data being produced in different formats. For example, environmental sensors might output semi-structured data such as *.json* or *.xml*, while cameras and microphones generate unstructured data like video and audio streams, respectively [140]. These variations increase data complexity, necessitating flexible data collection and handling strategies tailored to smart environments. A solution however would be using specialized time-series databases designed for these types of sensor deployments.

In smart environments, the collected sensor data can be analyzed in real-time to provide services to occupants or used for advanced analytics to enhance the environment’s functionality. The continuous data streams from multiple sensors generate large volumes of data, often referred to as “big data” which demand significant computational resources-especially when applying AI and machine learning (AI/ML) models to deliver AI-based services. To offer AI/ML-based services, several data processing functions must be executed. These include data preprocessing, which involves noise removal, data conversion, normalization, and labeling; feature extraction; classification; and the execution of specific ML models [141]. Each of these functions requires fast and accurate data processing to ensure the delivery of real-time services to users.

Edge computing and cloud computing both offer effective solutions to meet the data storage and processing demands of smart environments [142]. The choice between these computing approaches depends on the specific design of the smart space, requiring careful consideration of the respective advantages and disadvantages of each. Naturally, cloud computing offers substantial computational power but raises concerns regarding latency, privacy, and bandwidth. In contrast, edge computing places the computing facilities closer to where sensor devices are located, which helps improve privacy and bandwidth concerns and reduce latency, enabling real-time services [143].

In smart environments, edge computing can serve as the main computing facility, connecting sensors, actuators to collect and process data, enhancing the performance and efficiency of these environments, by delivering timely and accurate data analytics and decision-making at the network’s edge. Edge computing can also leverage machine learning and deep learning techniques, facilitate communication and coordination among devices, including user-connected devices such as smartphones, and adapt to dynamic and complex settings [144]. For instance, it can enable the implementation of models that learn from user behaviors and the deployment of LLM-based models to provide customized and personalized services for the users of the environment.

Cloud computing, on the other hand, can function as a processing platform either by directly receiving data from sensors or by serving as a combined processing facility alongside edge computing, following the edge-cloud computing continuum architecture [145,146]. While cloud computing may introduce some latency and privacy concerns, it provides substantial storage capacity and facilitates the execution of AI/ML models that require large datasets, thereby enhancing the quality and functionality of services within smart environments [147].

3.4.2. Data quality

The quality of sensor data is fundamental to the accurate interpretation of environmental events in smart environments. Data quality, both in terms of its qualitative and quantitative aspects, is assessed based on key criteria such as accuracy, completeness, validity, consistency, uniqueness, and timeliness. In smart environments, the integrity of sensor data is crucial, as it directly influences decision-making processes. However, several factors, such as sensor degradation, measurement drift, network failures, and battery depletion (in battery-powered devices), can lead to deteriorating data quality, resulting in incomplete, inaccurate, inconsistent, noisy, outdated, or redundant data streams [148].

Sensor technologies that are suitable to be deployed at indoor environments, typically designed to be low-cost, often have limitations in terms of sensing accuracy. Fortunately, recent advancements in artificial intelligence (AI) and machine learning (ML) have enabled the development of effective calibration models that significantly enhance the accuracy of these sensors [149]. For example, state-of-the-art ML-based methods, including regression models, ensemble techniques, and neural networks, have been applied to improve the measurement accuracy of key environmental variables such as particulate matter ($PM_{2.5}$) and carbon dioxide (CO_2) levels-both critical for indoor air quality monitoring [38,150,151]. Furthermore, specialized ML models, such as Long Short-Term Memory (LSTM) networks, have been employed to process sequential and temporal data (e.g., air quality measurements) [151]. Convolutional Neural Networks (CNNs) are commonly used for image and video analysis, while Recurrent Neural Networks (RNNs) and Autoencoders (AEs) are leveraged for tasks such as data compression and reconstruction. Variational Autoencoders (VAEs) offer additional benefits for data modeling, and Generative Adversarial Networks (GANs) assist with data augmentation and synthesis. In addition, Deep Reinforcement Learning (DRL) is utilized for control and decision-making within the environment [152].

In addition to sensor calibration, the quality of sensor data can be affected by errors, outliers, anomalies, and missing data points. AI and ML-based techniques provide robust solutions for addressing these data quality challenges. For instance, data cleaning algorithms can detect and correct anomalies by removing or repairing invalid, missing, duplicate, or inconsistent data points [148]. Data fusion techniques can also enhance the comprehensiveness and accuracy of environmental data by combining inputs from multiple sensors, thus enriching datasets that might otherwise be incomplete, noisy, or uncertain. Additionally, the application of data provenance-techniques that track and record the origin, history, and ownership of data-offers a way to further improve data quality. By attaching metadata that describes the source, context, reliability, and trustworthiness of the data, provenance systems provide additional layers of validation.

In nutshell, maintaining and improving sensor data quality in smart environments is essential for reliable decision-making. This can be achieved by implementing AI/ML-based data processing pipelines that enhance data quality at both the edge and cloud, where sensor data is collected and processed. By integrating advanced ML techniques, these pipelines can improve the accuracy and reliability of sensor data, which in turn optimizes the overall performance of smart environments.

3.5. Challenges and limitations

3.5.1. Sensor technologies challenges

Commercially available sensor technologies are affordable and provide a wide range of environmental data from the locations where they are deployed. Their low cost and availability make it feasible to deploy them in large numbers, enabling the creation of smart environments. However, despite these advantages, they come with the following limitations:

Sensing capabilities: Low-cost sensors typically have a limited number of sensor units. Some sensors are designed to measure a specific variable, such as temperature, while others can measure multiple variables on a single sensor board, such as temperature, relative humidity, and CO_2. In reality, there are many variables that can be measured to fully understand the state of an indoor environment, but not all sensors are equipped to capture all of these variables. For example, some sensors can measure temperature, relative humidity, and CO_2, while others measure variables such as temperature, noise, light, and $PM_{2.5}$. To overcome this limitation, virtual sensors powered by AI and machine learning can be developed. These virtual sensors estimate unmeasured variables by using available data as a proxy [153]. For instance, $PM_{2.5}$ and temperature measurements can be used to estimate CO_2 and black carbon concentrations, providing a more comprehensive understanding of indoor environmental conditions [154].

Communication capabilities: Low-cost sensors, when not using cables for communication, typically employ at least one type of wireless communication technology. Each technology has its own set of characteristics, such as connectivity robustness, bandwidth, coverage, and energy consumption. For instance, while one low-cost sensor might offer Bluetooth connectivity, another might provide both Bluetooth and Wi-Fi options. Bluetooth is known to have connectivity robustness issues [155], while Wi-Fi might be more demanding in terms of energy consumption [156], which is a crucial consideration for battery-powered sensors. Therefore, when planning the connectivity and communication strategy for sensors, it is necessary to consider the availability and features of each communication technology.

Sensing accuracy: Low-cost sensors often experience reduced sensing accuracy over time. Although these sensors are initially calibrated in a controlled laboratory environment, their measurement precision tends to degrade, leading to measurement drifts. However, in smart environments, AI and machine learning-based calibration techniques can be employed to re-calibrate these sensors more frequently [79], helping to ensure accurate sensing over time.

Energy consumption: Most low-cost sensors are battery-powered and have limited energy resources. To create smart environments and improve the management of sensors deployed in these settings, it is essential to plan an effective data transmission strategy.

While some sensors generate small amounts of data that require less power to transmit, the frequency of data transmission can be adjusted based on the importance of the measured variables. This helps to extend the battery life. For sensor devices that capture images and require more bandwidth and transmission power, mechanisms can be implemented to selectively transmit a limited number of images, ensuring the extension of battery life [157].

Security and privacy: In smart environments, the use of cameras and other sensor devices that can potentially reveal individuals' identities is often known as privacy-intrusive, raising significant security and privacy concerns. To address these issues and protect individuals' privacy, a range of privacy-preserving techniques can be employed, including encryption, anonymization, access control, and differential privacy [158].

3.5.2. Data communications challenges

The communication technologies and protocols that enable data transmission in smart environments come with their own set of challenges and limitations, some of which we will briefly address here.

Communication Technologies: Most short-range wireless communication technologies are designed for low-power, low data-rate, and constrained IoT devices, which introduces significant challenges when creating smart environments. Below, we highlight a few examples. BLE faces difficulties in maintaining stable connections in dynamic channel environments, leading to performance degradation. When a connection is lost, nodes must restart the recovery process, which involves exchanging multiple control packets and results in increased power consumption [159]. Zigbee operates in the same frequency bands as WLAN, which increases the likelihood of interference in wireless environments. This interference leads to packet transmission delays and reduced reliability, necessitating adaptive cooperation mechanisms between Zigbee and WLAN to mitigate these issues [160].

Wi-Fi, despite its widespread availability and robust connectivity, is power-hungry in IoT environments. Designed primarily to optimize bandwidth, range, and throughput, Wi-Fi is not well-suited for power-constrained applications, particularly those relying on battery-powered devices [161]. LoRa (Long Range), although specifically designed for low-power, low-data-rate communication, operates in unlicensed frequency bands. This can lead to an increase in packet collisions due to shared spectrum use [162]. Additionally, as the number of connected devices in a LoRa network grows, contention and interference increase, ultimately degrading network performance [163].

Short-range communication technologies also face security threats and vulnerabilities [164]. For instance, Z-Wave has been shown to be susceptible to cyberattacks. By exploiting crafted data, attackers can potentially disable key features of Z-Wave-enabled IoT devices. One significant vulnerability arises during wireless firmware updates, where attackers can remotely take control of Z-Wave device operations [165]. Other example technologies that are often vulnerable to security threats include RFID and NFC. RFID tags can carry sensitive personal information, making them a target for exploitation by adversaries [166]. Similarly, while NFC operates within a few centimeters and requires physical device contact-such as tapping devices together to exchange personal information-this approach, though secure in proximity, can be less convenient compared to alternatives like app-based controls.

Communication Protocols: Although highly effective in enabling data exchange within IoT networks, various communication protocols present distinct challenges and limitations. For instance, HTTP faces difficulties in IoT environments due to its complexity, the large header size of TCP, and its high power consumption [118]. XMPP, on the other hand, relies on XML, leading to larger message sizes, which are inefficient for bandwidth-constrained networks. Additionally, XMPP's dependence on a persistent TCP connection, combined with its lack of efficient binary encoding, makes it unsuitable for lossy, low-power wireless networks commonly found in IoT environments. In terms of security, while XMPP supports basic SASL and TLS protocols, it lacks advanced native features like end-to-end encryption [167].

Other protocols, such as MQTT and CoAP, also have security vulnerabilities. Although MQTT is an efficient protocol for IoT, it is susceptible to various cyberattacks because it was originally designed for trusted IoT networks [168]. Basic security measures such as username/password authentication and SSL/TLS encryption are often inadequate, necessitating additional protective measures [169]. Similarly, CoAP, which operates over UDP, lacks a handshake mechanism, making it vulnerable to IP spoofing attacks. This vulnerability can escalate into more severe threats, such as Distributed Denial-of-Service (DDoS) and amplification attacks, within IoT environments [170].

Additionally, the emerging QUIC transport layer protocol comes with its own limitations. One of its key features, 0-RTT (zero-round-trip time), reduces latency but introduces the risk of replay attacks. Since QUIC operates over UDP, it is also susceptible to amplification attacks, where attackers exploit the absence of initial handshake verification to overwhelm servers. Furthermore, adoption challenges arise due to compatibility issues with legacy network infrastructure, as some firewalls and middleboxes may not be optimized for handling UDP traffic efficiently. While QUIC encrypts most packet contents for security, certain metadata remains unencrypted to facilitate routing, raising concerns about potential privacy risks through traffic analysis [171].

3.5.3. Sensor network management challenges

Sensor Calibration: Low-cost sensors, which form the backbone of smart space sensing infrastructures, often experience degradation in sensing quality over time, even under normal operating conditions [172]. As a result, regular calibration is crucial to maintain sensor accuracy and ensure the reliability of collected data. Calibration involves establishing a precise correlation between the sensor's raw output and the actual measured value [150]. In practice, this process is typically conducted either in a laboratory setting or by comparing the measurements of sensors to highly reliable and precise reference sensors [173]. To ensure accurate and consistent sensor measurements, calibration is necessary before sensors are deployed in real-world environments [174]. In smart spaces, where sensors are deployed at large scales, regular calibration is required. However, removing and re-installing individual sensors for laboratory calibration is impractical. Therefore, in-situ calibration-where sensors are calibrated in their deployed locations-becomes essential [175]. To achieve this, there is a need to develop opportunistic calibration mechanisms, such as using a recently calibrated sensor to calibrate others within the smart space.

Data Security: Sensor networks are highly vulnerable to various cyber threats, including eavesdropping, data tampering, and denial of service (DoS) attacks. Ensuring robust security involves implementing encryption, authentication, and secure communication protocols to protect the integrity of both data and the network. Due to their limited computational capabilities, sensors are particularly susceptible to attacks such as node replication and eavesdropping. While symmetric-key cryptography can be employed to enhance security, it introduces additional challenges due to resource constraints [176]. In smart spaces, sensors often collect crucial and sensitive information about individuals, making cyber attacks a significant threat to the collected data. Without adequate security mechanisms, intruders can easily compromise sensors and networking infrastructure, such as Wi-Fi routers, to access this data [177]. Therefore, implementing an effective security strategy to counter cyber attacks and safeguard data collected by sensors is a critical challenge [178].

3.5.4. Data collection, processing, and quality challenges

Processing costs: The use of AI/ML techniques enable the learning of complex, nonlinear patterns from high-dimensional and unstructured data, allowing for effective sensor calibration, anomaly detection, and handling of missing data. However, the adoption of AI/ML in smart environments presents the following challenges. The first challenge relates to the data availability. The lack of large, labeled datasets specific to smart environments hinders the training of accurate models. Generally, data in IoT systems is often sparse, unlabeled, or incomplete, making it difficult to apply supervised learning methods without significant pre-processing efforts. The second challenge relates to the computation and memory costs. AI/ML models, especially deep learning techniques, require substantial computational resources to execute and large memory to store data. The third challenge refers to the model interpretability. Many AI/ML models operate as "black boxes" providing predictions without offering insights into their decision-making process within the smart environments. Enhancing the transparency of these models remains an ongoing challenge.

Privacy and security: In smart environments, IoT systems continuously collect and process large volumes of data, which may contain sensitive information about individuals and their surroundings. This raises significant privacy and security concerns, particularly with respect to safeguarding personal data and complying with regulatory frameworks such as the General Data Protection Regulation (GDPR) and the California Consumer Privacy Act (CCPA) [179]. Followings include some of the key privacy and security challenges in sensor data processing.

The data generated by IoT sensors can potentially reveal individuals’ identities and behaviors, making it critical to implement privacy-preserving approaches. For example, sensors deployed in smart environment inadvertently capture personal activities, leading to a increased risk of privacy violations. In addition, ensuring compliance with stringent data protection regulations is a challenge in smart environments. These regulations impose strict guidelines on how personal data should be collected, stored, and shared, necessitating advanced privacy-preserving techniques.

To meet the security requirements, the state-of-the-art encryption techniques, such as the Advanced Encryption Standard (AES) and Transport Layer Security (TLS), are commonly used to secure data during transmission. Additionally, homomorphic encryption schemes, such as the Brakerski-Gentry-Vaikuntanathan (BGV) cryptosystem [180], enable secure data aggregation without exposing the underlying information. However, while these methods enhance security, they also introduce computational overhead, making it challenging to implement them for the smart environments.

Moreover, robust data governance frameworks are essential to prevent unauthorized access and mitigate cyber threats. Inadequate access controls and improper data management strategies can result in data breaches, demanding the development of effective privacy-preserving algorithms.

4. Artificial intelligence

The artificial intelligence (AI) and machine learning (ML) are terms often used interchangeably. However, while AI broadly refers to the ability of computers to perform tasks in real-world contexts, ML specifically involves the development of algorithms that enable systems to analyze data, recognize patterns, and make informed decisions. Creation of smart spaces relies heavily on the ML tools to provide AI-driven functionalities. This section explores the AI and ML techniques, needed to enable the capabilities of smart spaces.

Fig. 3 illustrates the evolution of AI methodologies, tracing their development from traditional ML techniques to advanced transformer networks and large language models (LLMs). As shown in the figure, LLMs are built on three main architectures: decoder-only, encoder-only, and encoder-decoder. Decoder-only models, such as GPT and Gemini, are primarily designed for text generation and dialogue. Encoder-only models, including BERT and MiniLM, are well suited for classification and information extraction tasks. Encoder-decoder architectures combine both structures to perform sequence-to-sequence tasks, such as translation and summarization. The figure also indicates the companies that developed each model (referred in parentheses) and provides few examples of AI applications within smart spaces.

Fig. 3 serves as a guide for the subsequent discussion, which is structured into two distinct sections. The first section covers conventional machine learning, deep learning, and transformer networks, while the second section focuses specifically on advancements in LLMs. This division allows us to present the information clearly and concisely, avoiding lengthy sections. As shown in the figure, AI facilitates a wide range of applications within smart spaces. Among many, the example applications include occupancy detection, activity recognition, elderly care, health monitoring, and home automation. In the following two sections, we introduce those AI methodologies (presented in the figure) while highlighting their applications in smart spaces. In addition, we discuss the limitations associated with each AI methodology in relation to these specific applications. To provide a comprehensive overview, we also present

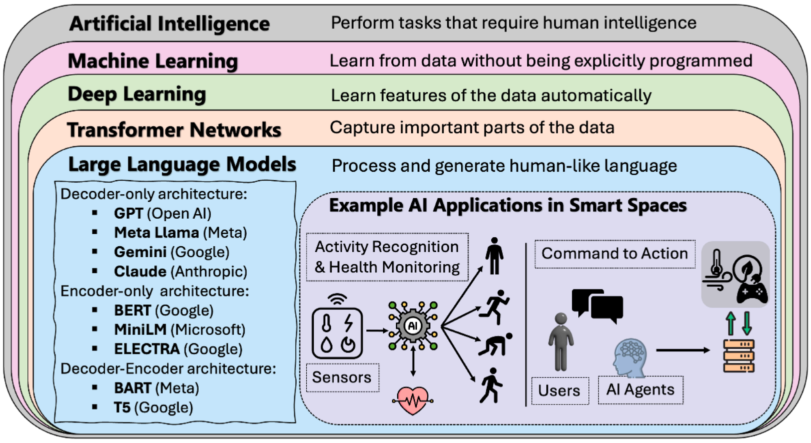
Artificial Intelligence
Perform tasks that require human intelligence
Machine Learning
Learn from data without being explicitly programmed
Deep Learning
Learn features of the data automatically
Transformer Networks
Capture important parts of the data
Large Language Models
Process and generate human-like language
Decoder-only architecture:
GPT (Open AI)
Meta Llama (Meta)
Gemini (Google)
Claude (Anthropic)
Encoder-only architecture:
BERT (Google)
MiniLM (Microsoft)
ELECTRA (Google)
Decoder-Encoder architecture:
BART (Meta)
T5 (Google)
Example AI Applications in Smart Spaces
Activity Recognition & Health Monitoring
AI
Sensors
Command to Action
Users
AI Agents

Fig. 3. Advancing AI techniques from conventional ML methods to LLMs.

a summary of the AI methodologies, their applications, and their limitations in smart spaces in Table 5.7, which is located at the end of the second section to encapsulate the information from both sections.

4.1. Conventional machine learning models

Conventional machine learning (ML) models use relatively simple structures of interconnected neurons to analyze and interpret data. These models are trained on historical datasets to develop predictive capabilities, enabling them to generalize and make accurate inferences on new data. Common examples of these models include perceptrons, multilayer perceptrons (MLPs), feed-forward artificial neural networks, random forests, logistic regression, naive Bayes, support vector machines (SVM), k-nearest neighbors (KNN), and extreme gradient boosting (XGBoost) [181]. A major strength of conventional ML models lies in their versatility for tasks such as classification, detection, regression, clustering, and pattern recognition, making them effective methods for human activity recognition in smart spaces.

For instance, the multilayer perceptron (MLP) model is a common neural network architecture, composed of an input layer, one or more hidden layers, and an output layer. Each neuron connects to every neuron in the subsequent layer, enabling the MLP to learn complex patterns through non-linear activation functions. By iteratively updating weights via backpropagation, the network can approximate any continuous function, achieving high classification accuracy. One study demonstrated that an MLP model with 256 neurons achieved 98% activity classification accuracy when trained on wearable device data from sensors on the ankle and wrist [182]. MLPs, when combined with k-pattern algorithms, can further enhance the accuracy of activity classification and prediction [183], enabling user-centric solutions in smart spaces.

Another effective model is the random forest, an ensemble learning algorithm used for both classification and regression. A random forest comprises multiple decision trees, each trained on a randomly selected subset of the data, and aggregates their outputs to produce a robust prediction. Random forest models have shown high accuracy in applications such as fall detection and activity recognition using human skeleton features [184]. Random forest shows an accuracy of 91% for people counting and 98.13% for occupancy detection [185].

Spiking Neural Networks (SNNs) process information through discrete spikes, allowing asynchronous processing and high activation sparsity, which is beneficial for real-time applications in smart environments [186,187]. SNNs can be used to cluster user activities based on sensor readings, enhancing the adaptability and responsiveness of smart environments [188,189].

Additionally, other conventional ML models such as XGBoost, and SVMs have proven to be powerful for anomaly detection [190], which is essential for improving life quality through continuous monitoring in smart spaces. By identifying irregularities in daily activities, these models can detect early signs of health issues, particularly for the elderly and individuals with disabilities.

Moreover, conventional ML models show efficiency in significantly reducing unnecessary electricity consumption within smart spaces. For example, ANNs with backpropagation and variable learning rate (VLR) can automate energy control by toggling electrical components on and off based on occupancy or use patterns [191].

4.2. Deep learning models

Deep learning (DL) architectures enhance traditional machine learning (ML) approaches by incorporating complex models designed to handle large datasets. DL architectures consist of multiple hidden layers, allowing for automated feature extraction from raw data and enabling the processing of diverse unstructured data types, including images, text, speech, and video. Such capabilities make DL models highly effective for tasks such as image recognition, activity recognition, natural language processing, and autonomous systems, making them powerful tools for real-time data analysis in smart spaces. DL architectures are generally classified into supervised learning models, such as Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs), and unsupervised learning models, such as Autoencoders (AEs). Based on these foundational architectures, numerous DL models have been developed, some of which we address in this subsection.

CNN as a fundamental deep learning architecture, consists of three primary components including the convolutional layer, the pooling layer, and the fully connected layer. Convolutional layers extract features from input data, identifying essential patterns such as edges, textures, and shapes. Pooling layers downsample the feature maps, effectively reducing spatial dimensions and computational complexity. Finally, fully connected layers integrate the features learned across the network to enable predictions or classifications. In smart spaces, CNNs can enable privacy-preserving, real-time applications such as occupancy monitoring, activity recognition, and people counting [55]. For instance, YOLO (You Only Look Once) is a widely used CNN architecture designed for real-time object detection by combining object localization and classification tasks within a single convolutional network. The lightweight design of YOLO makes it an ideal choice for deployment in smart spaces [192].

RNNs are another important DL architecture, specifically designed for processing sequential data. RNNs incorporate self-looping connections that enable them to retain a hidden state, capturing information from previous inputs. This unique feature makes RNNs well-suited for tasks where context and order are crucial, such as time series analysis, natural language processing, activity recognition, and speech recognition. In smart spaces, RNNs can offer a range of applications; for example, they can improve indoor localization by generating movement path data based on Wi-Fi fingerprints [193], and enable human activity recognition [194].

Long Short-Term Memory (LSTM) networks are a specialized variant of RNNs designed to capture long-range dependencies in sequential data. The LSTM architecture enables these models to retain context over extended sequences, making them ideal for analyzing time-sensitive sensor data. By effectively capturing temporal dependencies, LSTM models can provide more accurate predictions. In smart spaces, LSTMs are particularly useful for applications such as human activity recognition [195] and energy consumption forecasting, supporting optimized energy management strategies [196].

The Gated Recurrent Unit (GRU) network is an optimized variant of the RNN architecture, offering similar capabilities to LSTM networks for processing sequential data. With a simplified design and fewer parameters, GRUs improve computational efficiency while effectively capturing temporal dependencies. This makes them well-suited for tasks such as time series analysis, natural language processing, and speech recognition, where modeling long-term dependencies is essential. As a result, GRUs are ideal for developing applications such as sensor-based activity recognition [197], air quality prediction [198], and electricity consumption forecasting [199].

Autoencoders (AEs) are DL models for unsupervised learning, capable of capturing essential features of data while filtering out irrelevant information. These features makes AEs particularly effective for functions such as anomaly detection. Hence, in smart spaces, AEs can be employed to identify anomalies in time-series sensor data. For example, AEs can detect unusual behaviors such as irregular power consumption, control system failures, sensor malfunctions [200], and anomalies at indoor air quality [201].

4.3. Transformer networks

Transformer networks, a deep learning architecture innovation proposed by Google researchers, are distinguished by their attention mechanism, enabling a model to dynamically weigh the significance of elements within a sequence to capture complex contextual relationships [202]. Unlike earlier models such as RNNs and LSTMs, which process data sequentially, transformers operate in parallel, enhancing computational efficiency and scalability. This parallelism allows transformer networks to outperform prior deep learning models in both accuracy and speed, making them particularly effective for tasks involving time-series data, natural language processing, and other applications requiring the modeling of complex and context-dependent relationships in data. Additionally, transformers demonstrate strong performance even in supervised tasks with limited labeled data, a valuable feature for various smart space applications [203].

Compared to traditional deep learning models like convolutional neural networks (CNNs), LSTMs, and support vector machines (SVMs), transformer networks exhibit superior performance in user behavior and activity recognition in indoor environments [203, 204]. This advantage arises from the self-attention mechanism of transformers, which enables the modeling of spatial-temporal dependencies, enhancing classification and activity detection accuracy [205,206]. Furthermore, transformers excel in forecasting, allowing for predictive applications such as estimating users' next activities and their durations, a critical capability for enabling context-aware services in smart spaces that offer personalized and anticipatory interactions [207,208]. Beyond activity detection, transformers facilitate indoor localization and mobility pattern recognition, enabling user classification, location prediction, and schedule estimation [209]. Such insights can foster healthier environments, particularly benefiting elderly individuals, patients, children, and people with special needs by monitoring mobility and behavior in smart spaces [16,210].

Due to their long-range sequence modeling capabilities, transformers are increasingly employed in indoor activity detection using WiFi signal strength and coverage, for example, in modeling indoor human localization and mobility through WiFi connectivity patterns [211,212]. Transformers utilize self-attention to process WiFi channel state information (CSI), achieving high indoor localization accuracy [213]. This method also offers a privacy-preserving alternative to camera-based systems by reducing reliance on visual data for pose estimation [214] Notably, transformers facilitate privacy-preserving techniques that mitigate privacy concerns-a critical requirement for smart spaces [215,216].

Other studies leverage wearable-based sensing, incorporating sensors like inertial measurement units (IMUs) [205], or utilize smartphone-based approaches, applying Vision Transformers (ViT) for image processing in human localization tasks [217]. RFID-based methods, which measure received signal strength indicator (RSSI) values, are also used for activities like fall detection for elderly individuals indoors [218].

Transformer networks contribute to energy efficiency and operational optimization in smart spaces. For instance, transformers can optimize HVAC systems and forecast energy demands by using models such as the Temporal Fusion Transformer, which captures long-range dependencies in sequential energy data through attention mechanisms [219]. These models can deliver accurate energy load predictions based solely on smart meter readings [220]. Furthermore, multi-task learning frameworks based on transformers, utilizing environmental variables, support short-term multi-load energy forecasting [221,222]. Integrating transformer-based, context-aware systems with deep reinforcement learning (DRL) enables multi-zone HVAC control optimization, balancing energy consumption with thermal comfort, thereby enhancing quality of life in smart environments [223].

Transformers also excel in transfer learning, enabling knowledge transfer across tasks to improve model performance in dynamic environments [224]. In smart spaces, transformer networks can adapt to user preferences and behaviors, supporting user-centric interactions and assistance [225]. They even facilitate immersive metaverse interactions, where users can control and engage with their surroundings through avatars in virtual environments [214]

To further enhance transformer performance, recent studies have explored integrating the self-attention mechanism into other neural network types, improving the model's ability to focus on relevant features in the input data. Hybrid models that combine transformers with RNNs or CNNs utilize attention mechanisms to filter out irrelevant information, boosting prediction accuracy for human activity recognition in smart spaces [226]. These hybrid approaches provide a flexible and powerful framework for solving complex problems, significantly enhancing smart space applications.

4.4. Challenges and limitations

AI Implementation Complexity: In smart spaces, the effectiveness of DL and ML models relies heavily on data collected from various sensors, followed by a training phase to perform specific tasks. Conventional ML models typically require supervised learning, necessitating extensive labeled datasets and data preprocessing-such as outlier detection, data imputation, and feature selection-to ensure data integrity and model accuracy. Due to noise and inconsistent data formats, preprocessing may also involve filtering and scaling techniques [227]. Selecting and configuring models for smart spaces is challenging, as DL/ML models like CNNs and RNNs need tuning to accommodate memory and processing limitations of the computing devices. In addition, these models must be optimized for low latency and energy efficiency, often requiring specialized edge computing configurations and hardware expertise [228]. To address these constraints, strategies such as model compression and efficient memory management are essential, as limited resources can impact the performance and scalability of conventional ML models, complicating deployment in resource-constrained computing environments [229]. Architectures such as on-device computing, edge server-based solutions, and hybrid systems combining edge and cloud resources need careful selection based on specific use cases and device limitations. Achieving optimal performance on resource-limited hardware therefore demands substantial expertise in both ML and embedded systems [230]. As a result, deploying and managing DL/ML models in resource-constrained environments is challenging and often requires specialized expertise due to the complexity of model training and optimization.

Data Privacy and Security: Deploying AI in smart spaces raises significant privacy challenges, as these models often process sensitive user data, including identity, health information, personal habits, and preferences. Since smart spaces inherently involve local data collection, on-device processing and localized data handling are essential for mitigating privacy risks. Existing studies proposes methods such as Federated Learning (FL)-which enables model training across decentralized devices, such as IoT nodes, without sharing raw data-to address the privacy challenges of IoT deployments. By training AI/ML models directly on local devices without transmitting raw data, FL helps preserve user privacy. However, FL remains vulnerable to "poisoning" attacks, in which training data or model parameters are deliberately manipulated to mislead the model [231]. Despite this vulnerability, FL is particularly valuable in smart indoor spaces where sensitive location data is involved, as it reduces data transmission risks while maintaining model performance [232,233]. On the security side, many IoT devices within smart environments are resource-constrained, which limits the use of complex security protocols without impacting performance. This limitation highlights the need for lightweight, efficient security solutions that balance strong protection with operational efficiency [234]. In summary, preserving data privacy and developing robust security approaches are central challenges for smart spaces, where safeguarding users' information is essential.

5. Large language models

Large language models (LLMs) – that are deep learning models trained on big datasets – are gaining the research momentum. With billions of parameters, LLMs can capture complex linguistic patterns and structures, and leverage the self-attention mechanisms and scalability of transformer architectures to process and understand languages across extensive sequences. This makes LLMs invaluable tools for various natural language processing (NLP) tasks such as text classification, sentiment analysis, and machine translation [235, 236].

Integrating LLMs within smart spaces enables a wide range of applications that enhance operational performance, comfort, and promote healthier living environments. For instance, in smart spaces, LLMs can promote sustainable behaviors and improve user interactions by dynamically adjusting environmental settings and minimizing energy waste through predictive analytics [237]. LLMs can facilitate user-centric interactions, offering assistance with functions such as pose estimation and activity recognition [238]. Interestingly, in multilingual environments, LLMs support communication by providing user-friendly notifications in various languages [239]. LLMs can also interact directly with the physical world through sensors deployed in the spaces, inferring user activities from sensor data to improve contextual awareness [240–242]. Additionally, LLMs can be utilized to develop recommendation systems that provide personalized suggestions-optimizing for instance workspace setups, promoting energy-saving habits, and guiding users to suitable areas based on behavioral patterns-all of which contribute to sustainable practices [15,243].

5.1. Retrieval-Augmented generation (RAG)

Applying LLMs to real-world applications often requires integrating Retrieval-Augmented Generation (RAG) techniques, which equip generative AI models with information retrieval capabilities. RAG enables LLMs to access external data sources in real-time without additional training. This process involves two main stages including a "Retrieval Stage", where relevant documents or snippets are fetched from an external corpus based on the input query, and a "Generation Stage", where the LLM synthesizes information from the retrieved content to generate contextually accurate responses. This dual-stage process allows RAG-equipped LLMs to provide precise, real-time, and contextually enriched responses.

Presently, most studies in the literature apply RAG techniques to tasks like open-domain question answering and document retrieval. For example, the study in [244] explores RAG-equipped models "RepLLaMA and RankLLaMA" for retrieval tasks using a multi-stage pipeline. Whereas, RepLLaMA identifies relevant documents from a large corpus, and RankLLaMA re-ranks these based on relevance scores to prioritize the most contextually appropriate documents. Another study [236] enhances LLM reasoning in IoT applications by incorporating RAG to retrieve domain-specific IoT knowledge, thereby facilitating complex reasoning tasks.

In the literature, research on LLMs and RAG techniques applied to smart spaces is currently limited. In smart spaces, RAG-equipped LLMs can significantly improve tasks such as human activity recognition, anomaly detection, WiFi-based human sensing, and indoor localization by up to 65% [236]. Therefore, using LLMs with RAG capabilities can enhance the functionality of smart spaces, leading to more interactive and user-centered experiences. In the followings, we address well-known LLM models that can be utilized to enhance the capabilities of smart spaces.

5.2. OpenAI's generative pre-trained transformer (GPT)

The GPT is a state-of-the-art language model and a pioneering framework in generative artificial intelligence. Developed by OpenAI, GPT leverages advanced deep learning techniques. Since the release of its first version, GPT-1, in June 2018, the model has steadily evolved, with each new version demonstrating enhanced abilities in capturing language patterns and generating coherent, contextually relevant text. These capabilities provide opportunities to use GPTs to enhance the functionalities of smart spaces.

The GPT-1 model was trained using unsupervised learning with a 12-layer transformer architecture, where each layer included self-attention and feed-forward networks. It employed 117 million parameters and was trained on a dataset of over 7000 unpublished books [245].

The second evolution, GPT-2, was released with significantly increased capacity, consisting of 1.5 billion parameters and trained on a 40GB dataset of internet text. The GPT-2 architecture scaled up to 48 transformer layers, enabling it to handle tasks such as language translation, summarization, and question-answering without task-specific fine-tuning. In smart environments, GPT-2 can be applied to sensor event sequence prediction and real-time activity recognition of space users. This GPT-2-based approach predicts future sensor events in an autoregressive manner, outperforming traditional LSTM-based methods and providing anticipatory support in smart spaces [246].

GPT-3 was later introduced with 175 billion parameters and trained on a dataset of approximately 570 GB of internet text. With 96 transformer layers, GPT-3 was designed to support few-shot and zero-shot learning capabilities. Building on this foundation, the GPT-3.5 version was subsequently released, offering enhanced performance through optimized training techniques, including Reinforcement Learning from Human Feedback (RLHF) [247]. Existing literature demonstrates how GPT-3 enhances smart spaces by enabling context-aware and adaptable responses, surpassing the limitations of traditional rule-based systems. For example, GPT-3 can translate open-ended commands, such as "get ready for a party", into actionable device controls such as configuring lights and playing music. This is achieved through prompt engineering, where GPT-3 converts natural language commands into structured JSON outputs, which are then processed to control smart space devices [248]. In addition, similar to earlier versions, GPT-3.5 can be leveraged for zero-shot activity identification in smart spaces by utilizing sensor-based activity monitoring. For instance, the study in [249] applies GPT-3.5 to interpret environmental sensor data, transforming raw sensor inputs into descriptive text that enables GPT-3.5 to classify activities based on its pre-trained knowledge. Another study [250] introduces "Follow-Me AI", a GPT-3.5-powered system integrated with centralized AI agents that gathers both user preferences and environmental data to enhance user experiences in smart spaces. Follow-Me AI is designed to enable real-time adjustments to temperature, lighting, and occupancy, aligning with user preferences while also optimizing energy efficiency and maintaining data privacy.

Released in 2023, GPT-4 further enhanced GPT's reasoning and generalization abilities through training on a diverse dataset that included internet text, books, and specialized sources, allowing it to handle complex, domain-specific tasks more effectively [251]. GPT-4 introduced multimodal input capabilities, processing both text and images. Its input is divided into smaller units, called tokens, which are analyzed by Transformer layers. With a context length of up to 32,000 tokens, GPT-4 can predict subsequent tokens based on learned patterns to generate human-like responses [252]. Building on GPT-4, two new versions were introduced: GPT-4 Turbo, optimized for faster response times and lower computational demands, and GPT-4o, an omni model capable of processing and generating multiple media types, including text, audio, images, and video. GPT-4o is particularly notable for its rapid response to audio inputs, with an average latency comparable to human response time (around 320 ms). Key features of GPT-4o include web access, multimodal data processing, and enhanced code and mathematical capabilities, which together support improved reasoning abilities [253].

Integrating GPT-4 with smart spaces can offer advanced features and support context-aware, real-time decision-making in these environments, bridging the gap between digital reasoning and physical-world interactions. For example, the use of ChatGPT-4 have the potential to understand natural language commands and translate them into functional code, enabling home automation systems to activate code generated directly from user input [254]. This capability democratizes smart environments by allowing users to control devices and systems through natural language, enhancing accessibility and user experience. The study in [240] introduces “Penetrative AI”, which leverages GPT-4 to interpret smartphone sensor data and infer user activities, such as walking or remaining stationary, both indoors and outdoors. This system achieves over 90% accuracy in distinguishing between indoor and outdoor movements. Another study [255] presents a smart assistant named “Sasha”, which uses GPT-4 to respond to commands with creative, goal-oriented action plans. For example, when given the command “make it cozy”, Sasha interprets this abstract request into actionable JSON-based plans, adjusting smart devices such as lighting and temperature to create a comfortable environment. The research in [256], GPT-4 is applied to autonomously control HVAC systems in an office building. Integrated with a building simulation model, GPT-4 receives real-time indoor and outdoor CO_2 data and energy consumption metrics via a Python-based co-simulation interface. Based on this information, GPT-4 determines optimal control actions for variables for instance damper positions and chilled water temperature, aiming to minimize energy consumption while keeping indoor CO_2 levels below 1000 ppm. This example underscores the potential of large language models as effective decision-making agents, leveraging pre-trained general knowledge to perform complex, domain-specific tasks.

5.3. Meta AI’s large language model (LLama)

The LLaMA is a powerful family of autoregressive language models designed to provide efficient, high-quality language understanding for both general and specialized applications. Since the release of its first version in February 2023, Meta has developed several versions, including models with up to 65.2 billion parameters. LLaMA models are based on a transformer architecture and are pre-trained on a mixture of publicly available data sources, including Common Crawl, C4, GitHub, Wikipedia, books, and scientific articles. The dataset comprises around 1.4 trillion tokens, carefully curated to prioritize high-quality content and filtered to avoid duplicates and irrelevant information. Parameter sizes in the first version included 7 billion, 13 billion, 33 billion, and 65 billion [257].

The improved LLaMA 2 model, released later, introduced enhanced context sensitivity, dialogue capabilities, and alignment with user preferences. LLaMA 2 scales up to 69 billion parameters and is trained on 2 trillion tokens of publicly available data, allowing it to handle longer, more complex inputs. Key advancements include grouped-query attention (GQA) for larger models, which enhances inference scalability. LLaMA 2 also includes specialized fine-tuning for dialogue through supervised training and Reinforcement Learning with Human Feedback (RLHF). The dialogue-optimized variant, LLaMA 2-Chat, incorporates techniques such as red-teaming, safety tuning, and rejection sampling to improve alignment with human expectations for helpfulness and safety [258].

In 2024, Meta introduced LLaMA 3, representing a new generation of foundation models designed to support multilingual capabilities, advanced reasoning, tool use, and multimodal functionality across text, image, and speech. With up to 70.6 billion parameters, LLaMA 3 was trained on a 15.6 trillion token dataset. Compared to earlier versions, LLaMA 3 features an improved data curation pipeline for pre-training and post-training, enhancing its language understanding and complex reasoning skills. Subsequent updates, including LLaMA 3.1 and LLaMA 3.2, have added further features and fine-tuned performance. For example, LLaMA 3.2, which reaches a maximum of 405 billion parameters, shows strong capabilities in handling long-context applications and demonstrates reduced rates of hallucination.

Similar to other large language models, LLaMA models offer a range of opportunities when integrated with smart environments, particularly for handling complex tasks. For example, the study in [259] introduces “Harmony”, an intelligent home assistant system powered by the LLaMA 3-8B model, designed to maintain user privacy and operate locally without requiring an Internet connection. Harmony’s architecture consists of three components: a Message Handler, an Agent, and a Controller. The Message Handler processes sensor data and user commands, inferring user needs through both short-term and long-term memory functions. The Agent then formulates action plans based on these inferences, consulting memory for contextual relevance. Finally, the Controller translates the Agent’s plans into JSON-formatted commands to control devices, ensuring actions align with the smart home’s setup. Harmony demonstrates high accuracy (about 90%) in executing tasks, comparable to cloud-based solutions such as GPT-4, while significantly reducing hallucination rates. Harmony exemplifies how small-scale LLaMA models can enable effective, privacy-preserving applications for smart spaces without even the need for cloud resources.

5.4. Google’S bidirectional encoder representations from transformers (BERT) and gemini

The BERT is another significant example of a large language model (LLM) pre-trained on an extensive text dataset, including sources such as Wikipedia and BookCorpus. The larger version, BERT-Large, contains up to 340 million parameters, enabling it to understand and process language with advanced bidirectional context analysis, where words are analyzed in relation to both their left and right contexts within a sentence [260]. This is achieved through the transformer architecture, which uses self-attention layers to capture complex relationships between words. The study in [224] demonstrates the use of BERT to detect behavioral changes in elderly adults living in smart homes, improving monitoring and personalized care. By utilizing BERT’s sequence modeling capabilities, the study identifies anomalies that may indicate health or behavioral changes. To do this, BERT processes sequences of activities as tokenized events (e.g., “sleep” or “meal preparation”) and uses its bidirectional attention layers to model typical behavior patterns, enabling it to flag deviations that might indicate health deterioration. Another study in [261] employs BERT to enhance indoor positioning accuracy in smart spaces and defend against adversarial attacks. This research presents a crowdsourced indoor localization system that utilizes Bluetooth Low Energy (BLE) fingerprints. BERT’s self-attention mechanism captures complex patterns within BLE signal data, distinguishing authentic fingerprints from adversarial ones. The result is high localization accuracy and improved resilience against database and online attacks, supporting secure and reliable indoor localization.

The Gemini family, developed by Google, represents a powerful series of multimodal LLMs that excel in understanding and processing image, audio, video, and text data. Gemini models are available in various sizes-Ultra, Pro, and Nano-catering to a range of applications from complex reasoning tasks to memory-constrained, on-device use cases. Developers can leverage Gemini's capabilities through the Vertex AI Gemini API and Google AI Gemini API, enabling seamless integration into diverse applications [262,263]. In the literature, for instance, Gemini has been employed to optimize HVAC control in smart office environments. By integrating real-time environmental data, including temperature, illuminance, and occupant location from office sensors, Gemini's multimodal capabilities support enhanced energy efficiency and occupant comfort. The system processes these multimodal inputs to dynamically predict optimal HVAC setpoints, balancing energy savings with thermal comfort. Results from this experiment demonstrate that the Gemini-based system achieved up to a 47.92% reduction in power consumption and a 26.36% improvement in occupant comfort compared to traditional control methods. These findings underscore Gemini's effectiveness in managing complex, real-world environments, illustrating its potential for applications like energy-efficient management in smart spaces [264].

5.5. Other LLM models

In addition to the most widely recognized LLM models discussed earlier, an increasing number of new models are being introduced, each with specific features and advanced functionalities. Integrating these models within smart environments can offer advantages such as detecting sensor anomalies, enhancing user privacy, and reducing processing power demands.

One example is Claude, a family LLMs developed by Anthropic, an AI research company focused on creating safe and reliable AI. Claude can be used to build AI applications that support various functions, such as engaging in conversations, brainstorming ideas, and analyzing documents. Notably, Claude has demonstrated strengths in handling sensitive topics and maintaining consistency across extended conversational threads. In therapeutic settings, such as ADHD (Attention Deficit Hyperactivity Disorder) support, Claude has been employed as a virtual therapy assistant, fostering an environment that validates patients' emotions and experiences [265]. This example illustrates the potential of integrating Claude models with indoor robots to assist with a range of applications in smart environments, beneficial in enhancing user comfort and engagement.

Another example is ChatGLM, developed by Zhipu AI, a customized language model designed for AI-driven applications. ChatGLM is pre-trained using an autoregressive blank-filling objective and can be fine-tuned for various natural language understanding and generation tasks. For instance, the study in [241] showcases ChatGLM in a multi-agent AI system, where each AI agent corresponds to an LLM model, enabling intelligent processing of complex IoT data in a collaborative manner. This system assigns specific roles to AI agents-such as data analysis and decision-making-allowing them to handle diverse data inputs, including temperature, humidity, and image data. Additionally, the system integrates functions for memory management, summarization, and classification to streamline communication between agents. As a result, the multi-agent LLM deployment system outperforms single-agent configurations in IoT environments, optimizing data processing and minimizing errors. This approach enhances accuracy and reliability for real-time monitoring and anomaly detection across distributed sensor networks.

5.6. Small and tiny language models

These models often referred to SLMs or on-device LLMs, are gaining attention for their ability to shift processing tasks directly onto devices, reducing dependence on the cloud and improving latency, data localization, and personalized user experiences. These on-device models, typically containing fewer parameters (e.g., around 10 billion), are optimized for edge devices and support the development of responsive technologies like smart environments. Key examples of these models include Gemini Nano, Nexa AI Octopus, Apple OpenELM, Ferret-v2, Microsoft Phi, MiniCPM, Gemma, LLaMA, ChatGLM, Qwen, Yi, Mistral, and InternLM [266].

In the literature, models like Gemma (2B) and Phi-2 (2.7B) have been employed to create human-centric smart devices with a focus on privacy and user-friendly interactions. These models allow devices to respond to loosely defined commands and explain actions independently of cloud connectivity. Their operations follow a five-step process, including state modeling, synthetic data generation, and fine-tuning, enabling small LLMs to interpret and respond effectively to user commands on-device, even on compact hardware such as a Raspberry Pi. Case studies on devices such as lamps and thermostats illustrate how these models efficiently adapt to varied user requests, providing meaningful responses and actions based on embedded knowledge, which demonstrates the potential for SLMs in smart environments while bypassing the computational demands of larger models [267].

MobileLLM is another on-device language model optimized for AI applications specifically designed for mobile devices. MobileLLMs leverage advanced techniques such as deep-thin architecture, layer sharing, and grouped query attention. The deep-thin architecture emphasizes model depth (the number of layers) over width (layer dimensions) in sub-billion parameter models, enhancing performance on zero-shot commonsense reasoning tasks [268]. In smart spaces, implementing MobileLLMs on widely used and versatile mobile devices would enable seamless device-to-environment interactions and provide an accessible medium for user engagement within the space.

In addition to SLMs, tiny language models are designed with even fewer parameters-often starting at around one billion-to operate in resource-constrained environments where both computational power and memory are limited. Examples of tiny models include TinyBERT [269], TinyLlama-1.1B [270], MobileLlama-1.4B, Qwen-1.8B, PanGu-1B, and Phi-2.7B [271]. These tiny models typically use methods such as model pruning, quantization, knowledge distillation, and efficient architecture design. Model pruning reduces the model size by eliminating unnecessary weights, while quantization represents weights and activations with lower data types

Table 5
Performance benchmarks for LLM deployment strategies in smart spaces.

Deployment Strategy	Platform	Latency Improvement	Energy & Resource Efficiency	Baseline
Edge-only Deployment				
Phase-aware task distribution ELLIE [272]	Intel Core Ultra (CPU, GPU, NPU)	0.17x increased latency	1.8× lower energy and 1.5× lower energy delay product	GPU-only on-device
Multi-dimensional optimization m^{2}LLM [273]	Snapdragon 8 Gen 3	2.99–13.5× TTFT reduction	2.28–24.3× energy savings, accuracy loss %2–7	On-device frameworks
Lower-bit quantization LLMPi [274]	Raspberry Pi 5 (8 GB)	71× speedup tokens-per-second	4.3× tokens-per-joule	FP16 on-device
Collaborative Edge Deployment				
Hybrid model parallelism Co-Former [275]	Multi-Jetson Nano, Orin, TX2	1.7–3.1× inference speedup	36.3–63.8% energy reduction and 76% memory savings	Single edge device
Co-located collaborative inference Galaxy [276]	Multi-Jetson cluster	1.3–2.5× latency reduction	Distributed memory sharing	Single edge device
Hybrid Edge–Cloud Deployment				
Adaptive workload offloading CE-CoLLM [277]	Edge and Cloud (A100)	13.81% lower inference latency	84.53% workload offloaded to edge, 99% transfer reduction	Cloud-only
Distributed model sharding EdgeShard [278]	Multi-Jetson and Cloud GPU	50% latency reduction	2× throughput	Edge-only and hybrid baselines
Speculative decoding [279]	Jetson Nano, RTX 2080 Ti and Cloud	35% latency reduction	52% API cost reduction, 21% speedup robot control	Cloud autoregressive

(e.g., using 8-bit integers instead of 32-bit floating points). Knowledge distillation trains smaller models to emulate larger ones, and efficient architecture design creates models requiring fewer computations and less memory. By incorporating these techniques, small and tiny language models deliver improved performance on resource-constrained devices, reducing energy consumption and broadening AI adoption. In smart spaces, tiny language models can be implemented on resource-constrained, battery-powered sensors and IoT devices which have limited processing capabilities.

5.7. LLM Deployment strategies

For smart space applications, deploying LLMs on the edge or in the cloud requires careful consideration of key metrics such as latency, energy efficiency, privacy, and scalability. Depending on the application's specific requirements, LLMs can be deployed using edge-only, collaborative edge, or hybrid edge-cloud strategies, each exhibiting distinct performance characteristics across these metrics. In this subsection, we review the literature and examine the latency, energy efficiency, and resource utilization of each deployment strategy. Table 5.7 provides quantitative benchmarks for the three approaches, highlighting improvements in latency, energy efficiency, and resource utilization. The table also facilitates a direct comparison with baseline deployments, including edge-only and cloud-only configurations.

Edge-only deployment: This strategy executes LLMs directly on local devices without relying on external connectivity. By processing data locally, it provides strong privacy guarantees, eliminating the need to transmit sensitive information to the cloud. Despite the inherent resource constraints of edge devices, techniques such as hardware-aware optimizations, model quantization, and parallel processing can achieve significant performance gains. For example, deploying m2LLM (an LLM inference algorithm) using an edge-only strategy achieves a 3.0-13.5 × reduction in time to first token (TTFT) and 2.3-24.3 × energy savings through multi-dimensional optimization [273]. LLMs typically employ mixed precision, using 16-bit floating-point (FP16) for most computations with some FP32 operations. Quantizing models to lower bit precision, such as 8-bit or 4-bit (INT8 or INT4), reduces memory usage and can increase processing speed on supported hardware, albeit at a potential cost to accuracy. For instance, the LLMPi quantization technique delivers a 71 × throughput improvement (tokens per second) and 4.3 × higher energy efficiency (tokens per joule) on a Raspberry Pi [274]. Similarly, ELLIE (Energy-Efficient LLM Inference at the Edge) demonstrates TTFT ranging from 43-552 ms with 1.8 × lower energy consumption by intelligently distributing tasks across CPU, GPU, and NPU [272]. Edge-only deployment is particularly well-suited for privacy-sensitive applications with moderate-scale models and predictable workloads. However, computational capacity remains a fundamental limitation for large or highly dynamic models.

Collaborative edge deployment: This deployment strategy distributes LLM inference across multiple co-located devices within the same physical smart space, overcoming the limitations of single-device processing while avoiding reliance on the cloud. By leveraging the combined computational power of multiple devices, it enables improved latency, energy efficiency, and resource utilization. For example, the Galaxy collaborative edge AI system achieves a 1.3-2.5 × reduction in inference latency by pooling local computational resources [276]. Galaxy employs hybrid model parallelism to orchestrate collaborative LLM inference for models such as GPT2-L, BERT-L, and DistilBERT across both homogeneous and heterogeneous edge devices, including multiple Jetson Nano units

Table 6
AI applications in smart spaces.

AI Methodology	Applications	Limitations
Conventional machine learning models		
	Real-time people counting [55,185]	Privacy concerns
	Human activity recognition [182,195]	Needs expert-labeled data
RF, SVM, kNN	Fall detection [184,218]	Requires wearable sensors
Deep learning models		
	Real-time people counting [55,185]	Privacy issues
	Smart IoT video surveillance [192]	Misclassification of similar activities
CNN, RNN, LSTM,	Human activity and gesture recognition [197,211,226]	Sensitive to environmental interference
GRU, YOLO	Indoor localization [16,193,213] and person identification [210]	High computational cost
Autoencoders	Anomaly detection in IoT sensor data [200,201]	High computational cost and requires labeled data
Transformer networks		
	Human activity recognition, action anticipation [197,204,207]	Requires labeled sensor data
Transformer networks	Indoor localization [16,213], person identification [210]	Misclassification of similar activities
Vision transformers	Continuous health monitoring, Wi-Fi fingerprinting [217,231]	Privacy concerns and computational complexity
	Power forecasting [219], load monitoring [220]	Reward function complexity
	Multi-zone HVAC optimization [223]	Sensitivity to seasonal variation
Transformer networks and	Energy management [191]	High computational complexity
DRL, ANN hybrid AI models	Building power consumption forecasting [221,222]	Limited generalization
Large Language Models		
		Requires extensive historical data
GPT-4, Gemini and hybrid LLM	Air quality prediction [198]	High computational cost and long training
models with DRL, LSTM, GRU	Autonomous and real-world office HVAC control [256,264]	Dependency on cloud and occupant feedback
	Health decline detection [224]	
	Activity sensing [240]	Computational burden and hallucination risk
BERT, GPT-4 and	Heterogeneous IoT data processing [241]	Privacy concerns with cloud use
Multi-agent LLM	Clinical-trial eligibility classification [252]	Ethical, bias concerns
		Command interpretation variability
	Smart home assistants [248,255]	Context misunderstanding
GPT-3, GPT-4, Llama3-8B	Smart home management [250,259]	Unstable reasoning, latency and hallucinations
BERT, GPT, LLaMA	Transparent recommendation explanations [15]	Biases in personalization and user data privacy

with varying memory configurations. Similarly, CoFormer, a collaborative edge strategy designed for heterogeneous edge devices, decomposes large transformer models such as GPT2-XL (1.6B parameters) into smaller sub-models for distributed inference [275]. CoFormer achieves 1.7-3.1× speedups, 36.3-63.8% energy reduction, and 76.3% lower per-device memory usage, demonstrating both efficiency and scalability. Collaborative edge deployment is particularly suitable for multi-device smart buildings and dense IoT environments with high-speed local networks. However, it requires careful orchestration to manage device heterogeneity, maintain network stability, and synchronize distributed computations, ensuring consistent and reliable performance across all participating devices.

Hybrid edge-cloud deployment: This strategy dynamically partitions LLM inference between local edge devices and remote cloud resources, achieving an optimal balance of latency, computational efficiency, and scalability. By intelligently distributing workloads, it combines the low-latency responsiveness of edge computing with the high processing power of the cloud, enabling efficient execution of large-scale, real-time models that would be infeasible on edge devices alone. For example, CE-CoLLM, a cloud-edge collaboration framework, reduces end-to-end inference latency by 13.8%, offloads 84.5% of the workload from the cloud to the edge, cuts transmitted data by 99%, and lowers cloud compute time by 70-85%, leveraging NVIDIA A100 GPUs for cloud inference [277]. Similarly, speculative edge-cloud decoding achieves 35% lower latency and 52% reduction in operational API costs [279]. EdgeShard, which is another hybrid edge-cloud strategy partitions LLMs such as Llama 2 models into smaller shards and deploys them across edge and cloud nodes, further demonstrates 45-50% latency reduction through optimized task allocation [278]. Hybrid edge-cloud deployment is particularly suited for computation-intensive applications, including multimodal processing, large-scale language understanding, and complex reasoning. In these strategies, edge devices handle real-time inference, while cloud resources perform heavy computation and aggregation, ensuring both speed and capacity. This approach provides a scalable, energy-efficient, and latency-aware solution, making it ideal for dynamic smart spaces with fluctuating workloads and diverse application demands.

Deployment recommendations: Based on the comparative performance of the three deployment strategies, each approach is best suited for specific application contexts and system constraints within smart spaces. Edge-only deployment is recommended for privacy-critical and latency-sensitive applications with moderate computational requirements, where local processing is sufficient and data confidentiality is paramount. Typical examples include occupancy detection, environmental sensing, and simple activity recognition, where real-time responsiveness and on-device inference outweigh the need for large-scale computation. Collaborative edge deployment is ideal when multiple capable devices are co-located and interconnected via high-bandwidth local networks, particularly in environments where cloud connectivity is unreliable, costly, or undesirable. This strategy is well-suited for distributed sensor networks, where collaboration among heterogeneous devices enhances throughput, efficiency, and system resilience. Hybrid edge-cloud deployment should be employed for large-scale, computation-intensive applications where reliable network connectivity exists and controlled cloud participation is acceptable under defined privacy and latency constraints. Such applications include advanced HVAC optimization, multimodal user interaction, and complex reasoning or planning tasks, where edge devices handle real-time preprocessing while the cloud manages heavy computation and model aggregation. Ultimately, the selection of an appropriate deployment strategy depends on several key factors, including real-time requirements, privacy considerations, LLM model size and capability (e.g., multimodal processing of text, image, or voice), memory and processing capacity, inference frequency, data sensitivity, network reliability, hardware cost, and energy consumption. Therefore, deployment decisions should be carefully evaluated during the design and planning phases of smart spaces to ensure optimal performance, scalability, and user privacy.

5.8. Challenges and limitations

Privacy Risks: LLMs have shown a tendency to unintentionally disclose personally identifiable information (PII) from both their training data and user inputs, posing significant privacy risks. For instance, LLMs can replicate PII directly from user inputs, meaning that sensitive information-such as personal or health data-may inadvertently appear in generated outputs, even when privacy compliance is emphasized [280]. Models like ChatGPT, for example, may retain and reveal names, addresses, or other private details, despite efforts to filter this information. This issue is particularly concerning in smart spaces, where LLMs interact dynamically with user data, amplifying the potential for privacy breaches if PII is not well-managed. Although existing privacy-preserving methods aim to reduce these risks, they have limitations [281]. Differential privacy algorithms, for example, propose adding noise into the data to minimize memorization of specific data points and reduce PII retention. In addition, Hybrid approaches, which combine LLMs with structured privacy-preserving modules, employ techniques such as PII detection, redaction, and differential privacy at the inference stage. However, these methods often create a trade-off between privacy and model performance, adding computational overhead that may not be feasible for the low-latency requirements of smart spaces. Ultimately, privacy concerns in LLMs highlight the need for innovation in model architectures. Developing solutions that protect sensitive data while maintaining efficiency and compliance will be essential for advancing LLM applications in sensitive environments such as smart spaces.

Security Risks: Malicious attacks may exploit LLMs to bypass detection mechanisms, manipulate outputs, or compromise the integrity of smart environments. One of the major threats is adversarial attacks, where attackers subtly modify input data to influence the model's responses. Even small alterations, such as minor word substitutions or paraphrasing, can bypass common monitoring approaches such as classifiers and watermarking systems that are typically used to secure LLM outputs. Attackers may also craft prompts that instruct LLMs to generate responses in a style or vocabulary that evades detection [282,283]. To mitigate these vulnerabilities, methods such as instructional prompt filtering, which adjusts input processing to detect adversarial patterns, and advanced watermarking, which embeds distinctive markers in LLM outputs, have shown promise in enhancing model robustness. However, these approaches add computational overhead and are not entirely foolproof. This highlights the need for lightweight, scalable security solutions that protect LLMs while meeting the real-time needs of smart environments.

Limitations in Reasoning and Planning: Despite their proficiency in generating human-like text, LLMs exhibit critical limitations in reasoning and planning capabilities, which are essential for real-world applications like smart spaces. In these environments, consistent and reliable decision-making based on real-time data is paramount, yet LLMs often produce unpredictable or inaccurate outputs due to their reliance on patterns from training data rather than true logical inference or causal understanding. LLMs struggle to generalize reasoning processes across varying contexts, often resulting in flawed or overly simplistic responses. Unlike symbolic systems that follow explicit logical rules, LLMs rely on predicting token sequences based on statistical patterns, limiting their ability to engage in the conditional or iterative reasoning necessary for complex, multi-step deductions. For instance, models like BERT focus on superficial patterns rather than learning transferable logical rules, leading to reasoning errors when encountering unfamiliar data [284]. In addition to reasoning limitations, LLMs face significant challenges in planning, which is crucial for dynamically adapting in smart spaces. While LLMs can mimic structured reasoning in sequence generation, they often lack an understanding of causality or state-dependent conditions. This leads to failures in generating executable plans in real-world settings [285]. Models such as GPT-4 and Claude demonstrate limited success in producing consistent and valid plans autonomously, particularly in scenarios that deviate from their training examples. Their inability to self-verify or correct plans further restricts their application in tasks requiring high precision and reliability [286].

To address these limitations, neurosymbolic and hybrid frameworks have been proposed. For instance, LLM-Modulo frameworks integrate symbolic reasoning tools to improve both reasoning and planning robustness. However, these methods add computational overhead and complexity, posing practical barriers for real-time applications in smart spaces, where efficiency and responsiveness are essential [287]. Despite advancements in neurosymbolic frameworks, they face challenges with tasks requiring multi-step logic, showing significant performance drops as complexity increases. Their inability to filter out irrelevant information amplifies errors when extraneous context is introduced [288]. These limitations address the need for hybrid frameworks that combine traditional AI techniques with the capabilities of LLMs, incorporating structured reasoning components to address the demands of smart environments [289]. Fortunately, the advancements in Large Reasoning Models (LRMs), such as OpenAI's o1 model, can overcome these reasoning limitations of LLMs. By leveraging techniques like Chain-of-Thought (CoT) fine-tuning and reinforcement learning, the o1 model demonstrates enhanced proficiency in solving complex reasoning tasks, including multi-step logic problems and mathematical computations. For example, the Marco-o1 LRM [290] has been successfully applied to real-world problem-solving tasks, showcasing its potential for tackling complex challenges. This capability aligns with the needs of smart spaces, where dynamic and adaptive decision-making is critical.

Hallucination Risk: Hallucinations in LLMs refer to instances where the model generates information that is not based in factual or accurate data. This phenomenon occurs when the model produces content absent from the input data, often influenced by internal biases or overconfidence. Hallucinations pose significant risks in applications requiring precise and contextually accurate information [291]. Hallucinations persist across LLM generations due to the probabilistic nature of token prediction and the lack of true comprehension. Even with extensive training data, LLMs show limitations in fact-verification and struggle to differentiate between high-confidence predictions and unverified associations [292]. Techniques such as retrieval-augmented generation (RAG), which allows LLMs real-time access to authoritative databases, and real-time verification frameworks such as EVER, which detect and correct inaccuracies as they arise, have shown promise in reducing hallucination rates. However, these methods increase computational demands, which can hinder their practicality in latency-sensitive applications [293]. In smart spaces, where ongoing and accurate interaction with the physical world is essential, hallucinations can introduce substantial risks, pinpointing the need for solutions that ensure LLMs provide reliably accurate outputs in real-time contexts.

Computational Overhead, Real-Time Constraints, and Energy Demands: The use LLMs in smart environments places significant demands on processing power, memory, and response time. These requirements pose particular challenges for resource-constrained edge devices, which often lack the capacity to meet the high computational needs of LLMs, leading to increased latency [294]. To address these limitations, strategies such as model compression, edge-cloud collaboration, the use of smaller and tiny language models, and quantization techniques are under development to better manage computational resources.

The latency associated with generating outputs from LLMs can compromise their effectiveness in smart spaces. To mitigate this, hybrid approaches have been proposed that integrate LLMs with locally deployed sensors to perform preliminary inference, before transferring data to cloud-based LLMs for more complex analysis. This approach would reduce latency and enhance responsiveness [295]. However, these approaches are still in early stages and may not be suitable for some smart space applications that require processing large datasets.

LLMs also have high energy demands, consuming substantial power during both training and inference. For example, studies have shown that models such as LLaMA require significant computing and memory resources across multiple GPUs, resulting in high energy consumption and increased carbon emissions [296]. In response, studies introduce carbon-efficient architectures, such as distributed computing and parallel execution, to help reduce computational loads. However, these architectures present practical challenges in real-time applications for smart spaces [297]. Distributed models require extensive inter-device communication, which increases network traffic and bandwidth utilization. Hyperparameter tuning has shown promise as an alternative, reducing unnecessary computations by optimizing model parameters to lower power consumption during training and inference. However, tuning itself can require significant computation, which may not align with the energy constraints of smart space applications [298]. In summary, while LLMs hold promise for enhancing smart spaces, their deployment is constrained by energy-related challenges that may limit practical use in these settings.

6. Conclusion

Advancements in AI methodologies, combined with innovations in sensing, communication, and computing technologies, pave the way for the development of interactive AI-enabled smart spaces. In this survey article, we reviewed the creation of AI-driven smart environments, focusing on essential components such as sensor and communication technologies, protocols, conventional machine learning, and large language models-all crucial for creating the smart spaces. In each section, we also provided a summary of the challenges and limitations associated with each of these components. Throughout our discussion, we highlighted real-world applications powered by these technologies. Our findings indicate that effective AI-driven smart spaces necessitate a careful selection of sensor devices, communication technologies, and computing platforms. In addition, reliable data collection through regular sensor calibration, robust connectivity, and privacy and security mechanisms are essential for smart spaces. The use of AI methodologies is essential for empowering applications, improving efficiency, and enhancing the functionality of smart spaces. By enabling advanced features such as personalized comfort settings, interactive living environments, and automation of system controls, AI significantly enriches the indoor experiences of users. Furthermore, the use of LLMs can facilitate interactive living environments, ultimately enhancing user experience and promoting sustainable living practices.

CRediT authorship contribution statement

Aygün Varol: Writing – original draft, Resources, Methodology, Investigation; **Naser Hossein Motlagh:** Writing – original draft, Supervision, Funding acquisition, Conceptualization; **Mirka Leino:** Writing – review & editing; **Sasu Tarkoma:** Writing – review & editing; **Johanna Virkki:** Writing – review & editing, Resources, Funding acquisition.

Data availability

No data was used for the research described in the article.

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests:

Aygün Varol reports financial support was provided by Tampere University and Tampere University of Applied Sciences. Aygün Varol reports a relationship with Tampere University and Tampere University of Applied Sciences that includes: employment. If there are other authors, they declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Grant information

This research is supported by the Research Council of Finland with grant number 362594 and by the Ministry of National Education of the Republic of Türkiye, and by the Jane and Aatos Erkko Foundation under the EVIL-AI project.

References

[1] M.W. Aziz, A.A. Sheikh, E.A. Felemban, Requirement engineering technique for smart spaces, in: Proceedings of the International Conference on Internet of Things and Cloud Computing, 2016, pp. 1–7. https://doi.org/10.1145/2896387.2896439

[2] N.H. Motlagh, M.A. Zaidan, L. Lovén, P.L. Fung, T. Hänninen, R. Morabito, P. Nurmi, S. Tarkoma, Digital twins for smart spaces-beyond IoT analytics, IEEE Internet Things J. 11 (1) (2023) 573–583. https://doi.org/10.1109/JIOT.2023.3287032

[3] S.B. Goyal, P. Bedi, D.K. Yadav, N.A. Vakil, Internet of things information analysis using fusion based learning with deep neural network, in: Journal of Physics: Conference Series, 1714, IOP Publishing, 2021, p. 012022. https://doi.org/10.1088/1742-6596/1714/1/012022

[4] A. Akintola, J. Ma, N.H. Motlagh, G. Bouloukakis, H. Flores, Autonomous indoor customization: the future of human comfort and productivity, Computer: Publicat. IEEE Comput. Soc. (2025).

[5] A. Sleem, I. Elhenawy, Survey of artificial intelligence of things for smart buildings: a closer outlook, J. Intell. Syst. IoT 8 (2) (2023). https://doi.org/10.54216/JISIoT.080206

[6] A. Almusaed, I. Yitmen, A. Almssad, Enhancing smart home design with AI models: a case study of living spaces implementation review, Energies 16 (6) (2023) 2636. https://doi.org/10.3390/en16062636

[7] C.I. Nwakanma, G.O. Anyanwu, L.A.C. Ahakonye, J.-M. Lee, D.-S. Kim, A review of thermal array sensor-based activity detection in smart spaces using AI, ICT Express 10 (2) (2024) 256–269. https://doi.org/10.1016/j.icte.2023.11.007

[8] I.Y.L. Lee, T. Nguyen-Duc, R. Ueno, J. Smith, P.Y. Chan, Use of a multiscale vision transformer to predict nursing activities score from low-Resolution thermal videos in an intensive care unit, IEEE Sensors Lett. 8 (7) (2024) 1–4. https://doi.org/10.1109/LSENS.2024.3408320

[9] T. Sutjarittham, H.H. Gharakheili, S.S. Kanhere, V. Sivaraman, Experiences with IoT and AI in a smart campus for optimizing classroom usage, IEEE Internet Things J. 6 (5) (2019) 7595–7607. https://doi.org/10.1109/JIOT.2019.2902410

[10] B. Abade, D. Perez Abreu, M. Curado, A non-intrusive approach for indoor occupancy detection in smart environments, Sensors 18 (11) (2018) 3953. https://doi.org/10.3390/s18113953

[11] M. Saleem, M.S. Khan, G.F. Issa, A. Khadim, M. Asif, A.S. Akram, H.K.G. Nair, Smart spaces: occupancy detection using adaptive back-propagation neural network, in: 2023 International Conference on Business Analytics for Technology and Security (ICBATS), IEEE, 2023, pp. 1–6. https://doi.org/10.1109/ICBATS57792.2023.10111286

[12] N.H. Motlagh, P. Toivonen, M.A. Zaidan, E. Lagerspetz, E. Peltonen, E. Gilman, P. Nurmi, S. Tarkoma, Monitoring social distancing in smart spaces using infrastructure-based sensors, in: 2021 IEEE 7Th World Forum on Internet of Things (WF-IoT), IEEE, 2021, pp. 124–129. https://doi.org/10.1109/WF-IoT51360.2021.9595897

[13] A. Almeida, U. Bermejo, A. Bilbao, G. Azkune, U. Aguilera, M. Emaldi, F. Dornaika, I. Arganda-Carreras, A comparative analysis of human behavior prediction approaches in intelligent environments, Sensors 22 (3) (2022) 701. https://doi.org/10.3390/s22030701

[14] S. Khelifi, A. Morris, Mixed reality IoT smart environments with large language model agents, in: 2024 IEEE 4Th International Conference on Human-Machine Systems (ICHMS), 2024, pp. 1–7. https://doi.org/10.1109/ICHMS59971.2024.10555610

[15] L. Wu, Z. Zheng, Z. Qiu, H. Wang, H. Gu, T. Shen, C. Qin, C. Zhu, H. Zhu, Q. Liu, et al., A survey on large language models for recommendation, World Wide Web 27 (5) (2024) 60. https://doi.org/10.1007/s11280-024-01291-2

[16] F. Jovan, C. Morgan, R. McConville, E.L. Tonkin, I. Craddock, A. Whone, Multimodal indoor localisation in Parkinson's disease for detecting medication use: observational pilot study in a free-Living setting, in: Proceedings of the 29Th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, 2023, pp. 4273–4283. https://doi.org/10.1145/3580305.3599872

[17] A. Saleh, S. Tarkoma, P.K. Donta, N.H. Motlagh, S. Dustdar, S. Pirttikangas, L. Lovén, Usercentrix: An agentic memory-augmented ai framework for smart spaces, arXiv preprint (2025). arXiv:2505.00472

[18] Y. Lu, L. Zhou, A. Zhang, S. Zha, X. Zhuo, S. Ge, Application of deep learning and intelligent sensing analysis in smart home, Sensors 24 (3) (2024) 953. https://doi.org/10.3390/s24030953

[19] W. Alsafery, O. Rana, C. Perera, Sensing within smart buildings: a survey, ACM Comput. Surv. 55 (13s) (2023) 1–35. https://doi.org/10.1145/3596600

[20] A.G. Putrada, M. Abdurohman, D. Perdana, H.H. Nuha, Machine learning methods in smart lighting toward achieving user comfort: a survey, IEEE Access 10 (2022) 45137–45178. https://doi.org/10.1109/ACCESS.2022.3169765

[21] G. Diraco, G. Rescio, A. Caroppo, A. Manni, A. Leone, Human action recognition in smart living services and applications: context awareness, data availability, personalization, and privacy, Sensors 23 (13) (2023) 6040. https://doi.org/10.3390/s23136040

[22] F. Gu, M.-H. Chung, M. Chignell, S. Valaee, B. Zhou, X. Liu, A survey on deep learning for human activity recognition, ACM Comput. Surv. (CSUR) 54 (8) (2021) 1–34. https://doi.org/10.1145/3472290

[23] G. Diraco, G. Rescio, P. Siciliano, A. Leone, Review on human action recognition in smart living: sensing technology, multimodality, real-time processing, interoperability, and resource-constrained processing, Sensors 23 (11) (2023) 5281. https://doi.org/10.3390/s23115281

[24] L. Babangida, T. Perumal, N. Mustapha, R. Yaakob, Internet of things (IoT) based activity recognition strategies in smart homes: a review, IEEE Sens. J. 22 (9) (2022) 8327–8336. https://doi.org/10.1109/JSEN.2022.3161797

[25] D. Bouchabou, S.M. Nguyen, C. Lohr, B. LeDuc, I. Kanellos, A survey of human activity recognition in smart homes based on IoT sensors algorithms: taxonomies, challenges, and opportunities with deep learning, Sensors 21 (18) (2021) 6037. https://doi.org/10.3390/s21186037

[26] L. Yu, S. Qin, M. Zhang, C. Shen, T. Jiang, X. Guan, A review of deep reinforcement learning for smart building energy management, IEEE Internet Things J. 8 (15) (2021) 12046–12063. https://doi.org/10.1109/JIOT.2021.3078462

[27] H. Zhang, S. Seal, D. Wu, F. Bouffard, B. Boulet, Building energy management with reinforcement learning and model predictive control: a survey, IEEE Access 10 (2022) 27853–27862. https://doi.org/10.1109/ACCESS.2022.3156581

[28] Z. Nagy, G. Henze, S. Dey, J. Arroyo, L. Helsen, X. Zhang, B. Chen, K. Amasyali, K. Kurte, A. Zamzam, et al., Ten questions concerning reinforcement learning for building energy management, Build. Environ. 241 (2023) 110435. https://doi.org/10.1016/j.buildenv.2023.110435

[29] P. Lissa, C. Deane, M. Schukat, F. Seri, M. Keane, E. Barrett, Deep reinforcement learning for home energy management system control, Energy AI 3 (2021) 100043. https://doi.org/10.1016/j.egyai.2020.100043

[30] A. Shaqour, A. Hagishima, Systematic review on deep reinforcement learning-based energy management for different building types, Energies 15 (22) (2022) 8663. https://doi.org/10.3390/en15228663

[31] J. Ogundiran, E. Asadi, M. Gameiro da Silva, A systematic review on the use of AI for energy efficiency and indoor environmental quality in buildings, Sustainability 16 (9) (2024) 3627. https://doi.org/10.3390/su16093627

[32] P.W. Tien, S. Wei, J. Darkwa, C. Wood, J.K. Calautit, Machine learning and deep learning methods for enhancing building energy efficiency and indoor environmental quality–a review, Energy AI 10 (2022) 100198. https://doi.org/10.1016/j.egyai.2022.100198

[33] G.H. Merabet, M. Essaaidi, M.B. Haddou, B. Qolomany, J. Qadir, M. Anan, A. Al-Fuqaha, M.R. Abid, D. Benhaddou, Intelligent building control systems for thermal comfort and energy-efficiency: a systematic review of artificial intelligence-assisted techniques, Renew. Sustain. Energy Rev. 144 (2021) 110969. https://doi.org/10.1016/j.rser.2021.110969

[34] A. Sivanathan, H.H. Gharakheili, F. Loi, A. Radford, C. Wijenayake, A. Vishwanath, V. Sivaraman, Classifying IoT devices in smart environments using network traffic characteristics, IEEE Trans. Mob. Comput. 18 (8) (2018) 1745–1759. https://doi.org/10.1109/TMC.2018.2866249

[35] N.S. Baqer, H. Mohammed, A.S. Albahri, et al., Development of a real-time monitoring and detection indoor air quality system for intensive care unit and emergency department, Signa Vitae 19 (1) (2023). https://doi.org/10.22514/sv.2022.013

[36] D. Bousiotis, L.-N.S. Alconcel, D.C.S. Beddows, R.M. Harrison, F.D. Pope, Monitoring and apportioning sources of indoor air quality using low-cost particulate matter sensors, Environ. Int. 174 (2023) 107907. https://doi.org/10.1016/j.envint.2023.107907

[37] M. Lopes, J. Reis, A.P. Fernandes, D. Lopes, R. Lourenço, T. Nunes, C.H.G. Faria, C. Borrego, A.I. Miranda, Indoor air quality study using low-cost sensors, WIT Trans. Ecol. Environ. 244 (2020) 1–13.

[38] N.H. Motlagh, M.A. Zaidan, E. Lagerspetz, S. Varjonen, J. Toivonen, J. Mineraud, A. Rebeiro-Hargrave, M. Siekkinen, T. Hussein, P. Nurmi, et al., Indoor air quality monitoring using infrastructure-based motion detectors, in: 2019 IEEE 17Th International Conference on Industrial Informatics (INDIN), 1, IEEE, 2019, pp. 902–907. https://doi.org/10.1109/INDIN41052.2019.8972332

[39] M. Kong, B. Dong, R. Zhang, Z. O'Neill, HVAC Energy savings, thermal comfort and air quality for occupant-centric control through a side-by-side experimental study, Appl. Energy 306 (2022) 117987. https://doi.org/10.1016/j.apenergy.2021.117987

[40] N. Morresi, V. Cipollone, S. Casaccia, G.M. Revel, Measuring thermal comfort using wearable technology in transient conditions during office activities, Measurement 224 (2024) 113897. https://doi.org/10.1016/j.measurement.2023.113897

[41] A. Aryal, B. Becerik-Gerber, Thermal comfort modeling when personalized comfort systems are in use: comparison of sensing and learning methods, Build. Environ. 185 (2020) 107316. https://doi.org/10.1016/j.buildenv.2020.107316

[42] D. Li, C.C. Menassa, V.R. Kamat, E. Byon, HEAT-Human embodied autonomous thermostat, Build. Environ. 178 (2020) 106879. https://doi.org/10.1016/j.buildenv.2020.106879

[43] Y. Feng, J. Wang, N. Wang, C. Chen, Alert-based wearable sensing system for individualized thermal preference prediction, Build. Environ. 232 (2023) 110047. https://doi.org/10.1016/j.buildenv.2023.110047

[44] Q. Fan, X. Xu, P. Liu, H. Zhang, S. Tang, A data-driven framework for thermal comfort assessment method based on user interaction, J. Build. Eng. 82 (2024) 108294. https://doi.org/10.1016/j.jobe.2023.108294

[45] J. Ma, D. Panic, R. Yus, G. Bouloukakis, Co-zybench: using co-simulation and digital twins to benchmark thermal comfort provision in smart buildings, in: 2024 IEEE International Conference on Pervasive Computing and Communications (PerCom), IEEE, 2024, pp. 25–35.

[46] L. Hallett, M. Tatum, G. Thomas, S. Sousan, K. Koehler, T. Peters, An inexpensive sensor for noise, J. Occup. Environ. Hyg. 15 (5) (2018) 448–454.

[47] J. Picaut, A. Can, N. Fortin, J. Ardouin, M. Lagrange, Low-cost sensors for urban noise monitoring networks-a literature review, Sensors 20 (8) (2020) 2256.

[48] A. Kumar, A. Singh, A. Kumar, M.K. Singh, P. Mahanta, S.C. Mukhopadhyay, Sensing technologies for monitoring intelligent buildings: a review, IEEE Sens. J. 18 (12) (2018) 4847–4860.

[49] B. Dong, V. Prakash, F. Feng, Z. O'Neill, A review of smart building sensing system for better indoor environment control, Energy Build. 199 (2019) 29–46.

[50] N.H. Motlagh, S.H. Khajavi, A. Jaribion, J. Holmstrom, An IoT-based automation system for older homes: a use case for lighting system, in: 2018 IEEE 11Th Conference on Service-Oriented Computing and Applications (SOCA), IEEE, 2018, pp. 1–6.

[51] T. Dissanayake, T. Maekawa, D. Amagata, T. Hara, Detecting door events using a smartphone via active sound sensing, Proc. ACM Interact. Mobile, Wearable Ubiquitous Technol. 2 (4) (2018) 1–26.

[52] L. Wu, Y. Wang, A low-power electric-mechanical driving approach for true occupancy detection using a shuttered passive infrared sensor, IEEE Sens. J. 19 (1) (2018) 47–57.

[53] C. Perra, A. Kumar, M. Losito, P. Pirino, M. Moradpour, G. Gatto, Monitoring indoor people presence in buildings using low-cost infrared sensor array in doorways, Sensors 21 (12) (2021) 4062.

[54] A. Fasolino, P. Vitolo, R. Liguori, L. Di Benedetto, A. Rubino, G.D. Licciardo, D. Pau, Object classification using ultra low resolution time-of-Flight sensor and tiny convolutional neural network, in: 2024 IEEE Sensors Applications Symposium (SAS), IEEE, 2024, pp. 1–6.

[55] C. Xie, F. Daghero, Y. Chen, M. Castellano, L. Gandolfi, A. Calimera, E. Macii, M. Poncino, D.J. Pagliari, Efficient deep learning models for privacy-preserving people counting on low-resolution infrared arrays, IEEE Internet Things J. 10 (15) (2023) 13895–13907.

[56] C. Raghavachari, V. Aparna, S. Chithira, V. Balasubramanian, A comparative study of vision based human detection techniques in people counting applications, Procedia Comput. Sci. 58 (2015) 461–469.

[57] C. Xie, D.J. Pagliari, A. Calimera, Energy-efficient and privacy-aware social distance monitoring with low-resolution infrared sensors and adaptive inference, in: 2022 17Th Conference on Ph. D Research in Microelectronics and Electronics (PRIME), IEEE, 2022, pp. 181–184.

[58] V. Tsakanikas, T. Dagiuklas, Video surveillance systems-current status and future trends, Comput. Electric. Eng. 70 (2018) 736–753.

[59] O. Elharrouss, N. Almaadeed, S. Al-Maadeed, A review of video surveillance systems, J. Vis. Commun. Image Represent. 77 (2021) 103116.

[60] A.A. Adams, J.M. Ferryman, The future of video analytics for surveillance and its ethical implications, Secur. J. 28 (2015) 272–289.

[61] Z. Shao, J. Cai, Z. Wang, Smart monitoring cameras driven intelligent processing to big surveillance video data, IEEE Trans. Big Data 4 (1) (2017) 105–116.

[62] R. Gade, T.B. Moeslund, Thermal cameras and applications: a survey, Mach. Vis. Appl. 25 (2014) 245–262.

[63] F. Altay, S. Velipasalar, The use of thermal cameras for pedestrian detection, IEEE Sens. J. 22 (12) (2022) 11489–11498.

[64] A. Naser, A. Lotfi, J. Zhong, Multiple thermal sensor array fusion toward enabling privacy-preserving human monitoring applications, IEEE Internet Things J. 9 (17) (2022) 16677–16688.

[65] M. Javaid, A. Haleem, S. Rab, R.P. Singh, R. Suman, Sensors for daily life: a review, Sensors Int. 2 (2021) 100121.

[66] S.C. Hsia, S.-H. Wang, S.-W. Hsu, Smart water-meter wireless transmission system for smart cities, IEEE Consum. Electron. Mag. 10 (6) (2020) 83–88.

[67] K.S. Cetin, Z. O'Neill, Smart meters and smart devices in buildings: a review of recent progress and influence on electricity use and peak demand, Curr. Sustain./Renew. Energy Rep. 4 (2017) 1–7.

[68] S.A. Nabavi, N.H. Motlagh, M.A. Zaidan, A. Aslani, B. Zakeri, Deep learning in energy modeling: application in smart buildings with distributed energy generation, IEEE Access 9 (2021) 125439–125461.

[69] X. Tang, M.-C. Huang, S. Mandal, An "internet of ears" for crowd-aware smart buildings based on sparse sensor networks, in: 2017 IEEE SENSORS, IEEE, 2017, pp. 1–3.

[70] J. Liu, H. Liu, Y. Chen, Y. Wang, C. Wang, Wireless sensing for human activity: a survey, IEEE Communications Surv. Tutor. 22 (3) (2019) 1629–1645.

[71] S. Savazzi, S. Sigg, M. Nicoli, V. Rampa, S. Kianoush, U. Spagnolini, Device-free radio vision for assisted living: leveraging wireless channel quality information for human sensing, IEEE Signal Process. Mag. 33 (2) (2016) 45–58.

[72] Y. Lin, D. Jiang, R. Yus, G. Bouloukakis, A. Chio, S. Mehrotra, N. Venkatasubramanian, Locater: cleaning wifi connectivity datasets for semantic localization, Proc. VLDB Endow. 14 (3) (2020) 329-341. https://doi.org/10.14778/3430915.3430923

[73] J. Wang, Q. Gao, M. Pan, Y. Fang, Device-free wireless sensing: challenges, opportunities, and applications, IEEE Netw. 32 (2) (2018) 132–137.

[74] A. Khalili, A.-H. Soliman, M. Asaduzzaman, A. Griffiths, Wi-Fi sensing: applications and challenges, J. Eng. 2020 (3) (2020) 87–97.

[75] O. Arshi, S. Mondal, Advancements in sensors and actuators technologies for smart cities: a comprehensive review, Smart Construct. Sustain. Cities 1 (1) (2023) 18.

[76] J. Torres-Sospedra, A. Ometov, Data from smartphones and wearables, 2021, https://doi.org/10.3390/data6050045

[77] F.J. Dian, R. Vahidnia, A. Rahmati, Wearables and the internet of things (IoT), applications, opportunities, and challenges: a survey, IEEE Access 8 (2020) 69200–69211.

[78] S.K. Phooi, L.-M. Ang, E. Peter, A. Mmonyi, Machine learning and AI technologies for smart wearables, Electronics (Basel) 12 (7) (2023) 1509.

[79] M.A. Zaidan, N.H. Motlagh, P.L. Fung, A.S. Khalaf, Y. Matsumi, A. Ding, S. Tarkoma, T. Petäjä, M. Kulmala, T. Hussein, Intelligent air pollution sensors calibration for extreme events and drifts monitoring, IEEE Trans. Ind. Inf. 19 (2) (2022) 1366–1379.

[80] Bluetooth, Bluetooth Technology, (https://www.bluetooth.com/). (accessed 30 September 2025).

[81] G. Shan, H. Lee, B.-h. Roh, Indoor localization-based energy management for smart home, in: 2022 IEEE PES 14Th Asia-Pacific Power and Energy Engineering Conference (APPEEC), IEEE, 2022, pp. 1–5.

[82] W.-F. Alliance, Wi-Fi Technology, (https://www.wi-fi.org/). (accessed 30 September 2025).

[83] R. Karmakar, S. Chattopadhyay, S. Chakraborty, Impact of IEEE 802.11 n/ac PHY/MAC high throughput enhancements on transport and application protocols-a survey, IEEE Commun. Surv. Tutor. 19 (4) (2017) 2050–2091.

[84] M.-D. Dianu, J. Riihijärvi, M. Petrova, Measurement-based study of the performance of IEEE 802.11 ac in an indoor environment, in: 2014 IEEE International Conference on Communications (ICC), IEEE, 2014, pp. 5771–5776.

[85] E. Mozaffariahrar, F. Theoleyre, M. Menth, A survey of wi-Fi 6: technologies, advances, and challenges, Future Internet 14 (10) (2022) 293.

[86] L. Tian, S. Santi, A. Seferagić, J. Lan, J. Famaey, Wi-Fi halow for the internet of things: an up-to-date survey on IEEE 802.11 ah research, J. Netw. Comput. Appl. 182 (2021) 103036.

[87] S. Aust, R.V. Prasad, I.G. Niemegeers, Outdoor long-range WLANs: a lesson for IEEE 802.11 ah, IEEE Commun. Surv. Tutor. 17 (3) (2015) 1761–1775.

[88] Z. Zhou, H. Yu, H. Shi, Optimization of wireless video surveillance system for smart campus based on internet of things, IEEE Access 8 (2020) 136434–136448.

[89] L. Alliance, LoRa Technology, (https://lora-alliance.org/). (accessed 30 September 2025).

[90] M. Ahsan, M.A. Based, J. Haider, E.M.G. Rodrigues, Smart monitoring and controlling of appliances using lora based IoT system, Designs 5 (1) (2021) 17.

[91] D. Culler, S. Chakrabarti, 6LoWPAN: Incorporating IEEE 802.15. 4 into the IP architecture, White paper (2009).

[92] A. Sanila, B. Mahapatra, A.K. Turuk, Performance evaluation of RPL protocol in a 6LoWPAN based smart home environment, in: 2020 International Conference on Computer Science, Engineering and Applications (ICCSEA), IEEE, 2020, pp. 1–6.

[93] A. Zohourian, S. Dadkhah, E.C.P. Neto, H. Mahdikhani, P.K. Danso, H. Molyneaux, A.A. Ghorbani, IoT Zigbee device security: a comprehensive review, IoT 22 (2023) 100791.

[94] D.-G. Akestoridis, M. Harishankar, M. Weber, P. Tague, Zigator: analyzing the security of zigbee-enabled smart homes, in: Proceedings of the 13Th ACM Conference on Security and Privacy in Wireless and Mobile Networks, 2020, pp. 77–88.

[95] Z.-W. Alliance, Z-Wave Long Range, (https://z-wavealliance.org/). (accessed 30 September 2025).

[96] C. Vattheuer, C. Liu, A. Abedi, O. Abari, Is z-wave reliable for iot sensors?, IEEE Sens. J. 23 (24) (2023) 31297–31306.

[97] T.T. Group, Thread 1.2 for Smart Buildings, (https://www.threadgroup.org/). (accessed 30 September 2025).

[98] D. Lan, Z. Pang, C. Fischione, Y. Liu, A. Taherkordi, F. Eliassen, Latency analysis of wireless networks for proximity services in smart home and building automation: the case of thread, IEEE Access 7 (2018) 4856–4867.

[99] L. Lubna, H. Hameed, S. Ansari, A. Zahid, A. Sharif, H.T. Abbas, F. Alqahtani, N. Mufti, S. Ullah, M.A. Imran, et al., Radio frequency sensing and its innovative applications in diverse sectors: a comprehensive study, Front. Commun. Netw. 3 (2022) 1010228.

[100] NFC, NFC Technology, (https://nfc-forum.org/). (accessed 30 September 2025).

[101] B. Ozdenizci, V. Coskun, K. Ok, NFC Internal: an indoor navigation system, Sensors 15 (4) (2015) 7571–7595.

[102] O. Liberg, M. Sundberg, E. Wang, J. Bergman, J. Sachs, Cellular Internet of things: technologies, standards, and performance, Academic Press, 2017.

[103] M.A. Obeidat, A.M. Mansour, T. Hamadneh, J. Abdullah, Remotely controlled smart home system using GSM and IOT, in: 2021 International Conference on Information Technology (ICIT), IEEE, 2021, pp. 748–753.

[104] H. Joe, H. An, W. Wang, W. Lee, H. Park, In-band cellular IoT for smart home applications, in: 2017 IEEE International Conference on Consumer Electronics (ICCE), IEEE, 2017, pp. 336–337.

[105] D. Xue, H. Xu, P. Li, An indoor 3D positioning technology based on NB-Iot, in: Web, Artificial Intelligence and Network Applications: Proceedings of the Workshops of the 33Rd International Conference on Advanced Information Networking and Applications (WAINA-2019) 33, Springer, 2019, pp. 35–43.

[106] J. Dizdarević, F. Carpio, A. Jukan, X. Masip-Bruin, A survey of communication protocols for internet of things and related challenges of fog and cloud computing integration, ACM Comput. Surv. (CSUR) 51 (6) (2019) 1–29.

[107] K. Sachs, S. Appel, S. Kounev, A. Buchmann, Benchmarking publish/subscribe-based messaging systems, in: Database Systems for Advanced Applications: 15Th International Conference, DASFAA 2010, International Workshops: GDM, BenchmarX, MCIS, SNSMW, DIEW, UDM, Tsukuba, Japan, April 1–4, 2010, Revised Selected Papers 15, Springer, 2010, pp. 203–214.

[108] A. Banks, R. Gupta, MQTT Version 3.1. 1. OASIS Standard, (http://docs.oasis-open.org/mqtt/mqtt/v3.1.1/os/mqtt-v3.1.1-os.html, 2014). (accessed 30 September 2025).

[109] A. Yudidharma, N. Nathaniel, T.N. Gimli, S. Achmad, A. Kurniawan, A systematic literature review: messaging protocols and electronic platforms used in the internet of things for the purpose of building smart homes, Procedia Comput. Sci. 216 (1) (2023) 194–203.

[110] Z. Shelby, K. Hartke, C. Bormann, The constrained application protocol (CoAP), Internet Engineering Task Force (IETF), (http://www.rfc-editor.org/rfc/rfc7252.txt, 2014). (accessed 30 September 2025).

[111] S. Bansal, D. Kumar, Enhancing constrained application protocol using message options for internet of things, Cluster Comput. 26 (3) (2023) 1917–1934.

[112] M. Koster, A. Keranen, J. Jimenez, Publish-subscribe broker for the constrained application protocol (CoAP), Internet Engineering Task Force, Internet-Draft draft-ietf-core-coap-pubsub-12, (https://www.ietf.org/archive/id/draft-ietf-core-coap-pubsub-12.html, 2023). (accessed 30 September 2025).

[113] O. Standard, OASIS Advanced Message Queuing Protocol (AMQP) Version 1.0, (http://docs.oasis-open.org/amqp/core/v1.0/os/amqp-core-overview-v1.0-os.html, 2012). (accessed 30 September 2025).

[114] T. Adiono, B.A. Manangkalangi, R. Muttaqin, S. Harimurti, W. Adijarto, Intelligent and secured software application for IoT based smart home, in: 2017 IEEE 6Th Global Conference on Consumer Electronics (GCCE), IEEE, 2017, pp. 1–2.

[115] Object Management Group (OMG), Data distribution service (DDS), (https://www.omg.org/spec/DDS/1.4/PDF, 2015). (accessed 30 September 2025).

[116] A. Saleh, S. Tarkoma, S. Pirttikangas, L. Lovén, Publish/subscribe for edge intelligence: systematic review and future prospects, Available at SSRN 4872730 (2024).

[117] R. Fielding, J. Gettys, J. Mogul, H. Frystyk, L. Masinter, P. Leach, T. Berners-Lee, Hypertext transfer protocol–HTTP/1.1, (http://www.rfc-editor.org/rfc/rfc2616.txt, 1999). (accessed 30 September 2025).

[118] C. Severance, Roy t. fielding: understanding the rest style, Computer (Long Beach Calif) 48 (6) (2015) 7–9.

[119] P. Saint-Andre, Extensible messaging and presence protocol (XMPP): Core. RFC 3920, Technical Report, Internet Engineering Task Force (IETF), 2011.

[120] T.I. Q.W. Group, The QUIC Protocol, (https://quicwg.org/). (accessed 30 September 2025).

[121] J. Iyengar, M. Thomson, et al., QUIC: A UDP-based multiplexed and secure transport, in: Rfc 9000, Internet Engineering Task Force (IETF) Fremont, CA, USA, 2021.

[122] E. Zanaj, G. Caso, L. De Nardis, A. Mohammadpour, Ö. Alay, M.-G. Di Benedetto, Energy efficiency in short and wide-area IoT technologies-a survey, Technologies 9 (1) (2021) 22.

[123] H. Jayakumar, K. Lee, W.S. Lee, A. Raha, Y. Kim, V. Raghunathan, Powering the internet of things, in: Proceedings of the 2014 International Symposium on Low Power Electronics and Design, 2014, pp. 375–380.

[124] M. Mansour, A. Gamal, A.I. Ahmed, L.A. Said, A. Elbaz, N. Herencsar, A. Soltan, Internet of things: a comprehensive overview on protocols, architectures, technologies, simulation tools, and future directions, Energies 16 (8) (2023) 3465.

[125] J.F. Landivar, K. Botirov, H. Sallouha, M. Katz, S. Pollin, Batteryless BLE and light-based IoT sensor nodes for reliable environmental sensing, in: 2024 IEEE 35Th International Symposium on Personal, Indoor and Mobile Radio Communications (PIMRC), IEEE, 2024, pp. 1–6.

[126] C. Orfanidis, K. Dimitrakopoulos, X. Fafoutis, M. Jacobsson, Towards battery-free LPWAN wearables, in: Proceedings of the 7Th International Workshop on Energy Harvesting & Energy-Neutral Sensing Systems, 2019, pp. 52–53.

[127] M. Katz, T. Paso, K. Mikhaylov, L. Pessoa, H. Fontes, L. Hakola, J. Leppäniemi, E. Carlos, G. Dolmans, J. Rufo, et al., Towards truly sustainable IoT systems: the SUPERIOT project, J. Phys.: Photon. 6 (1) (2024) 011001.

[128] N. Belapurkar, J. Harbour, S. Shelke, B. Aksanli, Building data-aware and energy-efficient smart spaces, IEEE Internet Things J. 5 (6) (2018) 4526–4537.

[129] G. Callebaut, G. Leenders, J. Van Mulders, G. Ottoy, L. De Strycker, L. Van der Perre, The art of designing remote iot devices-technologies and strategies for a long battery life, Sensors 21 (3) (2021) 913.

[130] W. Tuming, Y. Sijia, W. Hailong, A dynamic voltage scaling algorithm for wireless sensor networks, in: 2010 3Rd International Conference on Advanced Computer Theory and Engineering (ICACTE), 1, IEEE, 2010, pp. V1–554.

[131] A.-A.A. Boulogeorgos, P.D. Diamantoulakis, G.K. Karagiannidis, Low power wide area networks (lpwans) for internet of things (iot) applications: Research challenges and future trends, arXiv preprint (2016). arXiv:1611.07449

[132] L. Bereketeab, A. Zekeria, M. Aloqaily, M. Guizani, M. Debbah, Energy optimization in sustainable smart environments with machine learning and advanced communications, IEEE Sens. J. 24 (5) (2024) 5704–5712.

[133] A. Tsanousa, C. Moschou, E. Bektsis, S. Vrochidis, I. Kompatsiaris, Fusion of environmental sensors for occupancy detection in a real construction site, Sensors 23 (23) (2023) 9596.

[134] A. Martins, I. Fonseca, J.T. Farinha, J. Reis, A.J.M. Cardoso, Online monitoring of sensor calibration status to support condition-based maintenance, Sensors 23 (5) (2023) 2402.

[135] R. Laidi, D. Djenouri, UDEPLOY: User-driven learning for occupancy sensors DEPLOYment in smart buildings, in: 2018 IEEE International Conference on Pervasive Computing and Communications Workshops (PerCom Workshops), IEEE, 2018, pp. 209–214.

[136] M.A. Zaidan, Y. Xie, N.H. Motlagh, B. Wang, W. Nie, P. Nurmi, S. Tarkoma, T. Petäjä, A. Ding, M. Kulmala, Dense air quality sensor networks: validation, analysis, and benefits, IEEE Sens. J. 22 (23) (2022) 23507–23520.

[137] S.O. Guclu, T. Ozcelebi, J. Lukkien, Distributed fault detection in smart spaces based on trust management, Procedia Comput. Sci. 83 (2016) 66–73.

[138] R.N. Duche, N.P. Sarwade, Sensor node failure detection based on round trip delay and paths in WSNs, IEEE Sens. J. 14 (2) (2013) 455–464.

[139] P. Wongthongtham, J. Kaur, V. Potdar, A. Das, Big data challenges for the internet of things (IoT) paradigm, Connect. Environ. IoT: Challeng. Solut. (2017) 41–62.

[140] P. Azad, N.J. Navimipour, A.M. Rahmani, A. Sharifi, The role of structured and unstructured data managing mechanisms in the internet of things, Cluster Comput. 23 (2020) 1185–1198.

[141] M.S. Mahdavinejad, M. Rezvan, M. Barekatain, P. Adibi, P. Barnaghi, A.P. Sheth, Machine learning for internet of things data analysis: a survey, Digit. Commun. Netw. 4 (3) (2018) 161–175.

[142] V. Hayyolalam, M. Aloqaily, Ö. Özkasap, M. Guizani, Edge-assisted solutions for IoT-based connected healthcare systems: a literature review, IEEE Internet Things J. 9 (12) (2021) 9419–9443.

[143] J. Pan, J. McElhannon, Future edge cloud and edge computing for internet of things applications, IEEE Internet Things J. 5 (1) (2017) 439–449.

[144] J. Bhatia, K. Italiya, K. Jadeja, M. Kumhar, U. Chauhan, S. Tanwar, M. Bhavsar, R. Sharma, D.L. Manea, M. Verdes, M.S. Raboaca, An overview of fog data analytics for IoT applications, Sensors 23 (1) (2022) 199. https://doi.org/10.3390/s23010199

[145] H. Kokkonen, L. Lovén, N.H. Motlagh, A. Kumar, J. Partala, T. Nguyen, V.C. Pujol, P. Kostakos, T. Leppänen, A. González-Gil, et al., Autonomy and intelligence in the computing continuum: Challenges, enablers, and future directions for orchestration, arXiv preprint (2023). arXiv:2205.01423

[146] A. Varol, N.H. Motlagh, M. Leino, J. Virkki, Performance of large language models across edge and cloud platforms in smart spaces, in: Proc. SpliTech 2025, IEEE, Split, Croatia, 2025, pp. 1–6. https://doi.org/10.23919/SpliTech65624.2025.11091771

[147] K. Cao, S. Hu, Y. Shi, A.W. Colombo, S. Karnouskos, X. Li, A survey on edge and edge-cloud computing assisted cyber-physical systems, IEEE Trans. Ind. Inf. 17 (11) (2021) 7806–7819.

[148] S.C. Mukhopadhyay, S.K.S. Tyagi, N.K. Suryadevara, V. Piuri, F. Scotti, S. Zeadally, Artificial intelligence-based sensors for next generation IoT applications: a review, IEEE Sens. J. 21 (22) (2021) 24920–24932.

[149] Y. Cheng, X. He, Z. Zhou, L. Thiele, Ict: in-field calibration transfer for air quality sensor deployments, Proc. ACM Interact. Mobile Wearable Ubiquit. Technol. 3 (1) (2019) 1–19.

[150] K. Aula, E. Lagerspetz, P. Nurmi, S. Tarkoma, Evaluation of low-cost air quality sensor calibration models, ACM Trans. Sens. Netw. 18 (4) (2022) 1–32.

[151] M.A. Zaidan, N.H. Motlagh, P.L. Fung, D. Lu, H. Timonen, J. Kuula, J.V. Niemi, S. Tarkoma, T. Petäjä, M. Kulmala, et al., Intelligent calibration and virtual sensing for integrated low-cost air quality sensors, IEEE Sens. J. 20 (22) (2020) 13638–13652.

[152] K. Lakshmanna, R. Kaluri, N. Gundluru, Z.S. Alzamil, D.S. Rajput, A.A. Khan, M.A. Haq, A. Alhussen, A review on deep learning techniques for IoT data, Electronics (Basel) 11 (10) (2022) 1604.

[153] M.A. Zaidan, N.H. Motlagh, B.E. Boor, D. Lu, P. Nurmi, T. Petäjä, A. Ding, M. Kulmala, S. Tarkoma, T. Hussein, Virtual sensors: toward higH-resolution air pollution monitoring using ai and iot, IEEE IoT Mag. 6 (1) (2023) 76–81.

[154] X. Liu, F. Concas, N.H. Motlagh, M.A. Zaidan, P.L. Fung, S. Varjonen, J.V. Niemi, H. Timonen, T. Hussein, T. Petäjä, et al., Estimating black carbon levels with proxy variables and low-cost sensors, IEEE Internet Things J. 11 (10) (2024) 17577–17588.

[155] W. Bronzi, R. Frank, G. Castignani, T. Engel, Bluetooth low energy performance and robustness analysis for inter-vehicular communications, Ad Hoc Netw. 37 (2016) 76–86.

[156] G.D. Putra, A.R. Pratama, A. Lazovik, M. Aiello, Comparison of energy consumption in wi-Fi and bluetooth communication in a smart building, in: 2017 IEEE 7Th Annual Computing and Communication Workshop and Conference (CCWC), IEEE, 2017, pp. 1–6.

[157] D. Newell, M. Duffy, Review of power conversion and energy management for low-power, low-voltage energy harvesting powered wireless sensors, IEEE Trans. Power Electron. 34 (10) (2019) 9794–9805.

[158] K. Zhang, J. Ni, K. Yang, X. Liang, J. Ren, X.S. Shen, Security and privacy in smart city applications: challenges and solutions, IEEE Commun. Mag. 55 (1) (2017) 122–129.

[159] T. Lee, J. Han, M.-S. Lee, H.-S. Kim, S. Bahk, CABLE: Connection interval adaptation for BLE in dynamic wireless environments, in: 2017 14Th Annual IEEE International Conference on Sensing, Communication, and Networking (SECON), IEEE, 2017, pp. 1–9.

[160] K. Hong, S. Lee, K. Lee, Performance improvement in zigbee-based home networks with coexisting WLANs, Pervasive Mob Comput 19 (2015) 156–166.

[161] J. Mesquita, D. Guimarães, C. Pereira, F. Santos, L. Almeida, Assessing the ESP8266 wifi module for the internet of things, in: 2018 IEEE 23Rd International Conference on Emerging Technologies and Factory Automation (ETFA), 1, IEEE, 2018, pp. 784–791.

[162] M.A.B. Temim, G. Ferré, B. Laporte-Fauret, D. Dallet, B. Minger, L. Fuché, An enhanced receiver to decode superposed lora-like signals, IEEE Internet Things J. 7 (8) (2020) 7419–7431.

[163] G. Premsankar, B. Ghaddar, M. Slabicki, M. Di Francesco, Optimal configuration of lora networks in smart cities, IEEE Trans. Ind. Inf. 16 (12) (2020) 7243–7254.

[164] C.K. Nkuba, S. Woo, H. Lee, S. Dietrich, ZMAD: Lightweight model-Based anomaly detection for the structured Z-Wave protocol, IEEE Access 11 (2023) 60562–60577.

[165] K. Kim, K. Cho, J. Lim, Y.H. Jung, M.S. Sung, S.B. Kim, H.K. Kim, What's your protocol: vulnerabilities and security threats related to Z-Wave protocol, Pervasive Mob. Comput. 66 (2020) 101211.

[166] M.I. Ahmed, G. Kannan, Cloud-based remote RFID authentication for security of smart internet of things applications, J. Inf. Knowl. Manag. 20 (supp01) (2021) 2140004.

[167] A. Celesti, M. Fazio, M. Villari, Enabling secure XMPP communications in federated IoT clouds through XEP 0027 and SAML/SASL SSO, Sensors 17 (2) (2017) 301.

[168] J. Kotak, A. Shah, A. Shah, P. Rajdev, A comparative analysis on security of MQTT brokers, in: 2Nd Smart Cities Symposium (SCS 2019), 2019, pp. 1–5. https://doi.org/10.1049/cp.2019.0180

[169] E. Toé, D.F. Somé, T. Yélémou, Lightweight and robust MQTT protocol authentication model suitable for connected portals, in: 2023 IEEE Multi-conference on Natural and Engineering Sciences for Sahel's Sustainable Development (MNE3SD), IEEE, 2023, pp. 1–7.

[170] D. Ray, P. Bhale, S. Biswas, S. Nandi, P. Mitra, Daiss: design of an attacker identification scheme in coap request/response spoofing, in: TENCON 2021-2021 IEEE Region 10 Conference (TENCON), IEEE, 2021, pp. 941–946.

[171] Y.A. Joarder, C. Fung, Exploring QUIC security and privacy: a comprehensive survey on QUIC security and privacy vulnerabilities, threats, attacks, and future research directions, IEEE Trans. Netw. Serv. Manage. 21 (6) (2024) 6953–6973. https://doi.org/10.1109/TNSM.2024.3457858

[172] E. Lagerspetz, N.H. Motlagh, M.A. Zaidan, P.L. Fung, J. Mineraud, S. Varjonen, M. Siekkinen, P. Nurmi, Y. Matsumi, S. Tarkoma, et al., Megasense: feasibility of low-cost sensors for pollution hot-spot detection, in: 2019 IEEE 17Th International Conference on Industrial Informatics (INDIN), 1, IEEE, 2019, pp. 1083–1090.

[173] N.H. Motlagh, E. Lagerspetz, P. Nurmi, X. Li, S. Varjonen, J. Mineraud, M. Siekkinen, A. Rebeiro-Hargrave, T. Hussein, T. Petaja, et al., Toward massive scale air quality monitoring, IEEE Commun. Mag. 58 (2) (2020) 54–59.

[174] N.H. Motlagh, M.A. Zaidan, P.L. Fung, X. Li, Y. Matsumi, T. Petäjä, M. Kulmala, S. Tarkoma, T. Hussein, Low-cost air quality sensing process: validation by indoor-outdoor measurements, in: 2020 15Th IEEE Conference on Industrial Electronics and Applications (ICIEA), IEEE, 2020, pp. 223–228.

[175] F. Delaine, B. Lebental, H. Rivano, In situ calibration algorithms for environmental sensor networks: a review, IEEE Sens. J. 19 (15) (2019) 5968–5978.

[176] M.A. Burhanuddin, A.A.-J. Mohammed, R. Ismail, M.E. Hameed, A.N. Kareem, H. Basiron, A review on security challenges and features in wireless sensor networks: IoT perspective, J. Telecommun. Electron. Comput. Eng. (JTEC) 10 (1–7) (2018) 17–21.

[177] A.K. Ray, A. Bagwari, IoT Based smart home: security aspects and security architecture, in: 2020 IEEE 9Th International Conference on Communication Systems and Network Technologies (CSNT), IEEE, 2020, pp. 218–222.

[178] E. Fernandes, J. Jung, A. Prakash, Security analysis of emerging smart home applications, in: 2016 IEEE Symposium on Security and Privacy (SP), IEEE, 2016, pp. 636–654.

[179] G. Carlson, J. McKinney, E. Slezak, E.-S. Wilmot, General data protection regulation and california consumer privacy act: background, Currents: J. Int'l Econ. L. 24 (2020) 62.

[180] A. Alwarafy, K.A. Al-Thelaya, M. Abdallah, J. Schneider, M. Hamdi, A survey on security and privacy issues in edge-computing-assisted internet of things, IEEE Internet Things J. 8 (6) (2020) 4004–4022.

[181] J.D. Kelleher, B. Mac Namee, A. D'arcy, Fundamentals of machine learning for predictive data analytics: algorithms, worked examples, and case studies, MIT press, 2020.

[182] O. Majidzadeh Gorjani, R. Byrtus, J. Dohnal, P. Bilik, J. Koziorek, R. Martinek, Human activity classification using multilayer perceptron, Sensors 21 (18) (2021) 6207.

[183] S.T.M. Bourobou, Y. Yoo, User activity recognition in smart homes using pattern clustering applied to temporal ANN algorithm, Sensors 15 (5) (2015) 11953–11971.

[184] H. Ramirez, S.A. Velastin, I. Meza, E. Fabregas, D. Makris, G. Farias, Fall detection and activity recognition using human skeleton features, IEEE Access 9 (2021) 33532–33542.

[185] G. Violatto, A. Pandharipande, Anomaly classification in people counting and occupancy sensor systems, IEEE Sens. J. 20 (12) (2020) 6573–6581.

[186] M. Dampfhoffer, T. Mesquida, A. Valentian, L. Anghel, Backpropagation-based learning techniques for deep spiking neural networks: a survey, IEEE Trans. Neural Netw. Learn. Syst. 35 (9) (2023) 11906–11921.

[187] S.H. Choi, Spiking neural networks for biomedical signal analysis, Biomed. Eng. Lett. 14 (5) (2024) 955–966.

[188] H.H. Amin, W. Deabes, K. Bouazza, Clustering of user activities based on adaptive threshold spiking neural networks, in: 2017 Ninth International Conference on Ubiquitous and Future Networks (ICUFN), IEEE, 2017, pp. 1–6.

[189] V. Fra, E. Forno, R. Pignari, T.C. Stewart, E. Macii, G. Urgese, Human activity recognition: suitability of a neuromorphic approach for on-edge AIot applications, Neuromorphic Comput. Eng. 2 (1) (2022) 014006.

[190] S.A. Alnaqbi, H.M. Tawfik, Anomaly detection in smart homes based on kitchen activities and machine learning, in: 2023 16Th International Conference on Developments in eSystems Engineering (DeSE), IEEE, 2023, pp. 331–336.

[191] I.A. Akhinov, M.R.A. Cahyono, Development of smart home system based on artificial intelligence with variable learning rate to manage household energy consumption, in: 2021 International Conference on Artificial Intelligence and Mechatronics Systems (AIMS), IEEE, 2021, pp. 1–6.

[192] X. Zhou, X. Xu, W. Liang, Z. Zeng, Z. Yan, Deep-learning-enhanced multitarget detection for end–edge–cloud surveillance in smart IoT, IEEE Internet Things J. 8 (16) (2021) 12588–12596.

[193] H.-G. Shin, Y.-H. Choi, C.-P. Yoon, Movement path data generation from wi-Fi fingerprints for recurrent neural networks, Sensors 21 (8) (2021) 2823.

[194] J. Park, K. Jang, S.-B. Yang, Deep neural networks for activity recognition with multi-sensor data in a smart home, in: 2018 IEEE 4Th World Forum on Internet of Things (WF-IoT), IEEE, 2018, pp. 155–160.

[195] D. Singh, E. Merdivan, I. Psychoula, J. Kropf, S. Hanke, M. Geist, A. Holzinger, Human activity recognition using recurrent neural networks, in: Machine Learning and Knowledge Extraction: First IFIP TC 5, WG 8.4, 8.9, 12.9 International Cross-Domain Conference, CD-MAKE 2017, Reggio, Italy, August 29–September 1, 2017, Proceedings 1, Springer, 2017, pp. 267–274.

[196] D. Syed, H. Abu-Rub, A. Ghrayeb, S.S. Refaat, Household-level energy forecasting in smart buildings using a novel hybrid deep learning model, IEEE Access 9 (2021) 33498–33511.

[197] L. Lu, C. Zhang, K. Cao, T. Deng, Q. Yang, A multichannel CNN-GRU model for human activity recognition, IEEE Access 10 (2022) 66797–66810.

[198] N. Sarkar, R. Gupta, P.K. Keserwani, M.C. Govil, Air quality index prediction using an effective hybrid deep learning model, Environ. Pollut. 315 (2022) 120404.

[199] I. Jrhilifa, H. Ouadi, A. Jilbab, N. Mounir, Forecasting smart home electricity consumption using VMD-Bi-GRU, Energy Effic. 17 (4) (2024) 35.

[200] M. Dissem, M. Amayri, N. Bouguila, Neural architecture search for anomaly detection in time-series data of smart buildings: a reinforcement learning approach for optimal autoencoder design, IEEE Internet Things J. 11 (10) (2024) 18059–18073.

[201] Y. Wei, J. Jang-Jaccard, W. Xu, F. Sabrina, S. Camtepe, M. Boulic, LSTM-Autoencoder-based anomaly detection for indoor air quality time-series data, IEEE Sens. J. 23 (4) (2023) 3787–3800.

[202] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A.N. Gomez, Ł. Kaiser, I. Polosukhin, Attention is all you need, Adv. Neural Inf. Process. Syst. 30 (2017).

[203] D. Chen, S. Yongchareon, E.M.-K. Lai, J. Yu, Q.Z. Sheng, Y. Li, Transformer with bidirectional GRU for nonintrusive, sensor-Based activity recognition in a multiresident environment, IEEE Internet Things J. 9 (23) (2022) 23716–23727.

[204] X. Huang, S. Zhang, Human activity recognition based on transformer in smart home, in: Proceedings of the 2023 2Nd Asia Conference on Algorithms, Computing and Machine Learning, 2023, pp. 520–525.

[205] S. Xiao, S. Wang, Z. Huang, Y. Wang, H. Jiang, Two-stream transformer network for sensor-based human activity recognition, Neurocomputing 512 (2022) 253–268.

[206] Y. Mao, G. Zhang, C. Ye, A spatio-temporal graph transformer driven model for recognizing fine-grained data human activity, Alexandria Eng. J. 104 (2024) 31–45.

[207] D. Roy, B. Fernando, Action anticipation using pairwise human-object interactions and transformers, IEEE Trans. Image Process. 30 (2021) 8116–8129.

[208] H. Kim, D. Lee, TAP: A transformer based activity prediction exploiting temporal relations in collaborative tasks, in: 2021 IEEE International Conference on Pervasive Computing and Communications Workshops and Other Affiliated Events (PerCom Workshops), IEEE, 2021, pp. 20–25.

[209] Q. Zhang, H. Zhu, P. Wang, E. Chen, H. Xiong, Hierarchical wi-Fi trajectory embedding for indoor user mobility pattern analysis, Proc. ACM Interact. Mobile, Wearable Ubiquitous Technol. 7 (2) (2023) 1–21.

[210] F. Bader, L. Liebenow, A. Steinhage, Using deep learning to identify persons by their movement on a sensor floor, in: Proceedings of the 8Th International Workshop on Sensor-Based Activity Recognition and Artificial Intelligence, 2023, pp. 1–5.

[211] J. Chen, X. Xu, T. Wang, G. Jeon, D. Camacho, An AIot framework with multimodal frequency fusion for wifi-based coarse and fine activity recognition, IEEE Internet Things J. 11 (24) (2024) 39020–39029.

[212] A. Trivedi, K. Silverstein, E. Strubell, P. Shenoy, M. Iyyer, Wifimod: transformer-based indoor human mobility modeling using passive sensing, in: Proceedings of the 4Th ACM SIGCAS Conference on Computing and Sustainable Societies, 2021, pp. 126–137.

[213] Z. Zhang, H. Du, S. Choi, S.H. Cho, TIPS: Transformer based indoor positioning system using both CSI and DoA of wifi signal, IEEE Access 10 (2022) 111363–111376.

[214] Y. Zhou, H. Huang, S. Yuan, H. Zou, L. Xie, J. Yang, Metafi++: wifi-enabled transformer-based human pose estimation for metaverse avatar simulation, IEEE Internet Things J. 10 (16) (2023) 14128–14136.

[215] T.-H. Lee, H. Kim, D. Lee, Transformer based early classification for real-Time human activity recognition in smart homes, in: Proceedings of the 38Th ACM/SIGAPP Symposium on Applied Computing, 2023, pp. 410–417.

[216] B. Sheng, R. Han, H. Cai, F. Xiao, L. Gui, Z. Guo, CDFi: Cross-domain action recognition using wifi signals, IEEE Trans. Mob. Comput. 23 (8) (2024) 8463–8477.

[217] D. Gufran, S. Tiku, S. Pasricha, VITAL: Vision transformer neural networks for accurate smartphone heterogeneity resilient indoor localization, in: 2023 60Th ACM/IEEE Design Automation Conference (DAC), IEEE, 2023, pp. 1–6.

[218] M.Z. Khan, M. Usman, J. Ahmad, M.M.U. Rahman, H. Abbas, M. Imran, Q.H. Abbasi, Tag-free indoor fall detection using transformer network encoder and data fusion, Sci. Rep. 14 (1) (2024) 16763.

[219] H. Zhou, P. Zheng, J. Dong, J. Liu, Y. Nakanishi, Interpretable feature selection and deep learning for short-term probabilistic PV power forecasting in buildings using local monitoring data, Appl. Energy 376 (2024) 124271.

[220] L. Wang, S. Mao, R.M. Nelms, Transformer for nonintrusive load monitoring: complexity reduction and transferability, IEEE Internet Things J. 9 (19) (2022) 18987–18997.

[221] W. Ji, Z. Cao, X. Li, Multi-Task learning and temporal-Fusion-Transformer-Based forecasting of building power consumption, Electronics (Basel) 12 (22) (2023) 4656.

[222] S. Cen, C.G. Lim, Multi-Task learning of the patchTCN-TST model for short-Term multi-Load energy forecasting considering indoor environments in a smart building, IEEE Access 12 (2024) 19553–19568.

[223] X. Deng, Y. Zhang, H. Qi, Toward smart multizone HVAC control by combining context-Aware system and deep reinforcement learning, IEEE Internet Things J. 9 (21) (2022) 21010–21024.

[224] F. Akbari, K. Sartipi, A transformer-based model for older adult behavior change detection, in: 2022 IEEE 10Th International Conference on Healthcare Informatics (ICHI), IEEE, 2022, pp. 27–35.

[225] M. Tiwari, M. Kumar, A. Srivastava, A. Bala, Inbuilt chat GPT feature in smartwatches, in: 2023 International Conference on Circuit Power and Computing Technologies (ICCPCT), IEEE, 2023, pp. 1806–1813.

[226] Y. Tang, L. Zhang, H. Wu, J. He, A. Song, Dual-branch interactive networks on multichannel time series for human activity recognition, IEEE J. Biomed. Health Inform. 26 (10) (2022) 5223–5234.

[227] J. Bzai, F. Alam, A. Dhafer, M. Bojović, S.M. Altowaijri, I.K. Niazi, R. Mehmood, Machine learning-Enabled internet of things (IoT): data, applications, and industry perspective, Electronics (Basel) 11 (17) (2022) 2676.

[228] M. Abdel-Basset, V. Chang, H. Hawash, R.K. Chakrabortty, M. Ryan, Deep learning approaches for human-centered IoT applications in smart indoor environments: a contemporary survey, Ann. Oper. Res. (2021) 1–49.

[229] K.R. Dalal, Analysing the role of supervised and unsupervised machine learning in iot, in: 2020 International Conference on Electronics and Sustainable Communication Systems (ICESC), IEEE, 2020, pp. 75–79.

[230] M. Merenda, C. Porcaro, D. Iero, Edge machine learning for ai-enabled iot devices: a review, Sensors 20 (9) (2020) 2533.

[231] N. Hossein Motlagh, A. Zuniga, N. Thi Nguyen, H. Flores, J. Wang, S. Tarkoma, M. Prosperi, S. Helal, P. Nurmi, Population digital health: continuous health monitoring and profiling at scale, Online J. Public Health Inform. 16 (2024) e60261.

[232] Z. Wu, X. Wu, Y. Long, Prediction based semi-supervised online personalized federated learning for indoor localization, IEEE Sens. J. 22 (11) (2022) 10640–10654.

[233] A. Nikul Patel, G. Srivastava, P. Kumar Reddy Maddikunta, R. Murugan, G. Yenduri, T. Reddy Gadekallu, A trustable federated learning framework for rapid fire smoke detection at the edge in smart home environments, IEEE Internet Things J. 11 (23) (2024) 37708–37717. https://doi.org/10.1109/JIOT.2024.3439228

[234] H.M.S. Badar, K. Mahmood, W. Akram, Z. Ghaffar, M. Umar, A.K. Das, Secure authentication protocol for home area network in smart grid-based smart cities, Comput. Electr. Eng. 108 (2023) 108721.

[235] M.A.K. Raiaan, M.S.H. Mukta, K. Fatema, N.M. Fahad, S. Sakib, M.M.J. Mim, J. Ahmad, M.E. Ali, S. Azam, A review on large language models: architectures, applications, taxonomies, open issues and challenges, IEEE Access 12 (2024) 26839–26874.

[236] T. An, Y. Zhou, H. Zou, J. Yang, IoT-LLM: Enhancing Real-World IoT Task Reasoning with Large Language Models (2025). arXiv:2410.02429

[237] M. Giudici, G.A. Abbo, O. Belotti, A. Braccini, F. Dubini, R.A. Izzo, P. Crovari, F. Garzotto, Assessing LLMs responses in the field of domestic sustainability: an exploratory study, in: 2023 Third International Conference on Digital Data Processing (DDP), IEEE, 2023, pp. 42–48.

[238] F. Li, J. Huang, Y. Gao, W. Dong, Chatiot: zero-code generation of trigger-action based IoT programs with chatGPT, in: Proceedings of the 7Th Asia-Pacific Workshop on Networking, 2023, pp. 219–220.

[239] M. Simunec, R. Soic, Smart home notifications in croatian language: a transformer-Based approach, in: 2023 17Th International Conference on Telecommunications (ConTEL), IEEE, 2023, pp. 1–5.

[240] H. Xu, L. Han, Q. Yang, M. Li, M. Srivastava, Penetrative ai: making llms comprehend the physical world, in: Proceedings of the 25Th International Workshop on Mobile Computing Systems and Applications, 2024, pp. 1–7.

[241] N. Zhong, Y. Wang, R. Xiong, Y. Zheng, Y. Li, M. Ouyang, D. Shen, X. Zhu, Casit: collective intelligent agent system for internet of things, IEEE Internet Things J. 11 (11) (2024) 19646–19656.

[242] M. Romaszewski, P. Sekuła, P. Głomb, M. Cholewa, K. Kołodziej, Through the thicket: a study of number-Oriented LLMS derived from random forest models, J. Artif. Intell. Soft Comput. Res. 15 (3) (2025) 279–298.

[243] S. Lubos, T.N.T. Tran, A. Felfernig, S. Polat Erdeniz, V.-M. Le, LLM-Generated explanations for recommender systems, in: Adjunct Proceedings of the 32Nd ACM Conference on User Modeling, Adaptation and Personalization, 2024, pp. 276–285.

[244] X. Ma, L. Wang, N. Yang, F. Wei, J. Lin, Fine-tuning llama for multi-stage text retrieval, in: Proceedings of the 47Th International ACM SIGIR Conference on Research and Development in Information Retrieval, 2024, pp. 2421–2425.

[245] A. Radford, K. Narasimhan, T. Salimans, I. Sutskever, et al., Improving language understanding by generative pre-training (2018).

[246] N. Takeda, R. Legaspi, Y. Nishimura, K. Ikeda, A. Minamikawa, T. Plötz, S. Chernova, Sensor event sequence prediction for proactive smart home support using autoregressive language model, in: 2023 19Th International Conference on Intelligent Environments (IE), IEEE, 2023, pp. 1–8.

[247] T. Brown, B. Mann, N. Ryder, M. Subbiah, J.D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, et al., Language models are few-shot learners, Adv. Neural Inf. Process. Syst. 33 (2020) 1877–1901.

[248] E. King, H. Yu, S. Lee, C. Julien, " Get ready for a party": Exploring smarter smart spaces with help from large language models, arXiv preprint (2023). arXiv:2303.14143

[249] G. Civitarese, M. Fiori, P. Choudhary, C. Bettini, Large language models are zero-shot recognizers for activities of daily living, ACM Trans. Intell. Syst. Technol. 16 (4) (2025) 1–32.

[250] A. Saleh, P.K. Donta, R. Morabito, N.H. Motlagh, S. Tarkoma, L. Lovén, Follow-Me AI: energy-Efficient user interaction with smart environments, IEEE Pervasive Comput. 24 (1) (2025) 32–42. https://doi.org/10.1109/MPRV.2025.3539421

[251] J. Achiam, S. Adler, S. Agarwal, L. Ahmad, I. Akkaya, F.L. Aleman, D. Almeida, J. Altenschmidt, S. Altman, S. Anadkat, et al., Gpt-4 technical report, arXiv preprint (2024). arXiv:2303.08774

[252] A. Devi, S. Uttrani, A. Singla, S. Jha, N. Dasgupta, S. Natarajan, R.S. Punekar, L.A. Pickett, V. Dutt, Quantitative analysis of GPT-4 model: optimizing patient eligibility classification for clinical trials and reducing expert judgment dependency, in: Proceedings of the 2024 8Th International Conference on Medical and Health Informatics, 2024, pp. 230–237.

[253] A. Hurst, A. Lerer, A.P. Goucher, A. Perelman, A. Ramesh, A. Clark, A.J. Ostrow, A. Welihinda, A. Hayes, A. Radford, et al., Gpt-4o system card, arXiv preprint (2024). arXiv:2410.21276

[254] M. Andrao, D. Morra, T. Paccosi, M. Matera, B. Treccani, M. Zancanaro, " This sounds unclear": evaluating chatGPT capability in translating end-user prompts into ready-to-deploy python code, in: Proceedings of the 2024 International Conference on Advanced Visual Interfaces, 2024, pp. 1–4.

[255] E. King, H. Yu, S. Lee, C. Julien, Sasha: creative goal-oriented reasoning in smart homes with large language models, Proc. ACM Interact. Mobile, Wearable Ubiquitous Technol. 8 (1) (2024) 1–38.

[256] K.U. Ahn, D.-W. Kim, H.M. Cho, C.-U. Chae, Alternative approaches to HVAC control of chat generative pre-Trained transformer (chatGPT) for autonomous building system operations, Buildings 13 (11) (2023) 2680.

[257] H. Touvron, L. Martin, K. Stone, P. Albert, A. Almahairi, Y. Babaei, N. Bashlykov, S. Batra, P. Bhargava, S. Bhosale, et al., Llama 2: Open foundation and fine-tuned chat models, arXiv preprint (2023). arXiv:2307.09288

[258] H. Touvron, T. Lavril, G. Izacard, X. Martinet, M.-A. Lachaux, T. Lacroix, B. Rozière, N. Goyal, E. Hambro, F. Azhar, et al., Llama: Open and efficient foundation language models, arXiv preprint (2023). arXiv:2302.13971

[259] Z. Yin, M. Zhang, D. Kawahara, Harmony: A Home Agent for Responsive Management and Action Optimization with a Locally Deployed Large Language Model, arXiv preprint (2025). arXiv:2410.14252

[260] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, Bert: Pre-training of deep bidirectional transformers for language understanding (2019) 4171–4186. https://doi.org/10.18653/v1/N19-1423

[261] X. Sun, H. Ai, J. Tao, T. Hu, Y. Cheng, BERT-ADLOC: A secure crowdsourced indoor localization system based on BLE fingerprints, Appl. Soft Comput. 104 (2021) 107237.

[262] G. Team, R. Anil, S. Borgeaud, J.-B. Alayrac, J. Yu, R. Soricut, J. Schalkwyk, A.M. Dai, A. Hauth, K. Millican, et al., Gemini: a family of highly capable multimodal models, arXiv preprint (2025). arXiv:2312.11805

[263] G. Team, P. Georgiev, V.I. Lei, R. Burnell, L. Bai, A. Gulati, G. Tanzer, D. Vincent, Z. Pan, S. Wang, et al., Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context, arXiv preprint (2024). arXiv:2403.05530

[264] T. Sawada, T. Hasegawa, K. Yokoyama, M. Mizuno, Office-in-the-Loop for building HVAC control with multimodal foundation models, in: Proceedings of the 11Th ACM International Conference on Systems for Energy-Efficient Buildings, Cities, and Transportation, 2024, pp. 110–120.

[265] S. Berrezueta-Guzman, M. Kandil, M.-L. Martín-Ruiz, I.P. de la Cruz, S. Krusche, Exploring the efficacy of robotic assistants with chatGPT and claude in enhancing ADHD therapy: innovating treatment paradigms, in: 2024 International Conference on Intelligent Environments (IE), IEEE, 2024, pp. 25–32.

[266] J. Xu, Z. Li, W. Chen, Q. Wang, X. Gao, Q. Cai, Z. Ling, On-device language models: A comprehensive review, arXiv preprint (2024). arXiv:2409.00088

[267] E. King, H. Yu, S. Vartak, J. Jacob, S. Lee, C. Julien, Thoughtful Things: Building Human-Centric Smart Devices with Small Language Models, arXiv preprint (2024). arXiv:2405.03821

[268] Z. Liu, C. Zhao, F. Iandola, C. Lai, Y. Tian, I. Fedorov, Y. Xiong, E. Chang, Y. Shi, R. Krishnamoorthi, et al., Mobilellm: optimizing sub-billion parameter language models for on-device use cases, in: Forty-first International Conference on Machine Learning, 2024.

[269] X. Jiao, Y. Yin, L. Shang, X. Jiang, X. Chen, L. Li, F. Wang, Q. Liu, Tinybert: Distilling bert for natural language understanding, arXiv preprint (2020). arXiv:1909.10351

[270] P. Zhang, G. Zeng, T. Wang, W. Lu, Tinyllama: An open-source small language model, arXiv preprint (2024). arXiv:2401.02385

[271] Y. Tang, K. Han, F. Liu, Y. Ni, Y. Tian, Z. Bai, Y.-Q. Hu, S. Liu, S. Jui, Y. Wang, Rethinking optimization and architecture for tiny language models, in: Forty-first International Conference on Machine Learning, 2024.

[272] H. Fan, Y.-C. Lin, V. Prasanna, ELLIE: Energy-Efficient LLM inference at the edge via prefill-Decode splitting, in: 2025 IEEE 36Th International Conference on Application-specific Systems, Architectures and Processors (ASAP), IEEE, 2025, pp. 139–146.

[273] K. Liu, X. Zhou, L. Li, M^{2}2LLM: a multi-Dimensional optimization framework for LLM inference on mobile devices, IEEE Trans. Parallel Distrib. Syst. 36 (10) (2025) 2014–2029. https://doi.org/10.1109/TPDS.2025.3587445

[274] M. Ardakani, J. Malekar, R. Zand, LLMPi: Optimizing LLMs for high-Throughput on raspberry pi, in: Proceedings of the Computer Vision and Pattern Recognition Conference, 2025, pp. 6379–6388.

[275] G. Xu, Z. Hao, L. Shen, Y. Luo, F. Sun, X. Wang, H. Hu, Y. Wen, Coformer: collaborating with heterogeneous edge devices for scalable transformer inference, IEEE Trans. Comput. (2025) 1–14. https://doi.org/10.1109/TC.2025.3604473

[276] S. Ye, J. Du, L. Zeng, W. Ou, X. Chu, Y. Lu, X. Chen, Galaxy: a resource-efficient collaborative edge ai system for in-situ transformer inference, in: IEEE INFOCOM 2024-IEEE Conference on Computer Communications, IEEE, 2024, pp. 1001–1010.

[277] H. Jin, Y. Wu, Ce-collm: efficient and adaptive large language models through cloud-edge collaboration, in: 2025 IEEE International Conference on Web Services (ICWS), IEEE, 2025, pp. 316–323.

[278] M. Zhang, X. Shen, J. Cao, Z. Cui, S. Jiang, Edgeshard: efficient LLM inference via collaborative edge computing, IEEE Internet Things J. 12 (10) (2025) 13119–13131. https://doi.org/10.1109/JIOT.2024.3524255

[279] Y. Venkatesha, S. Kundu, P. Panda, Fast and Cost-effective Speculative Edge-Cloud Decoding with Early Exits (2025). arXiv:2505.21594

[280] A. Priyanshu, S. Vijay, A. Kumar, R. Naidu, F. Mireshghallah, Are chatbots ready for privacy-sensitive applications? an investigation into input regurgitation and prompt-induced sanitization, arXiv preprint (2023). arXiv:2305.15008

[281] N. Mireshghallah, M. Antoniak, Y. More, Y. Choi, G. Farnadi, Trust no bot: Discovering personal disclosures in human-llm conversations in the wild, arXiv preprint (2024). arXiv:2407.11438

[282] Z. Shi, Y. Wang, F. Yin, X. Chen, K.-W. Chang, C.-J. Hsieh, Red teaming language model detectors with language models, Trans. Associat. Comput. Linguistic. 12 (2024) 174–189.

[283] A. Shaikh, A. Varol, J. Virkki, From prompts to motors: man-in-the-Middle attacks on LLM-Enabled vacuum robots, IEEE Access 13 (2025) 137505–137513. https://doi.org/10.1109/ACCESS.2025.3595424

[284] H. Zhang, L.H. Li, T. Meng, K.-W. Chang, G.V.d. Broeck, On the paradox of learning to reason from data, arXiv preprint (2022). arXiv:2205.11502

[285] K. Valmeekam, A. Olmo, S. Sreedharan, S. Kambhampati, Large language models still can't plan (a benchmark for LLMs on planning and reasoning about change), in: NeurIPS 2022 Foundation Models for Decision Making Workshop, 2022.

[286] K. Valmeekam, K. Stechly, S. Kambhampati, LLMs Still Can't Plan; Can LRMs? A Preliminary Evaluation of OpenAI's o1 on PlanBench, arXiv preprint (2024). arXiv:2409.13373

[287] N. Borazjanizadeh, S.T. Piantadosi, Reliable reasoning beyond natural language, arXiv preprint (2024). arXiv:2407.11373

[288] I. Mirzadeh, K. Alizadeh, H. Shahrokhi, O. Tuzel, S. Bengio, M. Farajtabar, Gsm-symbolic: Understanding the limitations of mathematical reasoning in large language models, arXiv preprint (2025). arXiv:2410.05229

[289] S. Kambhampati, K. Valmeekam, L. Guan, M. Verma, K. Stechly, S. Bhambri, L. Saldyt, A. Murthy, Llms can't plan, but can help planning in llm-modulo frameworks, arXiv preprint (2024). arXiv:2402.01817

[290] Y. Zhao, H. Yin, B. Zeng, H. Wang, T. Shi, C. Lyu, L. Wang, W. Luo, K. Zhang, Marco-o1: Towards Open Reasoning Models for Open-Ended Solutions, arXiv preprint (2024). arXiv:2411.14405

[291] Y. Zhang, Y. Li, L. Cui, D. Cai, L. Liu, T. Fu, X. Huang, E. Zhao, Y. Zhang, Y. Chen, et al., Siren's song in the AI ocean: a survey on hallucination in large language models, Comput. Linguist. (2025) 1–46.

[292] Z. Xu, S. Jain, M. Kankanhalli, Hallucination is inevitable: An innate limitation of large language models, arXiv preprint (2025). arXiv:2401.11817

[293] S. Tonmoy, S.M. Zaman, V. Jain, A. Rani, V. Rawte, A. Chadha, A. Das, A comprehensive survey of hallucination mitigation techniques in large language models, arXiv preprint 6 (2024). arXiv:2401.01313

[294] M. Li, W. Zhang, D. Xia, Transformer inference acceleration in edge computing environment, in: 2023 IEEE/ACM 23Rd International Symposium on Cluster, Cloud and Internet Computing Workshops (CCGridW), IEEE, 2023, pp. 104–109.

[295] Y. Sun, J. Ortiz, An AI-Based system utilizing IoT-Enabled ambient sensors and LLMs for complex activity tracking, Acad. J. Sci. Technol. 11 (3) (2024) 277-281. https://doi.org/10.54097/dj2pt496

[296] S. Samsi, D. Zhao, J. McDonald, B. Li, A. Michaleas, M. Jones, W. Bergeron, J. Kepner, D. Tiwari, V. Gadepally, From words to watts: benchmarking the energy costs of large language model inference, in: 2023 IEEE High Performance Extreme Computing Conference (HPEC), IEEE, 2023, pp. 1–9.

[297] A. Faiz, S. Kaneda, R. Wang, R. Osi, P. Sharma, F. Chen, L. Jiang, Llmcarbon: Modeling the end-to-end carbon footprint of large language models, arXiv preprint (2024). arXiv:2309.14393

[298] E. Strubell, A. Ganesh, A. McCallum, Energy and policy considerations for modern deep learning research, in: Proceedings of the AAAI Conference on Artificial Intelligence, 34, 2020, pp. 13693–13696.